School of Engineering
InIT Institute of Computer Science

Bachelor thesis

Is Bern Lagging Behind Zurich?
An In-depth Analysis of Swiss Dialects

Author	Louis Berthele Alina Spangenberg
Main supervisor	Mark Cieliebak
Sub supervisor	Don Tuggener
External supervisor	Anja Hasse

Date	07.06.2024

Declaration of Originality

Bachelor's Thesis at the School of Engineering

By submitting this Bachelor's thesis, the undersigned student confirms that this thesis is his/her own work and was written without the help of a third party. (Group works: the performance of the other group members are not considered as third party).

The student declares that all sources in the text (including Internet pages) and appendices have been correctly disclosed. This means that there has been no plagiarism, i.e. no sections of the Bachelor thesis have been partially or wholly taken from other texts and represented as the student's own work or included without being correctly referenced.

Any misconduct will be dealt with according to paragraphs 39 and 40 of the General Academic Regulations for Bachelor's and Master's Degree courses at the Zurich University of Applied Sciences (Rahmenprüfungsordnung ZHAW (RPO)) and subject to the provisions for disciplinary action stipulated in the University regulations.

Generative AI systems and AI tools were used in various process phases of this bachelor's thesis. Specifically, ChatGPT was used to improve the scripts and the text. ChatGPT was also used to understand existing code snippets.

City, Date:	**Name Student:**
Fribourg, 07.06.24	Louis Berthele
Siggenthal Station, 07.06.24	Alina Spangenberg

Is Bern Lagging Behind Zurich?
An In-depth Analysis of Swiss Dialects

Louis Berthele
berthlou@students.zhaw.ch

Alina Spangenberg
spangali@students.zhaw.ch

Zurich University of Applied Sciences
School of Engineering
Switzerland

February 2024 - June 2024

Abstract

This project explores the variation in speech speeds across different Swiss dialects, challenging longstanding assumptions about regional speech patterns in Switzerland. It uses two large datasets comprising approximately 400'000 audio files. It employs data cleansing and advanced linguistic analysis to measure and compare the speech speeds of speakers from various Swiss dialect regions, expressed in words per second.

While the research was initially inspired by the idea that speakers from Bern might be slower than those from Zürich, the study broadened its focus to encompass a wider range of dialect regions. By conducting a comprehensive analysis, this study offers novel insights into the intricacies of Swiss German dialects.

Throughout the thesis, an additional focus has developed on the in-depth examination of transcriptions, using existing Standard German Speech-To-Text models and their characteristics to assess word retrieval rates and analyze phonetic proximity between different dialects and Standard German. This includes a data-driven approach to identifying and computing dialect-specific words, comparing them across various Swiss dialects. Additionally, the study investigates the replacement of Standard German words with dialect-specific alternatives within these dialects, offering a detailed comparison of regional linguistic variations.

The findings of this analysis suggest that the widespread belief that Zürich speakers talk faster than those in Bern is not universally justifiable, meaning the hypothesis could not be confirmed. Only minimal differences in speech speeds across dialect regions were present in the final results. Furthermore, the analysis of dialect-specific words reveals significant regional variations, highlighting the unique linguistic characteristics of each dialect.

This project enhances the understanding of linguistic diversity within Switzerland and lays a foundation for future research on dialect variations and their implications for communication.

Zusammenfassung

Dieses Projekt befasst sich mit der Vielfalt der Sprechgeschwindigkeiten verschiedener Schweizer Dialekten und hinterfragt langjährige Annahmen über regionale Sprachmuster in der Schweiz. In dieser Analyse werden zwei grosse Datensätze mit circa 400'000 Audioaufnahmen verwendet. Im Verlauf der Studie werden Daten bereinigt und fortschrittliche linguistische Analysen durchgeführt, um Sprechgeschwindigkeiten messen und vergleichen zu können. Gerechnet wird in Wörtern pro Sekunde der verschiedenen Aufnahmen beziehungsweise Sprecher aus den schweizer Dialektregionen.

Während dieses Projekt ursprünglich durch die Idee inspiriert wurde, dass Sprecher aus Bern langsamer sprechen würden als diejenigen aus Zürich, erweitert die Studie ihren Fokus auf eine breitere Palette von Dialektregionen. Durch umfassende Analysen werden Einblicke in die Grundlagen und Details der Schweizerdeutschen Dialekte dargelegt.

Im Verlauf des Projekts hat sich ein weiterer Schwerpunkt mit der Untersuchung von Transkriptionen herauskristallisiert. Hierbei werden bestehende Hochdeutsche Speech-To-Text-Modelle mit deren spezifischen Eigenschaften und Limitierungen zur Bewertung von Retrieval Rates und zur Berechnung der phonetischen Nähe zwischen verschiedenen Dialekten und dem Hochdeutschen verwendet. Dies umfasst unter anderem einen datengesteuerten Ansatz zur Identifizierung dialektspezifischer Wörter, die über verschiedene Schweizer Dialekte hinweg verglichen werden. Darüber hinaus untersucht die Studie mögliche Äquivalente der Standartdeutschen Wörter durch dialektspezifische Alternativen der einzelnen Dialekte und bietet einen detaillierten Vergleich regionaler sprachlicher Vielfältigkeit.

Die Resultate dieser Analyse deuten daraufhin, dass der weit verbreitete Glaube, Züricher sprechen schneller als Berner, nicht definitv bestimmt werden konnte. Diese Analyse konnte somit die Hypothese nicht bestätigen, da nur minimale Unterschiede der Sprechgeschwindigkeiten zwischen den Dialektregionen in den Endresultaten festgestellt werden konnten. Darüber hinaus zeigt die Analyse dialektspezifischer Wörter ausserdem regionale Differenzen auf, die die sprachlichen Merkmale eines jeden Dialekts hervorheben.

Dieses Projekt dient zur Verbesserung des Verständnisses der sprachlichen Vielfalt innerhalb der Schweiz und soll eine Grundlage für zukünftige Forschungen zur Dialektvielfalt und dessen Auswirkung auf die Kommunikation darstellen.

Preface

This thesis marks a significant milestone in our academic journey, offering profound insights into the phonetic distance and speech speed analysis within Swiss German dialects. We appreciate the opportunity to engage with this subject, which has fostered our personal and professional growth throughout the process. We are excited about the prospects of dialect research and the advancements it may bring to the linguistic community.

We are deeply thankful to our supervisors, Prof. Dr. Marc Cieliebak and Prof. Dr. Don Tuggener, for their support, invaluable guidance, and enriching discussions that greatly enhanced our research experience. Their expertise and insights were crucial in steering this project to a successful conclusion.

We are also grateful to our family and friends for their encouragement and constructive feedback. Their support has been a cornerstone of our motivation and success.

We look forward to contributing further to the field of linguistics with the knowledge and experience we have gained during our studies.

Contents

1 Introduction

Language dynamically mirrors the diverse cultural landscapes that shape societies. In Switzerland, renowned for its multilingual heritage, the nuances of dialect variation provide a rich area for linguistic exploration. This bachelor thesis investigates an intriguing aspect of this variation: the speech speeds, dialect differences, and dialect-specific vocabulary across various German-speaking Swiss dialect regions.

This research questions the common belief that speech tempo varies greatly between different regions in Switzerland. While the initial focus was on Bern and Zürich, the study quickly expanded to compare all Swiss dialect regions. Although many think Bernese speakers are slower than those from Zürich, there is limited data on this topic. This study aims to bridge this gap by using a thorough analytical approach to measure and compare speech speeds throughout various regions of Switzerland. Moreover, the research seeks to assess the linguistic similarity of Swiss dialects to Standard German. Audio data from the STT4SG-350 and SDS-200 datasets are used to transcribe Swiss German dialects with a Standard German Speech-To-Text system. These transcriptions are then compared to the original text to determine how words in the dialects differ from their Standard German counterparts to identify dialect-specific words. This method provides a data-driven approach to understanding which dialect words are used instead of Standard German words.

This thesis aims to understand how dialect-specific variations in Swiss German influence communication, and interact with Standard German. By analyzing and transcribing spoken language from the two audio datasets mentioned above, this research reveals the complexities of Swiss German, identifying which dialects closely resemble Standard German and which have a unique lexicon[1].

This research addresses key challenges in linguistic data analysis, including data integrity, transcription methodologies, and retrieval rate accuracy. It focuses on improving data cleansing processes, refining transcription accuracy, and enhancing analytical methods to provide reliable insights for future research in Swiss dialectology[2] and computational linguistics.

[1] In this context, 'lexicon' refers to words unlike Standard German, featuring a unique vocabulary specific to the dialect.

[2] Dialectology studies dialect variations caused by geographic or social isolation, often mapping these differences in linguistic atlases [1].

1.1 Initial situation

Exploring variations in speech speeds across Swiss dialects is a topic of academic and public interest. Anecdotal[3] evidence has suggested differences in speech tempos among Swiss regions, particularly that Bernese speech is slower than Zürich. However, empirical data supporting these claims is limited and often inconclusive.

To address this gap, the Zurich University of Applied Sciences (ZHAW) utilized two extensive datasets, which included around 400'000 records representing a broad spectrum of Swiss German dialects. Before conducting the main analysis, a thorough data cleansing was undertaken to eliminate any recordings that could distort the results of the speech speed analysis, such as those with excessive background noise or non-standard speech patterns.

The initial phase ensured the reliability of the subsequent analyses. It aimed to assess and compare speech speeds across the dialect regions objectively, challenge traditional views of regional speech tempos, and improve understanding of the linguistic diversity within Swiss German dialects.

By analyzing these aspects, the thesis sought to illuminate Switzerland's linguistic nuances and offer insights that could guide future research in dialect studies and computer linguistics.

1.2 Objective

The primary objective of this bachelor thesis is to investigate the linguistic similarities and differences between Swiss German dialects and Standard German. This study leverages two large datasets covering all major Swiss dialect regions, including Zürich, Bern, and Wallis. The research begins by accurately measuring and comparing speech speeds across these regions. The audio recordings are also transcribed using a Standard German speech-to-text (STT) system. These transcriptions are then compared to the original Standard German texts to identify which dialect words significantly differ from their Standard German equivalents and which are closely similar. This analysis will address several key research questions:

- How does speech speed vary across different Swiss dialects?
- How phonetically different are the various dialects compared to Standard German?
- Which Standard German words have their own dialect-specific word forms, what are these, and how often are they used?
- Do these dialect-specific word forms vary among regions, and how do they impact the phonetic distance from Standard German?

The findings from this study are expected to provide insights into the phonetic and structural peculiarities of Swiss German dialects, thereby contributing to a broader understanding of Switzerland's linguistic diversity.

[3] In this context, 'anecdotal' refers to information based on personal accounts, observations, or informal evidence rather than systematic scientific research.

2 Theoretical Basics

This chapter outlines the foundational elements of our research on the speech speeds and linguistic characteristics of Swiss German dialects. We begin by introducing and defining key linguistic terms that form the base for understanding the subsequent study. Next, we review core theories and concepts from previous studies that inform our analytical approach. Finally, we detail the corpora used, including their origins, composition, and relevance. This chapter provides the necessary framework for understanding the methods and concepts used in our analysis of Swiss German dialects.

2.1 Key Terms

Language Systems, Heteronyms, and Cognates A significant part of the study involves comparing Standard German words to their dialectal versions, so defining the designations used for these variations is essential. The following terms can have different meanings [2] and are used in multiple contexts within linguistic research. Within the scope of this thesis, it is necessary to clarify the meanings of these terms and to emphasize that Standard German and Swiss German dialects are considered distinct language systems.

When the designation 'heteronym' is used for a dialect-specific word form that is the counterpart of a Standard German word, it refers to a fundamentally different and not etymologically related form (as defined in Definition 1b of the Duden Dictionary [2]). For example, 'Müll' is in Standard German, and 'Ghüdder' is in Bernese Swiss German. Additionally, if the dialect-specific word form is also used in Standard German but differs in dialectal usage (e.g., 'Knabe' in Standard German and 'Bub' in Bernese Swiss German), it is also considered a 'heteronym'. These instances might be considered synonyms if Standard German and Swiss German dialects were regarded as the same system.

On the other hand, when the designation 'cognate' is used for a dialect-specific word form that is the counterpart of a Standard German word, it refers to an etymologically related form (as defined in Definition 3b of the Merriam-Webster Dictionary [3]). Cognates exist on a continuum, ranging from a perfect match with identical pronunciation (e.g., 'Das' in Standard German and 'Das' in Zurich Swiss German) to more divergent forms (e.g., 'Ball' in Standard German and 'Bölä' in Zurich Swiss German).

By clearly defining these terms, we can accurately analyze and compare the linguistic characteristics of Standard German and Swiss German dialects within the framework of our research.

Phonetic Distance Phonetic distance measures the difference in pronunciation characteristics between dialects or languages. This metric is crucial for evaluating the similarities and differences in pronunciation patterns within and between dialect regions. The Levenshtein distance theory, which quantifies the number of changes needed to transform one phonetic sequence into another, enhances this concept. Calculating the minimum number of insertions, deletions, or substitutions required provides a precise measurement of phonetic distance, offering valuable insights into the phonological variations that distinguish dialects. More details on the Levenshtein distance are provided in the Glossary (Chapter 8.4) [4].

Transcription In linguistics, transcription is a method for accurately converting spoken language into written text. This process is important in dialectology, where it captures variations in pronunciation and linguistic usage that standard writing systems do not typically represent. Such transcriptions are used for analyzing and documenting dialectal differences [5].

Automatic Speech Recognition (ASR) and Speech-to-Text (STT) ASR utilizes computer algorithms to process and understand spoken language. ASR involves various tasks related to the recognition and interpretation of human speech. STT, a specific application of ASR, converts spoken language into written text. This study uses STT to transcribe spoken Swiss dialects into text, enabling quantitative analysis of dialectal speech characteristics [6].

Text-to-Speech (TTS) TTS technology converts written text into spoken voice output.

These key terms form the foundational vocabulary necessary to engage with the theoretical and practical aspects of the research presented in this thesis. They recur throughout the discussions on methodology, data analysis, and results, providing a framework for understanding the complexities of linguistic variation in Swiss German dialects.

2.2 Core Theories and Concepts Overview

This subsection delves into the theoretical frameworks and conceptual models that are pivotal in analyzing the linguistic features of Swiss German dialects. It serves as a guide to understanding the principles and theories that underpin our methodological approach and provides a lens through which the collected data is interpreted and analyzed.

This section related the theories to the practical challenges and technological applications of the research, such as speech speed analysis and dialect transcription. Understanding these theoretical aspects was necessary to analyze linguistic data from diverse Swiss dialects and draw conclusions about language usage, evolution, and identity across Switzerland's linguistic landscapes.

2.2.1 State of research

In exploring the state of research, particular attention was paid to studies that had analyzed dialectal variations, speech tempo, and transcription accuracy across different linguistic regions. Each reference included in this review was selected for its relevance and contribution to the field, providing a foundation for the subsequent analysis presented in this thesis. Through this detailed examination, the section underscored the novelty and necessity of the current research, setting the stage for a deeper investigation into the nuanced interactions between speech speed and dialect comprehension.

NLP - Project 3, Swiss german: A key starting point for our theoretical exploration is the study conducted by Kilian Pfister and Marc Zuber [7], as detailed in their paper. They performed a statistical analysis of Swiss German dialects in the test part of the STT4SG-350 dataset, using extensive audio data to compare the speaking rates across various regions. Their findings, which challenge some traditional assumptions about dialectal speech speeds, provide a foundation for the hypotheses tested in this thesis. The approach they adopted in analyzing dialectal differences in speech speed using metadata offers methodological insights and underscores the complexity of linguistic diversity across Swiss dialects.

The results of their speed analysis are presented in Table 1 and lay a foundation for our further work. The table displays the various dialect regions of Switzerland, ordered from the slowest-speaking region, 'Ostschweiz', to the fastest, 'Wallis'. Both characters per second (CPS) and words per second (WPS) are shown. Bern is among the slower regions, ranking second to last, while Wallis is one of the fastest regions.

dialect_region	cps	wps
Ostschweiz	9.99607	1.55815
Bern	10.19271	1.58931
Zürich	10.48183	1.63605
Innerschweiz	10.57417	1.64993
Basel	10.86201	1.69648
Graubünden	11.16148	1.74231
Wallis	11.52777	1.79961

Table 1: Speed Analysis of the Report from Pfister and Zuber [7]

We utilized this information to conduct initial analyses and verify their accuracy. Subsequently, this method was applied to both datasets. The detailed functionality of this methodology will be discussed later in Chapter 5 of this thesis.

Quantitative Analysis of Speech Tempo in Various Swiss German Dialects: [4]

In their comprehensive study [8], Alessio Drigatti and Lea Keller investigated the speech tempo across various Swiss German dialects using a dataset consisting of 24'603 audio recordings equipped with metadata such as dialect region and demographic data. Their research, articulated in their report, involved developing a pipeline to analyze these records by measuring the number of spoken sounds per second (LPS), thereby examining differences in speech tempo between various Swiss dialects. These insights offer valuable guidance for further linguistic analysis and applications in NLP systems. This thorough approach advanced our understanding of vernacular[5] differences in Swiss German and enhanced the methodological tools available for future research in this area.

Drigatti and Keller's work is particularly relevant for this thesis as it established a foundation for examining how regional variations in speech tempo could influence the comprehension and transcription of Swiss German dialects. Their findings serve as a reference point for our analysis, especially in understanding the complexity of linguistic diversity and its implications on automated speech recognition technologies.

Dialect Transfer for Swiss German Speech Translation: In addition to insights from prior research, a paper titled 'Dialect Transfer for Swiss German Speech Translation', submitted to Empirical Methods in Natural Language Processing (EMNLP) 2023 [10], was incorporated into our analysis. This study examined the challenges of creating Swiss German speech translation systems, with a particular focus on dialect diversity and linguistic differences between Swiss German and Standard German.

Swiss German, lacking a formal writing system and consisting of diverse dialects, presents significant challenges for speech translation. This is particularly due to its status as a language with only about 5.4 million speakers, representing approximately 62% of the Swiss population in German-speaking regions [11]. The study posed two main research questions: the impact of including or excluding dialects during the training of speech translation models on performance across specific dialects and the effects of linguistic differences between Swiss German and Standard German on the system's effectiveness.

The findings indicated that the diversity of dialects and linguistic differences impact translation performance [10]. Models trained without specific dialect data performed poorly on those dialects, highlighting the necessity for inclusive training approaches that encompass the full range of dialectal variations within Swiss German.

This paper's contributions helped shape the methods used in our thesis. The empirical approach adopted to test these models across varied dialect settings provided a foundational methodology that informed our analysis of speech speed and dialect comprehension across Swiss German dialects. The study's focus on the impacts of dialect inclusion in model training offers valuable lessons for optimizing speech translation systems in linguistically diverse environments. This alignment with our research objectives enhanced our understanding of linguistic variability and its implications for natural language processing systems.

Dialäkt Äpp: These three studies established a solid foundation for our research. Additionally, other references further supported our analysis. One notable resource is the Dialäkt Äpp [12], which is freely available[6] and enables the geographical localization of Swiss dialects based on linguistic

[4] Original German tile: Quantitative Analyse des Sprechtempos verschiedener schweizerdeutscher Dialekte

[5] In this context, 'vernacular' refers to the everyday language or dialect spoken by people in a specific region of Switzerland [9].

[6] Dialäkt Äpp in the Apple Store

standards. Users can record their dialect, listen to others' dialects, and access historical recordings from the Phonogram Archive of the University of Zurich[7] via an interactive map. The app also helps archive contemporary Swiss dialects as valuable cultural assets. Furthermore, users can receive information about the 'Dialect Word of the Week', provided by the Swiss German Dictionary (Idiotikon) [13].

Anonymous recordings could be used by the Phonetics Laboratory of the University of Zurich for scientific purposes, adding a layer of practical application and data gathering to our theoretical exploration of Swiss German dialects. This tool exemplified how digital technologies could support the dynamic study and preservation of linguistic heritage in real-time. Although we did not use this app directly in our analysis, it provided valuable insights into how others tackle the study of dialects.

Dialect versus Accent: The distinction between a dialect and an accent became a focal point during the research. A dialect involves variations in pronunciation, vocabulary, and grammar from the standard language, such as the differences found in Swiss German dialects compared to Standard German. An accent pertains only to pronunciation and does not affect vocabulary or grammar. Dialects and accents often reflect geographic, social, or other shared characteristics among speakers [14].

An accent is often seen as part of a dialect, much like a dialect is a subset of a language. For example, Swiss German dialects in Zürich, Bern, and Wallis show unique vocabulary, grammar, and pronunciation. Additionally, lectal[8] varieties within these dialects can further distinguish local pronunciations.

American English, for instance, is a dialect of English due to its distinct vocabulary, grammar, and pronunciation compared to other English-speaking regions like Canada or Britain [14]. Similarly, Zürich, Bernese, and Wallis German illustrate broad linguistic differences, including grammar and vocabulary variations within Swiss German. Lectal varieties within these dialects further refine pronunciation differences, aiding in accurate speech data analysis and transcription by accounting for both dialectal and accent-specific variations.

Examining words in Standard German and comparing them to their Swiss German forms constituted a foundation for this analysis. Various variants were encountered, forming a continuum. At one end, there are heteronyms replaced by dialect-specific word forms that are inexistent in Standard German (e.g., 'Müll' -> 'Ghüdder') or that exist in Standard German (e.g., 'Knaben' -> 'Buben'). In the middle, there are compound cognates and heteronyms (e.g., 'herunterladen' -> 'abelade'), and at the other end, there are cognates with significant changes (e.g., 'Ball' -> 'Bölä'). No clear boundary defines when a replacement word in a dialect can be identified as a dialect-specific word. Generally, distinguishing between dialects and languages is a longstanding issue in linguistics and has been a persistent and critical challenge throughout history [16]. This ambiguity challenged the research of dialect-specific words in this thesis.

This brief overview provides an initial insight into linguistics and enhances the understanding of the results presented in this thesis. The differences between dialect and accent are not discussed in detail within the thesis but are included here to provide a deeper understanding of the work.

Preterite Tense Shrinkage: [9] To define the preterite tense[10] and discuss its shrinkage, we first present an example to illustrate the preterite tense. The preterite tense does not exist in

[7] The Phonogram Archive can be found online

[8] 'Lectal' refers to variations within a language, such as dialects or sociolects [15].

[9] This refers to the fading of the preterite tense, also known in Standard German as 'Präteritumschwund'.

[10] Preterite Tense is the German past tense

Swiss German dialects, which only use the present perfect tense[11]. In spoken Standard German, the present perfect often replaces the preterite tense [17], a phenomenon entirely adopted in Swiss German [18]. An example is the sentence 'Ich ging gestern zum Markt.' which is 'Ig bi gester a Märit gange.' in Bernese Swiss German. The verb 'ging' in the preterite tense is transformed to 'bi gange'/'bin gegangen' in Perfekt.

The phenomenon of the decline of the preterite tense has been extensively studied in the context of German linguistics. Werner Abraham and C. Jac Conradie (2002) explore this phenomenon in their book 'Präteritumschwund und Diskursgrammatik', [19] highlighting the syntactic and pragmatic factors that contribute to the decreasing use of the preterite tense in favor of the perfect tense in spoken German. Hanna Fischer also examines this shift in her article 'Präteritumschwund im Deutschen: Neue Erkenntnisse zu einem alten Rätsel', [20] offering new insights into the historical and regional variations of this trend within the German language. She created a graphic, shown below in Figure 1, which illustrates the geographical distribution of the preterite tense usage in spoken German. The graphic highlights areas in Germany where the past tense is still commonly used, regions experiencing a reduction in past tense usage, and areas where it has already faded.

Additionally, it can be observed in Figure 1 that areas near Switzerland and Austria are the most affected by the decline in preterite tense usage. A relevant question is whether the Swiss and Austrians[12] influence this linguistic change.

Figure 1: Geographical Distribution of Preterite Tense Usage in Spoken German

These studies inform our work by providing a detailed understanding of the grammatical shifts in Standard German dialects. This understanding aids in accurately transcribing and analyzing speech data in our research, ensuring that our linguistic analyses account for these variations.

[11] Present perfect tense is known in German as 'Perfekt'

[12] Austrians also appear to avoid using the past tense in spoken language [21]

Some sentences in the metadata provided with the datasets (STT4SG-350 and SDS-200) contain written sentences that include the preterite tense. This highlights another difficulty in our work: the given sentences are written in Standard German and then converted into the spoken everyday language of each Swiss German dialect. This is an issue, as verbs written in preterite tense greatly differ from their spoken form, being in direct conflict with some of our measurement techniques which will be illustrated later in this report.

Word: For most of our analyses, we used word-level measurements such as WPS or Retrieval Rate (RR), yet it raises the question of what exactly a word is and how it is defined. A word is a fundamental unit of language that conveys meaning or function and is used in sentences to communicate information. Words consist of one or more sounds and have distinct meanings or functions. They are the smallest units in a language that can independently convey meaning or perform a function. Words can be classified into lexical categories (nouns, verbs, adjectives, adverbs), grammatical categories (pronouns, prepositions, conjunctions, articles), and functional categories (interjections, particles) [22].

In linguistic studies, a lemma represents a word's base form, encompassing all its inflected forms in a dictionary entry. For example, 'run' is the lemma for 'runs', 'ran', and 'running' [23]. In this thesis, we did not perform lemmatization[13] and referred only to word forms as they appear in texts. By accurately defining a word, we ensure the precision of our calculations, which is fundamental to our study.

Conclusion Websites like Bern Tourism's page on 'Fun Facts' about Bern[14], the Adelboden Hotel Blog's discussion on 'Bernese Slowness'[15], and the University of Zurich's Citizen Science project blog[16] has provided cultural and societal contexts that highlight the perception and impact of these dialects in everyday life. Additionally, exploring transcription tools like the Dialäkt Äpp, which attempted to automate the transcription of Swiss German dialects, offered insights into the practical challenges of dialect processing in technological applications. These experiences illustrated the challenges of accurately capturing and transcribing the nuanced spoken forms of Swiss dialects, revealing limitations in current technologies.

This combination of academic research, cultural insights, and transcription tool evaluations provides a comprehensive understanding of Swiss German, setting the stage for the empirical research that follows in this thesis. By examining these sources, this thesis aims to bridge the gap between linguistic theory and the practical realities of Swiss dialects, contributing to a detailed appreciation of Switzerland's linguistic landscape.

[13] Lemmatization reduces inflected word forms to their base or dictionary form, known as a lemma, through comprehensive morphological analysis [24].

[14] Link to the Fun Facts of Bern

[15] Link to the Adelboden Blog's discussion

[16] Link to the University of Zurich's Citizen Science project blog

2.3 Corpora

The backbone of this research consists of two large corpora, which we will refer to as datasets from now on. Together, they comprise around 400'000 audio files accompanied by a detailed TSV metadata file. Both datasets cover various dialect regions within Switzerland, including Zürich, Bern, Innerschweiz, Ostschweiz, Graubünden, Basel, and Wallis, providing a comprehensive linguistic snapshot of the country's regional speech variations.

SDS-200 Dataset The SDS-200 dataset [25], curated by the Department of Centre for Artificial Intelligence (CAI) at the ZHAW, played a key role in dialect analysis. It included about 200 hours of audio records from approximately 4'000 different speakers, each representing one of the various Swiss dialect regions. These audio recordings were gathered from volunteer speakers across Switzerland, who contributed via a web-based tool[17] designed for this purpose. After the initial recording phase, these contributions were manually validated by other participants to ensure the accuracy and reliability of the dialect representations.

Detailed metadata associated with these recordings was documented in TSV format. This metadata included essential information about each audio clip. The thesis references key columns relevant to our study in Table 2. However, it is important to note that the TSV file contained additional columns that, while not central to the current analysis, provided supplementary data that could be useful for further research or a more detailed dataset exploration.

Column	Explanation
clip_path	path to Swiss German clip
sentence	Standard German sentence
clip_is_valid	True, False (validation of the clip)
client_id	unique speaker identifier
zipcode	zipcode of the origin municipality of a user's dialect
canton	canton of origin of a user's dialect
age	speaker's age bracket
gender	speaker's gender

Table 2: Columns in SDS-200 Metadata

STT4SG-350 Dataset In contrast, the STT4SG-350 dataset [26] offered a different structural composition and served as an even more extensive repository of Swiss dialect recordings, totaling approximately 343 hours from 316 speakers. As of its last update in 2023, it stood as the largest dataset of its kind, underscoring its significance in the field of linguistic research. This dataset was particularly geared towards applications in ASR, TTS synthesis, dialect recognition, and speaker identification.

The STT4SG-350 dataset's speakers were selected with attention to demographic and linguistic representation, ensuring approximately 50 speakers from each dialect region. Additionally, a balanced representation of genders across the dataset was achieved, reflecting a commitment to diversity in linguistic research.

This dataset was instrumental in training an ASR model, demonstrating its efficacy by achieving an impressive average BLEU score of 74.4 on its test set. This indicates a high level of accuracy in speech recognition.

[17] Was available on Dialektsammlung but not anymore.

Moreover, the metadata for the STT4SG-350 dataset was maintained in a TSV file as well, though it deviated slightly from the format used in the SDS-200 dataset. Notable changes included simplifying the clip_path to just 'path'. This TSV file also uniquely assigned the duration of the audio file and the dialect region directly to each audio clip, streamlining data handling and accessibility.
These datasets from SwissNLP[18] not only facilitated a deeper understanding of Swiss German dialects but also supported advancements in linguistic technology and dialectal studies, paving the way for future research and development in the field.

2.3.1 Exploratory Analysis

This section presents an initial exploratory analysis to get an overview of the STT4SG and SDS-200 datasets. This analysis includes the distributions of regions, age, gender, and duration of audio recordings (where metadata is available).

Region Distribution Figure 2 illustrates the distribution of records amongst the dialect regions within both the SDS-200 and the STT4SG-350 datasets. The STT dataset is well-balanced whereas the SDS dataset has far more records for the regions Zürich and Bern.

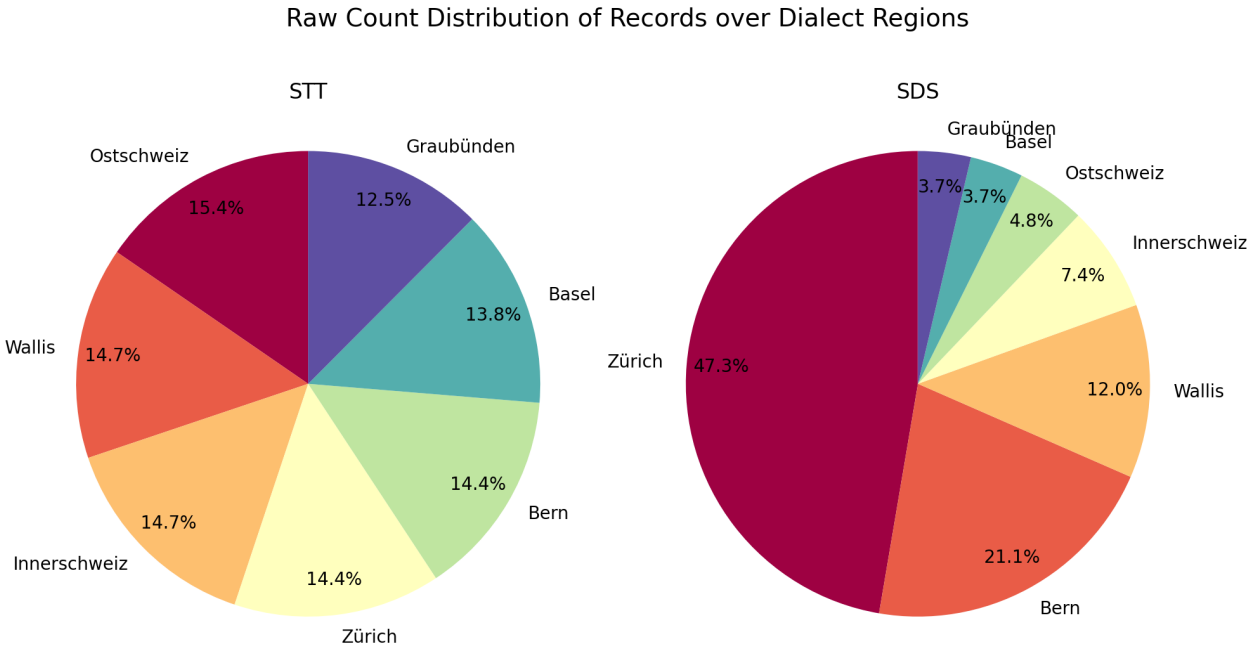

Figure 2: Record Counts amongst Dialect Regions in STT4SG-350 and SDS-200 datasets

Age Distribution Figure 3 shows the age distribution for both the STT4SG and SDS-200 datasets. The age groups are categorized as 'teens', 'twenties', 'thirties', 'forties', 'fifties', 'sixties', 'seventies', 'eighties', and 'No Info' for entries where age information is unavailable. The plots indicate that the datasets cover a wide range of age groups, with a notable amount of missing age information in the SDS-200 corpus.

[18]SwissNLP is an organization dedicated to advancing NLP, Computational Linguistics, and Text Analytics within Switzerland

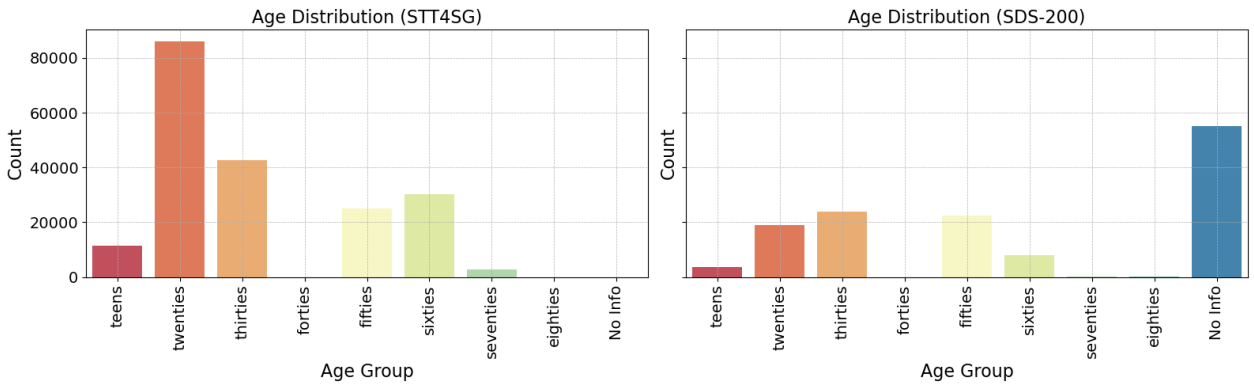

Figure 3: Age Distribution - STT4SG-350 and SDS-200

Gender Distribution Figure 4 presents the gender distribution within the STT4SG-350 and SDS-200 datasets. The categories include 'male', 'female', 'other', and 'No Info' for entries lacking gender data. This analysis reveals the proportion of male and female speakers in the corpora, highlighting the presence of a significant amount of missing gender information in the SDS-200 dataset.

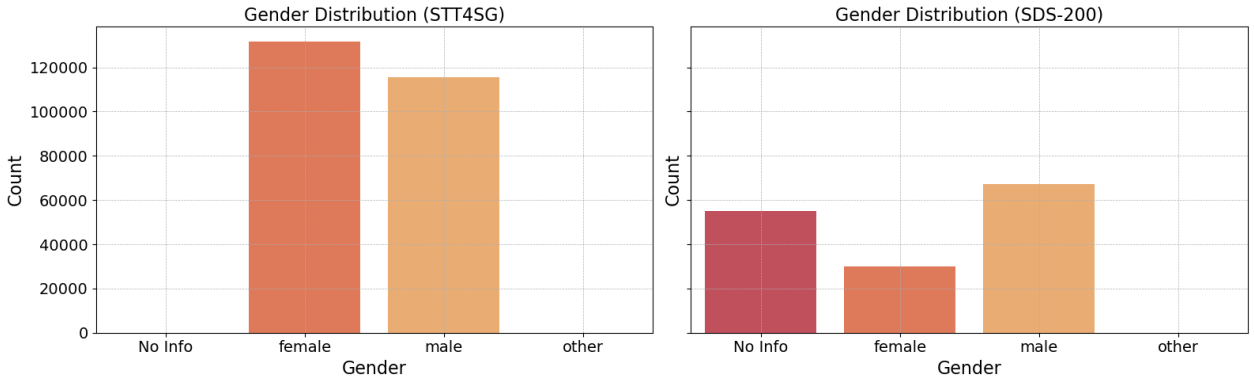

Figure 4: Gender Distribution - STT4SG-350 and SDS-200

Duration Distribution Figure 5 illustrates the distribution of audio recording lengths for the STT4SG-350 dataset. Unfortunately, the SDS-200 dataset does not include metadata for the duration of each audio recording and, hence, is not represented in this analysis. The histogram shows the spread of recording lengths, providing insight into the typical duration of recordings within the STT4SG-350 dataset.

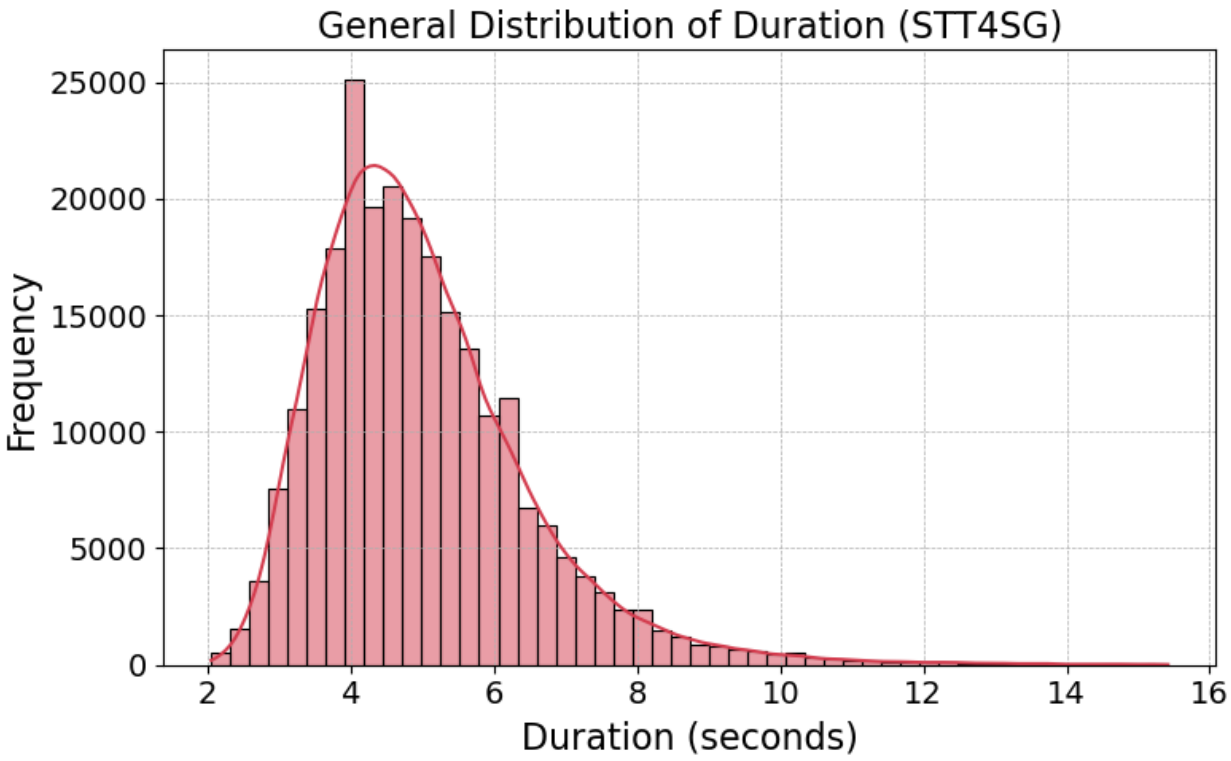

Figure 5: Audio File Duration Distribution - STT4SG-350 and SDS-200

3 Audio Clips Management

This chapter delves into the process of preparing the audio files included in the datasets for detailed analysis. The audio files were initially a mix of MP3 and FLAC formats, and our first step involved converting them to WAV format to standardize processing and ensure compatibility with our analytical tools. This conversion was essential for using the pyannote module[19]. After conversion, we removed silent parts at the beginning and end of all files to focus solely on relevant audio content. Each file was then reconverted to FLAC, facilitating further analysis, such as measuring the speech speed in words per second for each dialect, as the FLAC file format contains metadata like the clip length.

But why was the audio not converted back to MP3 or kept in WAV format? This question can be answered partly by the guidelines provided by SwissNLP and partly by the nature of WAV as an audio format. Due to their large file size, WAV files contain uncompressed raw audio data [28], making them impractical for sharing and analyzing. MP3, one of the most well-known audio formats, was used in parts of the SDS-200 dataset and the test section of the STT4SG-350. However, MP3 loses audio quality due to lossy compression [28]. Therefore, using FLAC is suitable for preserving quality while saving storage space. The data processing journey undertaken in this project was graphically depicted in the following Figure 6. This visual representation serves as a preliminary overview, setting the stage for a more detailed exploration of each phase involved in the data handling process. Subsequently, each process component was meticulously examined to provide a comprehensive understanding of the methodologies applied at every stage.

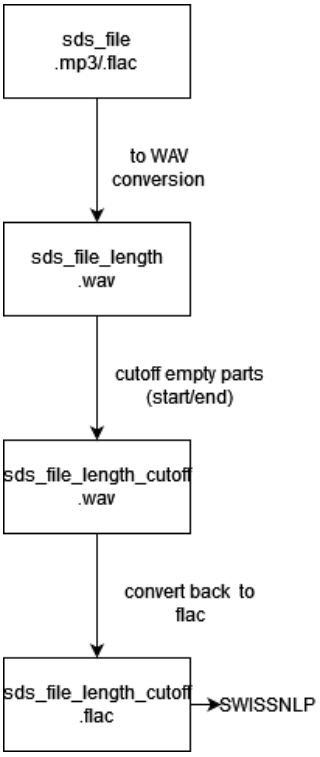

Figure 6: Audio Clips Pipeline

After data processing, all data is finally returned to SwissNLP for further analysis.

[19]Huggingface Pyannote a tool specializing in voice activity detection that supports only WAV files [27].

3.1 Audio Format Standardization

The initial phase involved converting all audio files from their original MP3/FLAC formats to WAV, a format better suited for detailed audio analysis and also the only one supported by the module for silence removal and certain STT models. For this process, we created a script designed to handle large-scale conversions efficiently without significantly losing audio quality. To achieve high efficiency, the script utilized multi-processing, enabling simultaneous processing of multiple audio files, which significantly sped up the conversion process across extensive audio libraries. Moreover, it employed the Pydub library[20], which is known for its ability to maintain the integrity of the audio quality during conversion. This ensured that the resultant WAV files retained the grade of the original recordings without significant quality loss.

3.2 Silence Removal

Since audio recordings often include silent intervals at the beginning and end, our next step was removing these segments. We developed and applied a silence removal script that automatically detected and trimmed these silent parts, streamlining the analysis of spoken content and reducing the time needed for further processing steps.

Removal with Pyannote The script began by loading an audio file into memory using the module torchaudio[21], which provided both the waveform (audio signal) and its corresponding sampling rate. Once the audio file was loaded, the script initialized a Voice Activity Detection (VAD) model using the Pyannote library, which is specifically trained to detect speech within audio files. A pre-trained model from Pyannote was employed to perform this task effectively.

After initializing the VAD model, the script applied it to the audio file. The model analyzed the waveform and sampling rate to identify segments where speech was present. Once these speech segments were detected, the script calculated their start and end times and applied a 0.5-second buffer to both the beginning and end of these segments. The buffer was added after the initial run-through because a review of random audio files revealed that spoken parts of the audio were being trimmed at both the beginning and the end. We removed the truncated files and started the process again. This buffer ensured that no speech was accidentally truncated. The buffer time was subtracted from the start and added to the end of the detected speech segments. The file was then saved with a new suffix to distinguish it from the original.

Removal with Pydub For 2'148 audio files (105 from the STT4SG-350 dataset and 2'043 from the SDS-200 dataset), Pyannote could not cut out the silence and skipped these files. Many of the audio files had to be discarded due to inadequate audio quality or the absence of sound. For these files, we used a new script that trims the audio file based on the volume at the beginning and the end with Pydub. Once the audio was loaded, the script employed a silence detection from Pydub to identify segments of the audio that were not silent. This function took parameters for the minimum length of silence to be considered, such as 1'000 milliseconds and a silence threshold, typically around 50 dB [30]. It then returned a list of start and end times for the non-silent sections identified within the audio.

If non-silent sections were found, the script calculated both the start of the first non-silent section and the end of the last one, determining the appropriate segment of the audio to trim. The audio

[20] The PyDub module is known for the audio manipulation with a high-level interface [29].
[21] PyTorch torchaudio

was then precisely trimmed to this specified section. If no non-silent sections were detected, the script retained the original audio to avoid data loss.

3.2.1 Results and Conclusion

The failure of the Pyannote tool to effectively process some files stemmed from multiple factors, as mentioned above. The incompatibility between the audio characteristics and the VAD model highlighted the importance of initial audio quality and proper alignment with the analytical tools' specifications for successful audio analysis.

Table 3 shows the trimmed time in total and on average per file. Despite the different number of files in the two data sets, the total truncated time is relatively similar. On average, records from the SDS-200 data set are truncated more.

Dataset	Total Time Trimmed (s)	Average Time Trimmed (s)	Files
STT4SG-350 - Valid	120001.139	0.54	222922
STT4SG-350 - Test	15238.74	0.62	24605
SDS-200	110110.374	0.72	152251

Table 3: Summary of Trimmed Time Data

3.3 FLAC Reconversion and Space Optimization

After the silence removal, the audio files were converted back to FLAC format. This step was for optimizing storage, as FLAC balances compression and preserving audio quality. The space gained through this compression was quantified to demonstrate the efficiency of the data management approach.

After reverting the audio files again, we determined a memory gain of 4452.81 MB (11%) for all datasets combined from the initial memory, which initially contained a combination of MP3 and FLAC files. All files were first converted to WAV, then the silence was trimmed out and finally converted back to FLAC. At first, some storage space was lost due to the conversion to WAV, which is indicated by a minus sign. The following Table 4 provides a detailed overview of the amount of storage space saved through this process.

State	Size (MB)	Space Gained (MB)	Percentage
Original Size	40861.15	-	-
Converted to WAV	82354.02	-41492.82	-51%
Trimmed silence	71386.89	-30525.74	-33%
Reconversion to FLAC	36408.34	4452.81	11%

Table 4: Summary of Gained Space

Furthermore, the reduction in data size resulting from these exclusions provided additional benefits for future project work utilizing this dataset. It allows students to download smaller datasets, easing their systems' computational and storage demands. This efficient use of storage facilitated smoother operational processes and enhanced the accessibility and manageability of the data for further academic pursuits.

4 Data Management

This chapter explores the management and enhancement of our datasets. Our primary focus was on improving the utility of the existing datasets through targeted techniques. Key to this enhancement was the addition of a 'usability' column and a 'reason' column in our datasets. These columns flagged and documented the specific reasons and phases during which files were deemed unusable, ensuring the analytical integrity of our project. Additionally, transcribing audio recordings into text was an important component of our data management strategy. This step resulted in enriched TSV files with a 'transcription' column. These enhancements facilitated a more comprehensive analysis and supported the goal of analyzing dialectical variations. By examining the detailed transcriptions, we aimed to identify and understand the linguistic features present in the records while ensuring that all data used was accurately vetted and recorded for transparency.

Figure 7 illustrates our entire Data Management Flow, the different steps undertaken were documented in detail in the following chapters. The data from both datasets was analyzed in three stages. In the first stage, the files were checked for silent audio files or those with an exceptionally high speech rate (>4 WPS). The selected values will be discussed in detail later. In the second stage, sentences were transcribed, and any sentences or audio files that could not be transcribed using either the NVIDIA model or the Whisper model were filtered out. The final stage focused on the preterite tense, which, as previously mentioned, does not exist in Swiss German. All sentences written in the preterite tense were marked to prevent distortion of the automated transcription results. These three stages of the project are explained in detail below.

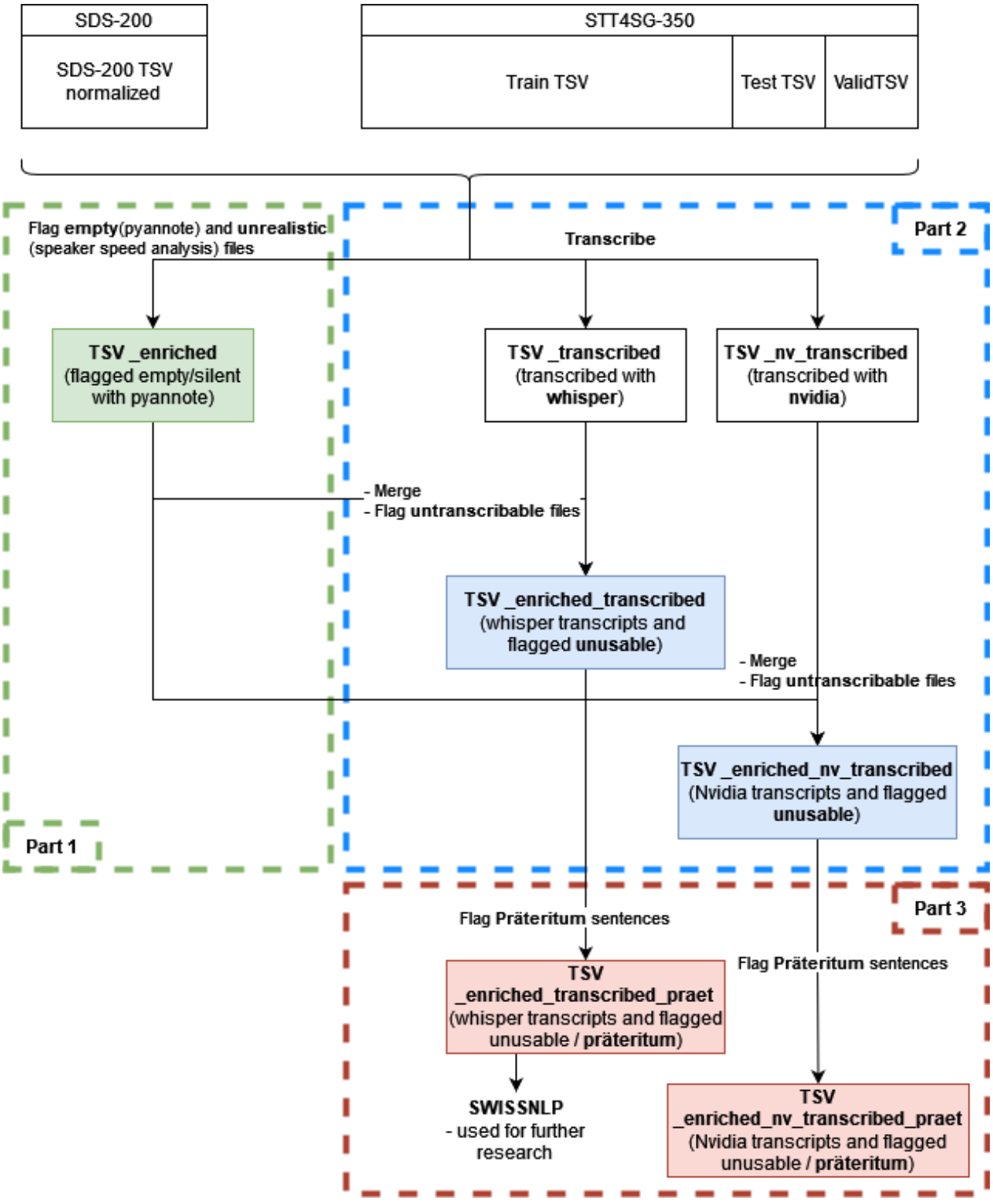

Figure 7: General Data Management Flow

4.1 Standardization

In the initial phase of our data management process, the primary task was to enable effective comparison across different dialect regions. To achieve this, we focused on normalizing the SDS-200 dataset, which originally lacked specific information about dialect regions as can be found in the STT4SG-350 dataset but contained information about canton and postal codes per record. Our approach enriched the dataset by dynamically mapping postal codes to their corresponding dialect regions. This mapping was computed dynamically by utilizing data from the STT4SG-350 dataset, which provided detailed insights into the dialect region associated with each record. Additionally, the code used to create this map was adapted from a previous project conducted by ZHAW students Claudio Frei and Philippe Schneider [31]. This step established a framework for subsequent analyses, ensuring that the data could be accurately categorized and compared based on regional dialectical variations.

4.2 Silent Files

Throughout our research's data and speed analysis phases, multiple cleansing scripts were used to ensure the integrity and relevance of the audio files examined. The first step in this process involved identifying and marking all silent files from the datasets, as they were unsuitable for accurate speech analysis. The detailed outcomes of this initial cleansing effort are documented in Table 5. The SDS-200 dataset produced results approximately eight times higher than the STT4SG-350 dataset. The table also shows that nearly 1'000 files were identified. Each file was manually checked using a playlist script that played all the detected audio files. If a file was found to be usable (non-silent), it was manually removed from the list of silent files. This review process ensured the accuracy and reliability of our dataset for further analysis.

Dataset	Number of Non-Usable Records	Reason
STT4SG-350 Valid/Train	98	silent file
STT4SG-350 Test	2	silent file
SDS-200	841	silent file

Table 5: Detection of Silent Files

4.3 Unrealistic Files

A subsequent round of data cleansing was initiated upon discovering that certain audio files exhibited unrealistically high speech speeds, indicating potential anomalies or errors in recording. These files were manually reviewed to determine the nature of the audio content, leading to the removal of those containing irrelevant content such as motor noise, environmental sounds, no speech, or improperly truncated speech, which could skew the analysis results. The specifics of this targeted removal are recorded in the second table below, Table 6. As with the previous results, the SDS-200 dataset again showed a higher detection rate than the STT4SG-350 dataset.

Dataset	Number of Non-Usable Records	Reason
STT4SG-350 Valid/Train	21	unrealistic speed (WPS)
STT4SG-350 Test	1	unrealistic speed (WPS)
SDS-200	555	unrealistic speed (WPS)

Table 6: Detection Files with Unrealistic High Speech Speed

With 'unrealistically high', all speakers who spoke faster than four words per second on average were referred to. This issue was addressed in great detail in the report by Dr. Amsel and F. W. Kaeding [32]. The working committee based their analysis on 20 million syllables, which required 10'910'777 words for easier calculation. According to the results presented below, the average word in Standard German consists of 1.83 syllables.

In order to calculate the words per second from syllables per second, formula 1 was used:

$$\text{WPS} = \frac{Syllables\ per\ second}{Average\ number\ of\ syllables\ per\ word} \tag{1a}$$

According to Heinrich Heine University in Düsseldorf [33], the average number of syllables per second ranged from four to six. Similarly, the Neue Zürcher Zeitung (NZZ) [34] reported comparable results of 5 to 6 syllables per second. Utilizing these figures, the average speech rate was calculated to range from 2.19 words per second to 3.28 words per second, indicating that four words per second would be quite high. Calculating speech speeds allowed us to benchmark our datasets against established academic and media-reported norms. This provided a basis for further comparative analysis of dialect variations among Swiss German speakers.

Following the detection and manual verification of these files, modifications were made to the metadata within the corresponding TSV files. Entries for the problematic audio files were marked with 'clip_is_usable' set to 'False', along with a 'drop_reason' detailing whether the exclusion was due to 'unrealistic speed (WPS)' or 'silent file'. These two parts of the Data Management pipeline are marked as Part 1 (green) in Figure 7.

4.4 Transcription

After the initial processing of the datasets to identify records with silent files and unrealistic speed values, the next step in our pipeline was transcription. The transcription process was designed to efficiently manage large datasets, adhere to API limitations, and maintain data integrity during unexpected disruptions.

Transcription Methodology We employed two advanced STT models for our transcription tasks: the locally run NVIDIA FastConformer-Hybrid Large (de) model [35] and the Whisper large-v2 model [36] accessed via the OpenAI API. Detailed information about the model research and choice will be presented in Chapter 6.2.

The transcription was executed over several days, necessitating the development of a robust pipeline to manage the high volume of data and the rate limitations imposed by the OpenAI API. We ran the transcription on a remote server continuously, including safeguards such as data backups for network issues, system overloads, or untranscribable files. To ensure completeness, any files that initially returned empty transcripts were re-processed to confirm that no transcription errors occurred unless the audio file was genuinely empty.

Integration and Flagging Following transcription, the newly created TSV files containing a new transcription column for the respective transcription model were merged with the TSVs previously enriched with information about unrealistic and silent files. This merging process retained all existing data, including the new 'usability' columns, while integrating the fresh transcription data. It is marked as Part 2 (blue) in Figure 7. The NVIDIA and Whisper models were employed to maximize accuracy and mitigate model-specific biases. This dual-model approach allowed for cross-verification, flagging files with empty transcripts from either model as unusable. Table 7 shows the number of records identified in this step.

Dataset	Number of Non-Usable Records	Reason
STT Train	142	Non-Transcribable
STT Valid	7	Non-Transcribable
STT Test	2	Non-Transcribable
SDS	156	Non-Transcribable

Table 7: Non-Usable Records Due to Non-Transcribability

To ensure the integrity of the datasets, each flagged file was manually reviewed through listening tests. This manual verification confirmed that no potentially usable files were incorrectly marked as unusable due to transcription errors. No more than three records were considered valid after manual investigation and were therefore kept and not included in the non-transcribable records.

These steps enabled us to uncover and analyze instances of transcription 'hallucinations', such as biased terms produced by the Whisper model in ambiguous or inaudible clips, further enhancing the reliability of the datasets. These 'hallucinations' or biased terms refer to anomalies where the transcription model, such as Whisper, erroneously generates text based on its training data rather than the actual audio content. This is particularly noticeable in silent or unclear audio portions, where the model might output phrases like 'Translated by Amara.org Community' — a line often seen in subtitled films, suggesting the model's reliance on film-based training materials. Additional text lines that were translated can be found on GitHub [37]. This phenomenon underscores the importance of thorough manual checks to ensure data integrity.

There appears to be a potential solution for these hallucinations, which involves adjusting certain parameters according to the user 'KaiserChr' on Github [38]. Due to time limitations, this was not further investigated nor implemented in our transcription process.

Final Outputs We identified 307 non-transcribable records across both datasets, refining our datasets for subsequent analytical tasks.

4.5 Preterite Tense

The Preterite Tense does not exist in any of the Swiss German dialects, as mentioned in Chapter 2.2.1.

This significant linguistic difference affects any analysis based on word matching for examining dialect variability. To address this, we implemented the third part (as marked in 7) of our Data Pipeline: flagging all sentences within the datasets to identify if they are written in the preterite tense, thus enabling the filtering of such entries.

Method The past tense component of the pipeline iterates through all elements of both the SDS-200 and STT4SG-350 corpora. It utilizes the Python library SpaCy[22] to determine the tense of each sentence. An additional field is appended to the enriched datasets, indicating whether an entry is written in the past tense.

SpaCy processes each word in a sentence to ascertain its tense. The sentence is marked as past tense if any word is in the past tense. SpaCy employs a morphological tagger, a tool that analyzes and identifies the morphological features of words, such as tense, number, gender, and case. The tagger relies on a combination of rules and statistical models derived from its training data. Table 8 lists the different datasets and the number of records with sentences written in the preterite tense, detected by the SpaCy library.

Dataset	Number of Past Tense Records	Total Records	Percentage
STT4SG-350 Train	62218	199705	31.15%
STT4SG-350 Valid	7051	23217	30.37%
STT4SG-350 Test	5817	24605	23.64%
SDS-200	48579	152251	31.91%

Table 8: Overview Records in the Preterite Tense

Final Outputs The culmination of this process is the output of the final TSV files, which are both enriched and transcribed. These files are flagged to denote unusable records and records written in the preterite tense.

4.6 Overview

Figure 8 gives an overview of the entire Data Management pipeline. The first bar represents the total records per dataset, followed by bars indicating the number of silent, unrealistic, non-transcribable, and preterite records. The last bar represents the remaining records from the SDS-200 dataset initially flagged as invalid during the dataset creation. To enrich the dataset and provide more details on individual records and the reasons for their invalidity, the detection pipeline ran over all records of the SDS-200 corpus. This means there is an overlap between the initially invalid files and our added invalid categories. In this graph, invalid records are counted towards our specific categories and the 'Invalid Clip' category[23], which had already been included in the dataset when we obtained it.

[22] SpaCy is an open-source library for Natural Language Processing in Python [39]

[23] clip _is _valid is a column in the TSV file, which indicates whether a clip is a correct Swiss German representation, whether it is not, or if there were too few votes. Other participants manually created this verification

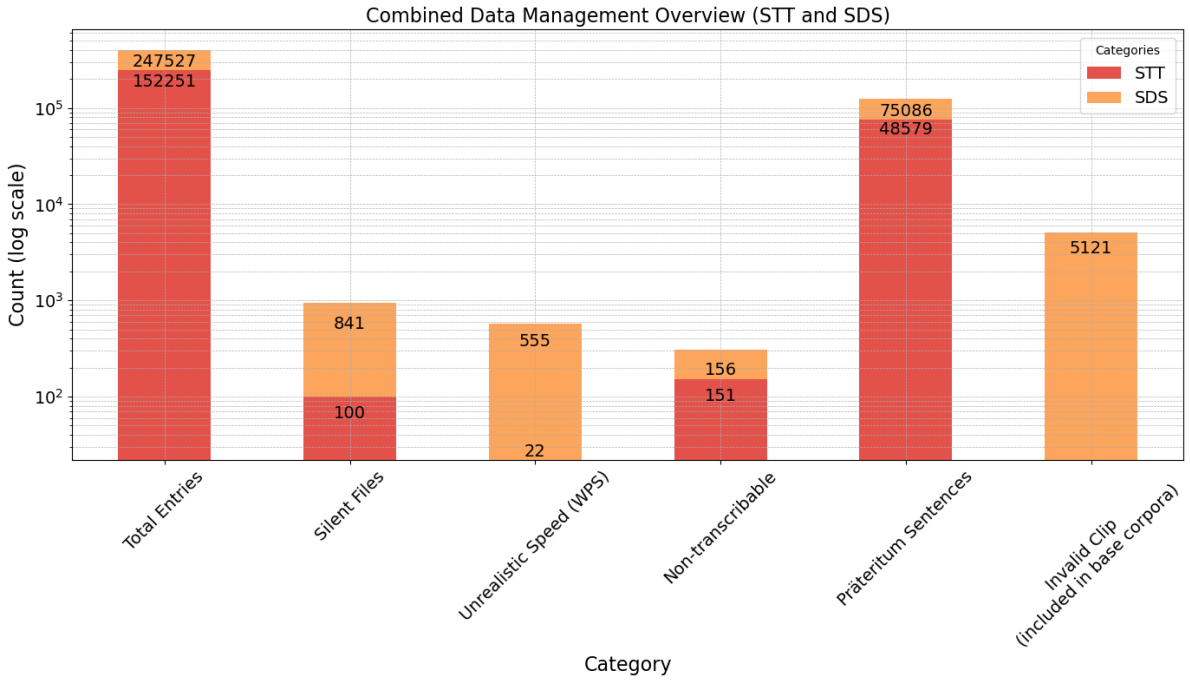

Figure 8: Data Management Overview

Figure 9 and 10 illustrate the relative distribution of invalid records along with their drop reason amongst the dialect regions for both datasets. The counts of invalid records are divided by the total amount of records within the region, representing the relative percentage score shown on the Y axis. Bern seems to have by far the most invalid records in the STT4SG-350 dataset relative to its amount of records. In SDS-200, most regions have a similar relative amount of invalid records, except Zürich and Wallis, with a notable low score despite representing large amounts of records within the dataset. It is also evident that Bern has a significantly high number of failed transcriptions but no specific explanation could be identified for why Bern exhibits the highest rate of failed transcriptions. All recordings detected as invalid in our pipeline were manually reviewed and deemed unusable.

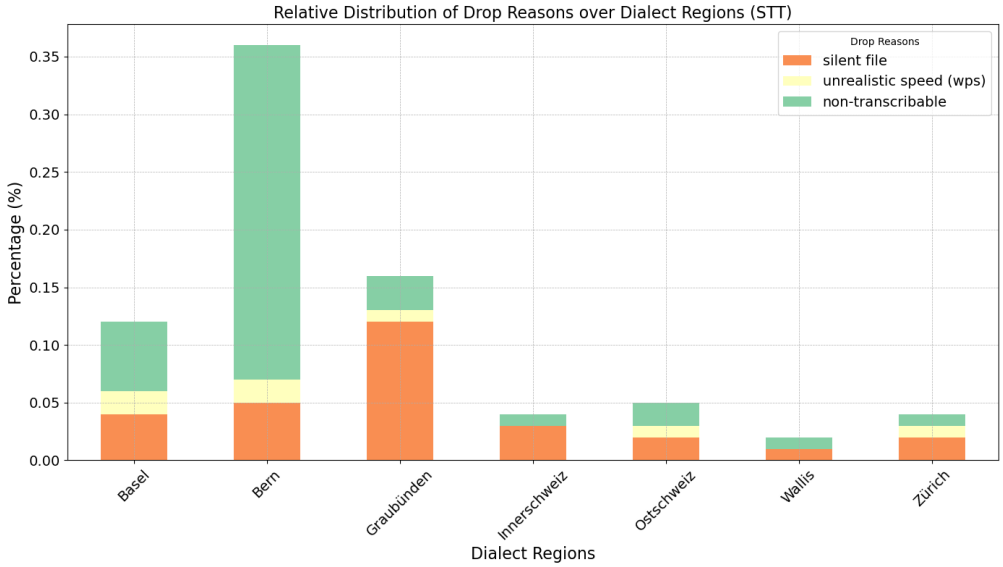

Figure 9: Data Management Drop Overview - STT4SG-350

Within the SDS-200 dataset displayed in Figure 10, a noticeable difference towards the STT dataset can be observed in the drop reason distribution: Many files were deemed unusable due to unrealistic speech speeds or being completely silent. When comparing these figures, the scale on the left of the chart reflects that the SDS-200 dataset generally has a higher relative amount of dropped records across all regions when compared to the STT4SG-350 dataset.

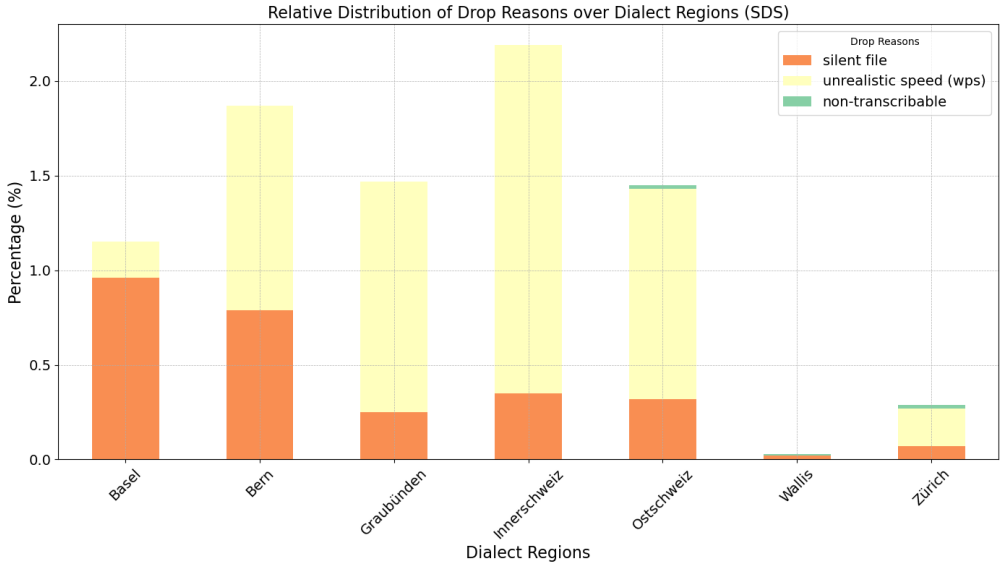

Figure 10: Data Management Drop Overview - SDS-200

The methodologies and outcomes described in this chapter are to enhance the reliability and analytical depth of our dialect research. They provide a solid foundation for accurately interpreting and understanding linguistic variations across regions.

5 Speed Analysis

This chapter presents a comprehensive analysis of speech speed across various Swiss dialects. Audio processing techniques were employed on the two datasets to determine the fastest Swiss dialect and investigate the hypothesis that Bernese speech is slower than Zürich's.

The core of this part, the speed analysis, involved calculating WPS for each record and mapping it to the respective regions to identify and compare the speeds between the dialects. This metric provided a direct measure of speech tempo across different dialects. The investigative phase of this study was dedicated to analyzing speech speeds, utilizing the advanced capabilities of the librosa audio processing library[24]. The results of the analysis of the trimmed audio files[25] are displayed in the three tables 9, 10 and 11. Each table illustrates specific aspects of the datasets: the first table focuses on the combined findings from both datasets to provide a holistic view. The second table focuses on the SDS-200 dataset, and the third on the STT4SG-350 dataset.

5.1 Speed Analysis calculated by Files

First, the combined approach calculated the WPS on each audio file for both datasets. The speech speed of each recording was determined using the formula 2 - Standard German words divided by time. This calculation was then averaged for each region to obtain the regional average speech speed.

$$\text{WPS} = \frac{\textit{Standard German Words (in base sentence)}}{\textit{Recording Duration}} \tag{2a}$$

This stage of the research brought to light several findings. Most notably, it was observed that the Bernese dialect emerged as the fastest among all the Swiss dialects studied. Given the preconceived notions about regional speech tempos, this was an unexpected outcome. The subsequent table, referenced as Table 9, presents the average words per second (WPS) for each dialect region, arranged in ascending order from the fastest to the slowest. This graphical representation not only highlights the relative speech speeds but also visually affirms the quantitative data derived from the analysis.

Dialect Region	WPS
Bern	2.12
Zürich	2.10
Ostschweiz	2.02
Innerschweiz	2.01
Wallis	2.01
Basel	1.97
Graubünden	1.96

Table 9: A First Speed Analysis over Both Datasets

Table 9 presents a snapshot of speech speeds combining both the STT4SG-350 and SDS-200 datasets, revealing variations in WPS among different Swiss dialect regions. It shows that Bern is the fastest with 2.12 WPS, and Graubünden is the slowest at 1.96 WPS.

To delineate differences between the datasets, further analyses were performed for each dataset separately. The following Table 10 includes the speed analysis for the test data partition from the

[24] Librosa is a package for Python to analyze audio [40]

[25] Audio file trimming was conducted in Chapter 3.2.

STT4SG-350 dataset. The results closely resemble the results of the Zuber and Pfister Paper shown in Chapter 2.2.1, as the same partition of the dataset is used.

Dialect Region	WPS
Wallis	1.77
Graubünden	1.72
Basel	1.67
Innerschweiz	1.63
Zürich	1.61
Bern	1.57
Ostschweiz	1.54

Table 10: Speed Analysis with Test Dataset of STT4SG-350

Subsequently, the speed analysis of the entire STT4SG-350 and SDS-200 datasets was conducted, as shown in Table 11.

Dialect Region	WPS
Bern	2.06
Ostschweiz	2.01
Zürich	2.00
Innerschweiz	2.00
Basel	1.97
Wallis	1.95
Graubünden	1.94

(a) Dataset STT4SG-350

Dialect Region	WPS
Wallis	2.20
Zürich	2.11
Bern	1.93
Basel	1.92
Graubünden	1.90
Ostschweiz	1.90
Innerschweiz	1.83

(b) Dataset SDS-200

Table 11: Speed Analysis with Both Datasets

In the ranking resulting from the entire SST4SG-350 dataset (shown in Table 11a) we have different outcomes than in the previous analyses, with Bern again among the faster regions. According to several sources, Wallis presumably is the fastest-speaking region, as has been discussed in Chapter 2.2.1. Right next to it in Table 11b displaying results from the entire SDS-200 dataset, there are different results again, with Wallis as the fastest-speaking region and Innerschweiz as the slowest.

The three tables from the analysis demonstrated varying speech rates across Swiss dialect regions, which did not correspond to the widely reported figures in various media and scholarly articles. Notably, the news platform '20 Minuten' highlighted the linguistic pace of different regions [41], claiming that Wallis had the fastest speakers, while Bern was characterized by a slower speech pace. According to '20 Minuten', the speech rates were quantified as 257 syllables per minute for Bern, 263 for Zürich, and 270 for Wallis. When converted using the Formula 1 used in Chapter 4.3, these figures suggested speech speeds of approximately 2.34 WPS for Bern, 2.4 WPS for Zürich, and 2.46 WPS for Wallis. This discrepancy between expected and observed values prompted a deeper investigation into the potential factors influencing these differences.

To provide a clearer understanding of the speech speed variances observed, the composition of the speaker samples from both datasets was examined. The compilation of data regarding the number of speakers per dialect region was thus undertaken, as outlined in the following Table 12, which presents the distribution of speakers across the dialect regions involved in the study.

Dialect Region	Speaker STT4SG-350	Speaker SDS-200
Zürich	46	340
Bern	46	309
Innerschweiz	43	135
Ostschweiz	47	119
Basel	44	59
Graubünden	46	54
Wallis	44	29

Table 12: Amount Speaker per Region in Both Datasets

The table shows that the STT4SG-350 dataset is relatively balanced, but the SDS-200 dataset is not. This could also have affected the results of the speed analysis if individual speakers have a greater influence on speed than other dialect regions.

5.2 SDS-200 Sampling

Due to the SDS-200 dataset being unbalanced regarding the distribution of speakers across dialect regions and the number of recordings uploaded by each speaker, additional processing was conducted to sample the dataset. It was observed that a few speakers had uploaded over 1'000 recordings each, while others had fewer than 10.

All speakers with fewer than ten records were excluded from this analysis to sample the dataset. Furthermore, for speakers who had made more than 50 records, only 50 random records were selected for inclusion in the analysis. This approach was implemented to ensure that individual speakers cannot excessively influence the result, thereby reducing potential biases arising from overrepresentation or underrepresentation in the dataset. The outcomes of this refined analysis are visualized and presented below in Table 13.

Dialect Region	WPS
Zürich	1.95
Ostschweiz	1.93
Bern	1.90
Innerschweiz	1.90
Wallis	1.90
Basel	1.90
Graubünden	1.88

Table 13: Sampled Speed Analysis of the SDS-200 Dataset

With only a difference of 0.07 WPS from the fastest to the slowest region, the results reveal minor differences between the dialect regions. This minor WPS variance strongly suggests that the sampled dataset is unsuitable for such an analysis or that an error was made during the analysis. After carefully reflecting on the analysis, we suspected the first one to be more likely. Since we sampled the data to achieve equal representation, we also lost a lot of valuable data, which could have skewed the result.

5.3 Speed Analysis Calculated by Average Speaker Speed

To address the possible skewed results stemming from the unequal representation of audio files in the datasets, an average speech speed was initially calculated for each speaker to obtain more precise analysis and avoid dependency on the number of speaker recordings. With this method, it is unnecessary to sample the dataset as described in Chapter 5.2. Subsequently, an average was computed across the regions. The following Figure 11 displays the distribution of speech speeds for individual speakers. By averaging speeds at the speaker level before regional aggregation, the analysis mitigated the influence of speakers who might disproportionately affect the results due to either a high or low number of contributions.

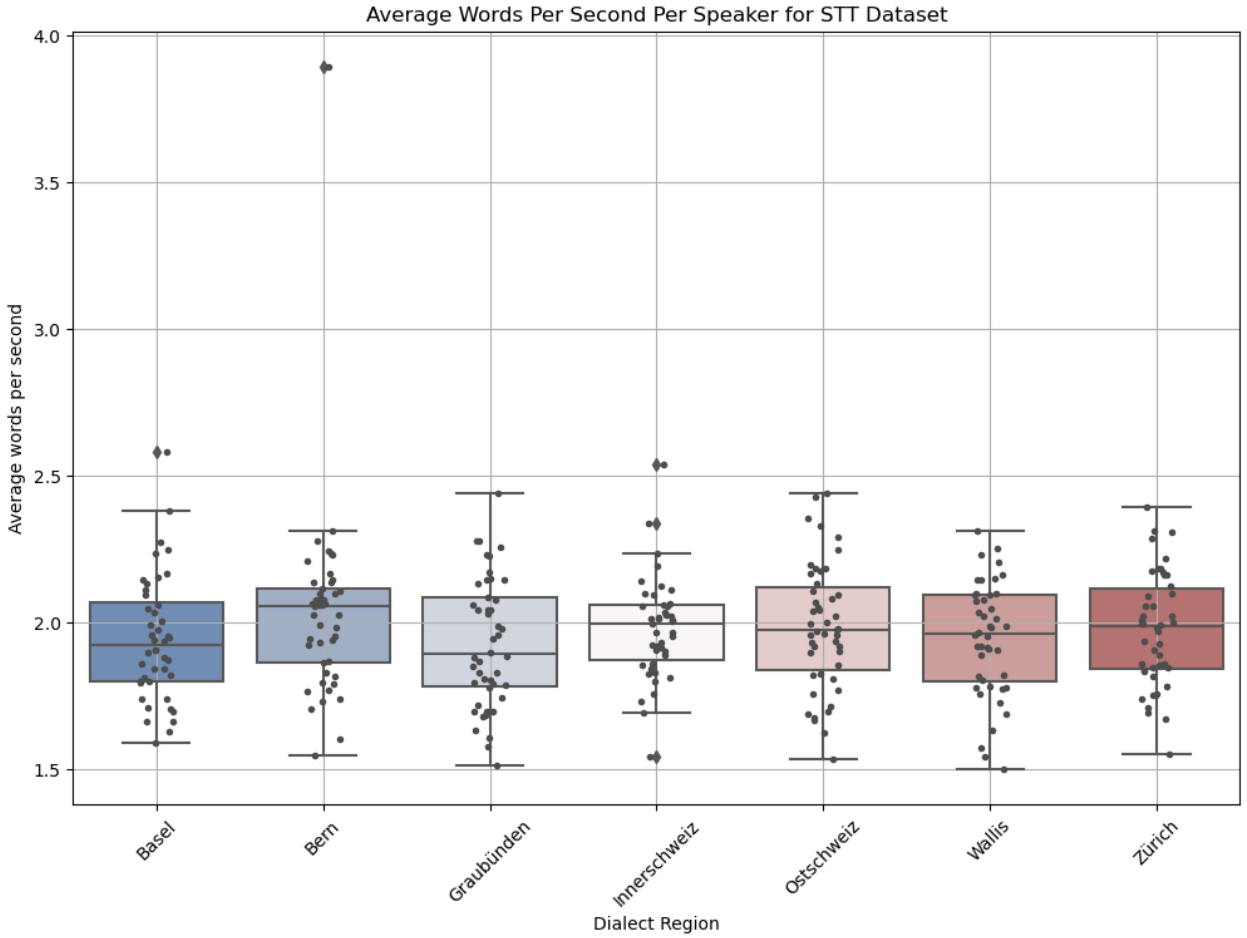

Figure 11: Average WPS per Speaker for STT4SG-350 Dataset

Each point in the boxplot represents a speaker. On the Y-axis are the speaking rates in WPS, and on the X-axis are the dialect regions. The same distribution analysis was also conducted for the SDS-200 dataset, as depicted in Figure 12 below.

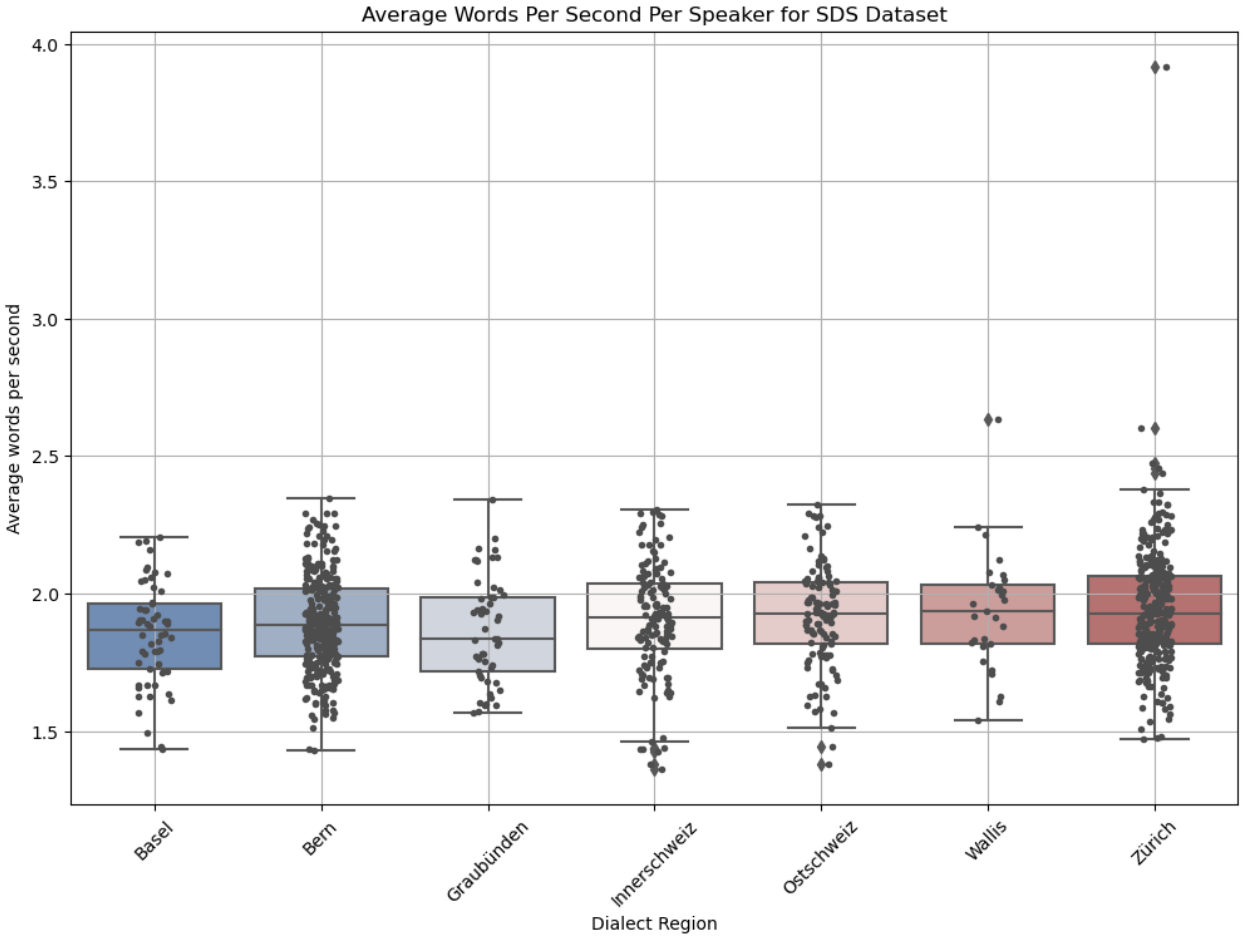

Figure 12: Average WPS per Speaker for SDS-200 Dataset

These two graphs show the distribution of the speaking speeds of the individual speakers within the region. This serves to provide an overview. The outliers with more than 3.5 WPS were already manually checked and kept as valid records, as mentioned in Chapter 4.3. Outliers with less than one WPS were not detected in any dataset.

Histograms were created for each region and dataset to gain a more precise overview of the regional speech speeds. Figure 13 exemplifies this approach with a histogram representing the Bern region with data from the STT4SG-350 dataset. The complete set of histograms for all other regions and datasets can be found in the appendix in Chapter 9.3. This graphical representation allows for visual analysis of the distribution of speech speeds, highlighting variations within each dialect region and providing a comprehensive view of the data's structure and tendencies.

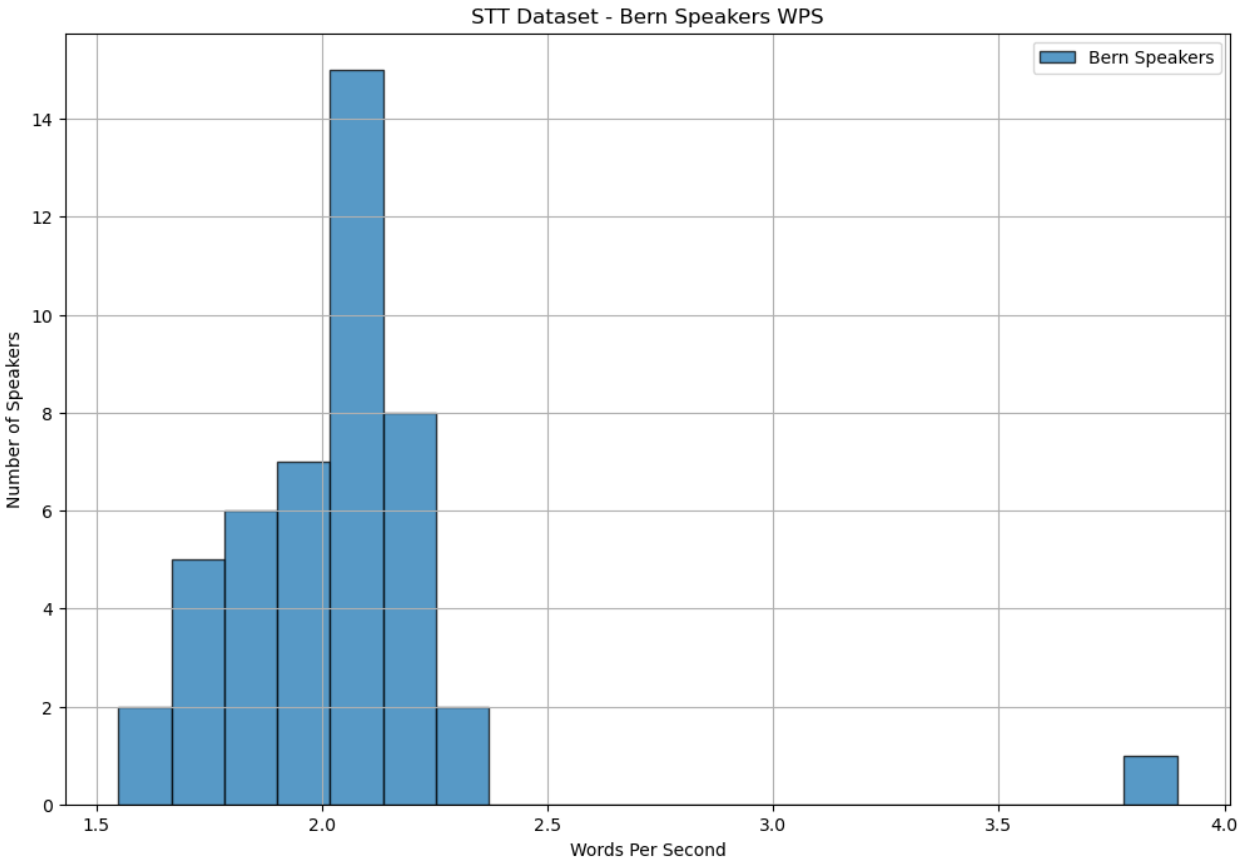

Figure 13: Histogram STT4SG-350 Region Bern

Accompanying the distribution analysis, speed analyses were conducted to assess the average speech rates across different dialect regions. The results of this analysis are presented in Table 14, showcasing the average WPS for each dialect region based on the mean values: First, a mean WPS was calculated for each speaker, followed by a mean over each dialect region.

Dialect Region	WPS
Zürich	2.07
Bern	2.02
Wallis	2.01
Ostschweiz	2.00
Innerschweiz	1.97
Basel	1.97
Graubünden	1.94

Table 14: Speed Analysis over both Datasets with Mean

As with previous analyses, the results were typically around approximately two WPS. Another method to represent the results is by using the median of speaker means. Table 15 below represents the values by first calculating the average speed for individual speakers and then computing the medians across the dialect regions. This median approach offered an alternative perspective,

potentially smoothing out outliers and providing a different view of the central tendency within the speech data of each region.

Dialect Region	WPS
Zürich	2.04
Wallis	1.99
Ostschweiz	1.97
Bern	1.95
Innerschweiz	1.94
Basel	1.93
Graubünden	1.89

Table 15: Speed Analysis over both Datasets with Median

This variation facilitated a detailed examination of regional speech dynamics, considering the diversity within the speaker data. With this table, we again obtained a new result, which provides similar values to the previous variants. However, the distribution differs from previous results.

5.4 Results

No definitive conclusion could be reached after a detailed analysis of speech speeds, as each approach yielded different results. This subsection delves deeper into the potential reasons for these discrepancies. The results are summarized in Table 16 below for each dataset and computation variant, highlighting the variance observed in the analyses.

Dialect Region	Avg. WPS for each File		WPS with Mean[1]		WPS with Median[2]	
	STT4SG-350	SDS-200	STT4SG-350	SDS-200	STT4SG-350	SDS-200
Zürich	2.00	2.11	2.00	2.11	1.97	2.10
Bern	2.06	1.93	2.06	1.93	1.97	1.91
Wallis	1.95	2.20	1.95	2.20	1.93	2.17
Ostschweiz	2.01	1.90	2.01	1.90	1.98	1.88
Innerschweiz	2.00	1.83	2.00	1.83	1.96	1.81
Basel	1.97	1.92	1.97	1.92	1.94	1.89
Graubünden	1.94	1.94	1.94	1.90	1.90	1.88

[1] An average WPS value was calculated for each speaker, followed by the computation of the mean WPS across speakers of each dialect region.

[2] An average WPS value was calculated for each speaker, followed by the computation of the median WPS across speakers of each dialect region.

Table 16: Summary of the Speed Analysis for Both Datasets

Interpretation of the Results: This examination aims to understand the underlying factors contributing to the inconsistent outcomes, such as data quality variations, analytical methodologies differences, or the inherent diversity in speech patterns across dialect regions. One straightforward conclusion for the imbalances across the datasets and results with different methods might be that the prevailing assumption itself is flawed, and the analysis has effectively demonstrated the contrary—that Bern isn't the slowest-speaking region. It is also possible that there is no significant,

consistent, and measurable difference in speech speed among the regions. However, further exploration would be needed to assert this with higher confidence. This could involve extending the research to include additional datasets or comparing the results with other studies that have performed dialect analyses across Switzerland. Such comparative studies could reinforce these findings or highlight specific variables that might have influenced these results.

Another aspect considered was the methodology employed in measuring speech speed. The speech rate was calculated in WPS, a common approach. However, alternative metrics such as CPS, syllable analysis, as discussed by Hartmut R. Pfitzinger [42], and sounds per syllable (or second), as noted by Christian Gebhard in his dissertation [43], were identified as potential avenues for providing different insights. Although, it can be expected that these measurements would correlate with the WPS measurements. Additionally, analyzing the duration of pauses between words is a method that could offer a more detailed understanding of dialectal tempo and rhythm, potentially uncovering variations that the WPS metric might not capture.

The reliability of the records used in the analysis also warranted scrutiny. The recordings, voluntarily provided by speakers, represented a small sample size within each dialect region in the STT4SG-350 dataset—approximately 40-50 individuals. This limited representation might not have adequately captured each region's full diversity and range of speech patterns. Moreover, there is reason to question whether the recordings accurately reflected natural speech or merely reading speed. If the latter is true, this could have skewed results towards uniformity across regions since reading speed doesn't necessarily equal the natural speaking speed. This could explain the observed minimal variance—approximately, on average, 0.2 WPS difference. The variance is shown for every analysis variant in Table 17.

Dialect Region	Avg. WPS for each File		WPS with Mean[1]		WPS with Median[2]	
	STT4SG-350	SDS-200	STT4SG-350	SDS-200	STT4SG-350	SDS-200
Fastest Region	2.00	2.20	2.06	2.20	1.98	2.17
Slowest Region	1.94	1.94	1.94	1.83	1.90	1.81
Difference	0.06	0.26	0.12	0.37	0.08	0.36

[1] An average WPS value was calculated for each speaker, followed by the computation of the mean WPS across speakers of each dialect region.

[2] An average WPS value was calculated for each speaker, followed by the computation of the median WPS across speakers of each dialect region.

Table 17: Summary of the Variance in the Speed Analysis for Both Datasets

When examining these results more closely, a pattern emerges: the results from the SDS-200 dataset show greater variation than those from the STT4SG-350 dataset. This suggests that the STT4SG-350 dataset may more accurately reflect reading speed, whereas the SDS-200 dataset requires closer inspection for detailed analysis.

When considering the analyses with mean and median of the average speech speed of speakers per region, based only on the SDS-200 dataset, we obtain results that align closely with the findings from sources such as the University of Bern [44]. The rankings for these analysis variants are displayed in Table 18. Both variants keep the same ranking, independently of mean or median.

Dialect Region	WPS with Mean (SDS-200)	WPS with Median (SDS-200)
Wallis	2.20	2.17
Zürich	2.11	2.10
Bern	1.93	1.91
Basel	1.92	1.89
Graubünden	1.90	1.88
Ostschweiz	1.90	1.88
Innerschweiz	1.83	1.81

Table 18: Summary of the Speed Analysis for SDS-200 Dataset based on speaker averages

This analysis indicates that the Wallis region has the fastest speakers, with a WPS lead of 0.09 when considering the mean and 0.07 considering the median over Zürich. Bern lags further behind with a gap of 0.18 WPS. The difference between Bern and Graubünden is minimal, whereas Innerschweiz shows a slightly larger gap. A factor to keep in mind is that Bern, having relatively more speakers (around 21%) in the SDS-200 dataset, achieves a more accurate WPS measurement, whereas regions like Graubünden, with only 3% to 7% of the speakers, have too few records or speaker variance to provide a precise result. This speaker distribution was outlined in detail in Chapter 4.6.
In conclusion, our study's results challenged the traditional assumption that Zürich speaks faster than Bern. With the data at hand and the specific analysis chosen, it was not possible to confirm this hypothesis. However, we obtained partially reliable results when using only the SDS-200 dataset. For the STT4SG-350 dataset, we cannot draw any conclusions due to its low variance among the dialect regions. A broader and more detailed examination, incorporating various speech metrics and larger, more diverse samples, might be necessary to draw definitive conclusions about regional speech speeds in Switzerland.

6 Linguistic Analysis

This chapter aims to understand and analyze Swiss German dialects by processing them through Standard German STT models. General concepts and metrics developed in the study are first explored to facilitate a comprehensive understanding of the subsequent analyses. Various STT model choices are then examined, discussing their advantages and disadvantages, including a thorough investigation of their performance and suitability for the projects' specific needs.

The development and the results of the dialect variation analysis will be presented. This section focuses on identifying regional differences compared to Standard German and developing a method to compute and detect dialect-specific words. This analysis highlights the unique characteristics of Swiss German dialects and their divergence from Standard German.

6.1 General Concepts

This section will explain the concepts and measurements we used specifically for this project. The glossary in Chapter 8.4 explains more general concepts.

6.1.1 Similarity Matching

All the analyses are based on matching words between the base sentence and the transcripts. This can be done using an exact match procedure, where word pairs are considered matches only if they are 100% identical. However, word forms can have minor variations that still represent the same meaning, especially when dealing with automated transcription models and heavily accented speech. Therefore, a method was implemented to match words based on their similarity, allowing the adjustment of similarity thresholds to enhance the experiments.

Levenshtein Distance: The Levenshtein distance between two strings is defined as the minimum number of single-character edit operations required to transform one string into the other. These operations include insertions, deletions, or substitutions [45]. This metric helps quantify how similar or different the strings are, with a lower distance indicating higher similarity. More information about Levenshtein Distance can be found in the Glossary (8.4).

Benefits of Levenshtein Distance:

- **Precision in Differences**: Levenshtein distance provides a precise measure of how different two strings are by counting the exact number of edits needed.
- **Simplicity**: It is straightforward to implement and understand.

TheFuzz Ratio Calculation The `thefuzz` library [46] extends the basic Levenshtein distance by scaling it relative to the strings' length, allowing for a normalized similarity score. The similarity ratio R is calculated as follows:

$$R(s_1, s_2) = \left(1 - \frac{d_{lev}(s_1, s_2)}{\max(\text{len}(s_1), \text{len}(s_2))}\right) \times 100 \tag{3}$$

where $d_{lev}(s_1, s_2)$ is the Levenshtein distance, and $\text{len}(s_1)$ and $\text{len}(s_2)$ are the lengths of the strings s_1 and s_2.

Benefits of TheFuzz Ratio:

- **Normalized Similarity**: The ratio is scaled between 0 and 100, making it easier and more intuitive to interpret as a percentage similarity.
- **Length Scaling**: By considering the length of the strings, the ratio provides a more balanced similarity score, especially for strings of different lengths.
- **Flexibility**: It allows for partial and token-based matching[26], enhancing its applicability in various contexts.

Decision: TheFuzz Ratio: While Levenshtein distance offers precise differences between strings, the `thefuzz` ratio provides a normalized measure that is more intuitive to interpret. The normalization of the string length ensures that the similarity score is relative, making it more meaningful in the case of different-length strings. Because of the benefits mentioned, we decided to base the word comparisons and matching functionalities on the TheFuzz Ratio for this project.

6.1.2 Retrieval Rate

The Retrieval Rate (RR) is a key metric that was implemented and adapted to evaluate the precision or accuracy of an STT model for a single word or a group of words. This metric indicates how well the model transcribes words from audio inputs. For example, if a word appears in the base sentence and is consistently found in the transcription, the RR for this word would be 100%. Conversely, if it appears in only nine out of ten sentences, the rate drops to 90%.

The RR is a word-by-word measurement without alignment, akin to the 'Bag of Words' approach, a model commonly used in NLP where words are treated as individual tokens without considering their order or context in a sentence [47]. This method provides a straightforward way to measure how frequently words are correctly found in the transcriptions.

Groups Functioning: In the algorithm, the RR can be measured for all words as an average or separately for two groups of words to compare them (e.g., dialect-specific vs. non-dialect-specific). To form a group, the RR can be averaged over multiple words. By categorizing words into these groups, the accuracy of a transcription model in handling different types of words can be assessed.

Word-Level Functioning: Additionally, the RR can be measured per word, termed the Word-Level RR. This algorithm does this by collecting the RR for each word and within each word for each speaker. This allows for computing an average per speaker for a specific word and then computing the average RR per word based on these speaker averages. This approach ensures balanced RR for speakers, meaning that a word will only have a truly low or high RR if multiple speakers pronounce the words similarly for the model.

Algorithm: The algorithm we developed to calculate RR includes the following key parameters for the RR measurement:

- **Similarity Thresholds:** This refers to the threshold of TheFuzz similarity ratio, above which two words are considered a match. For more details on similarity matching, refer to the Similarity Matching chapter 6.1.1.

[26] Partial matching refers to comparing substrings or segments of the text rather than requiring an exact match of the entire string. Token-based matching involves splitting the text into meaningful units or tokens (e.g., words, phrases) and comparing these individual units to find similarities.

- **Ignore Stopwords:** Ignoring stopwords is a common NLP technique that focuses on meaningful words that add actual content to the sentence. For more details on stopwords, refer to the Glossary (8.4). Stopwords are treated on a word basis, meaning during the processing of sentences and transcripts, if this option is activated (set to the boolean value 'true'), words found to be stopwords are not considered, and their RR values are not computed nor added to any group. For this project, the NLTK library and its German stopword list[27] was used to remove stopwords.
- **Ignore Preterite Tense Sentences:** As seen in chapter 4.5, the preterite tense in German poses issues for our analysis because Swiss German lacks this tense. This parameter allows the exclusion of all records containing preterite tense sentences. Preterite tense sentences are treated as entire sentences; if a sentence is written in preterite tense, it is completely excluded from the analysis when this parameter is activated.

Word-Level RR specific parameters: These supplementary parameters were only included in the Word-Level RR version of the algorithm due to the fundamental difference in looking at individual words.

- **Minimal Sentence Count**: This parameter defines how many records a word must appear in to be included in the analysis, ensuring that the data is robust and reliable. Any word occurring less than this parameter's value will not be included in the analysis output.
- **Ignore Numbers**: The datasets contain sentences with numbers written in different forms, either as words (e.g., 'drei') or as digits (e.g., '3'). When the base form does not match the transcribed form (e.g., 'drei' to '3' or vice versa), they cannot be matched with similarity. Therefore, this parameter is provided to exclude any numbers categorically from the Word-Level RR computation to prevent such instances from crowding the results with irrelevant low RR numbers that do not offer any useful insight. We used the text2num module[28] to identify numbers written as words.

The process developed to measure the RR is outlined with pseudo-code in algorithm 1.

Algorithm 1 Calculate Retrieval Rate

```
for each base-sentence do
    for each word in base-sentence do
        if word meets similarity threshold with any word in transcript then
            Append 1 to appropriate list of RRs*
        else
            Append 0 to appropriate list of RRs*
        end if
    end for
    Compute average RR for all lists*
end for
```

** Lists: If a group of words is given as a parameter, the algorithm will collect RRs for all words in that group in one list and RRs for all words not in that group in a second list. If no group is given, the RR over all words is computed, and only one list will be used. If the algorithm is used in Word-Level mode, it will compute RRs for each word individually, creating lists for each word.*

[27] The NLTK (Natural Language Toolkit) library is a commonly used module for building Python programs to work with human language data, and it includes a comprehensive list of German stopwords [48]

[28] The text2num module is a Python library that converts numbers written as words into their numerical form [49]

This generalized approach ensures that the RR is measured accurately and consistently across different groups of words or individual words, taking into account various parameters that allow the output to be tailored to specific requirements.

6.1.3 Replacement Word Frequency

To understand why certain words in the Standard German base sentences might have a low RR and to identify the words that have likely been transcribed instead of these words by the STT model, an algorithm was developed that encompasses the following steps:

1. One target word is chosen as the one to be analyzed. The goal is to find words in the transcripts that have the most probability of having been transcribed for it (e.g., for the target word 'Knaben', the alternative form commonly used in Swiss German, 'Buben', is expected to be found).
2. The algorithm collects all records where the base sentences contain the target word and analyzes their corresponding transcriptions. It focuses solely on instances where no match was found for the analyzed word, meaning records, where the word could successfully be found, are ignored. This can be run either the entire dataset or a subset, such as all records from a specific region, which can be used to find regional replacements.
3. It then extracts all words in those transcripts that cannot be aligned or matched with any word from their respective base sentence. The word being analyzed will not be found, as all records with a transcription containing it have been filtered out during the previous step.
4. These words are then collected, counted, and sorted by frequency to create a list of words, ordered by how often they appear. This represents a list of candidates likely to have been transcribed instead of the target word, from most likely to least.

Further specific functionalities of the algorithm are:

- The algorithm allows for the option to ignore stop words, ensuring that common words like 'and' or 'the' do not skew the results.
- It can match words exactly or based on a similarity threshold, which is useful for handling variations in spelling or minor transcription errors.
- To order the output words, the analysis can be conducted using the term frequency-inverse document frequency (TF-IDF)[29] weighting to consider the importance of words within the dataset or using simple word counts represented as a percentage of total sentences analyzed for more straightforward frequency analysis. More information about can be found in the Glossary (8.4). For this analysis, simple word counts were chosen, as the results of the algorithm were nearly identical to those obtained through TF-IDF sorting. Additionally, the computation was much faster, and the results were easier to understand.
- The output includes the most common words that were likely transcribed instead of the target word, the total number of distinct words identified, the number of transcriptions analyzed, and the total number of sentences that originally contained the target word.

Example: A very basic example output of this algorithm is outlined below, for the most common words replacing 'Ziege':

[29] TF-IDF is a statistical measure used to evaluate the importance of a word in a document relative to a collection of documents [50].

- geissen: 66.67%
- könnt: 11.11%
- ergibt: 11.11%
- frische: 11.11%
- aufpasst: 11.11%
- mitnehmen: 11.11%
- mehr: 11.11%
- züge: 11.11%

We can observe the word 'geissen' at the very top, which figures in 66.67% of the transcripts where the base sentence contains the word 'Ziege'. In some dialect regions, the word 'Geiss' replaces 'Ziege', which can be interpreted from this output. Additionally, the algorithm could also output the following values in this example:

- **Total distinct words found**: 8
- **Transcriptions analyzed**: 11
- **Total sentences containing word**: 13

Giving more detailed information on how the result came together: A total of 13 base sentences were found containing the word 'Ziege', eleven sentences were analyzed (where 'Ziege' did not figure in the transcriptions when respecting the similarity threshold), and eight distinct words found, which form the candidates for replacement shown above.

6.2 Speech To Text Model Research

This chapter explores existing STT models and includes a comprehensive performance analysis. We examined various models to understand their capabilities, advantages, and limitations. This investigation aimed to identify the most suitable STT models for the study, focusing on analyzing Swiss German dialects using Standard German STT models. Through this exploration, the accuracy and reliability of the linguistic analysis were enhanced.

6.2.1 Vision and Goals

The primary vision behind employing an STT model trained exclusively on Standard German was to leverage its specific limitations—its inability to recognize Swiss German dialect words—to our advantage. Using a model that was not trained in Swiss German and without attempting to develop such a model ensured that the transcription process focused on individual words without inadvertently translating or restructuring sentences into Standard German syntax. This approach allowed most words to be captured accurately due to their close phonetic proximity to Standard German. The model's inherent constraints were particularly useful in highlighting words and syntactic structures peculiar to Swiss German. This methodology ensured that each word was transcribed as it was spoken, without the model applying learned sentence placements and word form translations that could mistakenly align Swiss German expressions closer to Standard German configurations.

It was also important to distinguish between accented speech and dialects. Swiss accents maintain the same grammar and vocabulary as Standard German but differ in pronunciation, whereas Swiss dialects involve pronunciation, vocabulary, and grammar/structural changes [14]. However, the delimitation between accents and dialects remains a critical problem, as discussed in Chapter 2.2.1. While the models could have been trained to recognize accented speech — which would be beneficial for understanding the pronunciation variations of Standard German — it did not extend to capturing dialect-specific words and structures. This distinction underscored the value of using a model trained solely on Standard German, as it avoided conflating dialect with mere accentuation, thereby preserving the linguistic peculiarities inherent in Swiss German dialects.

- **Fundamentally Different Words**: Our objective was to scrutinize the variations in vocabulary and pronunciation between dialects and Standard German. Words that were distinctive to dialects (heteronyms such as 'Knaben' -> 'Buben') or markedly differently pronounced (cognates such as 'Ball' -> 'Bölä') were likely to be transcribed inaccurately by a trained model with Standard German, thereby illuminating these fundamental differences.

- **Word Order**: Another key analytical focus was examining the alterations in word order specific to Swiss German dialects. By analyzing the transcriptions generated by the SG model, we could contrast these with the original Swiss German spoken texts to identify and understand syntactic variations.

6.2.2 Model Options

Selecting an STT model with high accuracy for Standard German was necessary to minimize errors unrelated to dialectical differences. The basis for the evaluation and model selection was the comprehensive evaluation of Standard German models by the GitHub user 'domcross' [51], which assessed various freely available and open-source models. Based on the findings from the domcross evaluation and our research, the following models were selected for their performance and training characteristics:

- **NVIDIA FastConformer-Hybrid Large (de)**: Known for its robust performance across standard and accented German, this model was trained on a blend of different linguistic datasets, including Mozilla's Common Voice (MCV) [52], Multilingual LibriSpeech (MLS) [53], and VoxPopuli [54], making it a suitable candidate for this project. Its predecessor featured a WER[30] of 5.77% in the domcross evaluation [35].
- **Jaco-Assistant Scribosermo (QuartzNet)**: This promising STT model was included due to its high ranking in comparative evaluations. It was trained on 37 datasets, including MCV [52], MLS [53], and Voxforge [55], and showed a WER of 9.43% in the domcross evaluation [56].
- **Whisper large-v1 to v3**: OpenAI's models were known for their versatility and high-quality transcriptions. Due to the lack of publicly available training data details, the focus for these models was more on testing to establish their efficacy in handling German dialects [36].
- **Silero**: This model offered enterprise-grade STT in a compact form factor and was designed to be robust against a variety of dialects, codecs, domains, and noises. Despite its versatility, the Silero model's performance in the domcross evaluation showed a WER of 18.98%, which was significantly higher than the other models evaluated, suggesting it might be less effective for the specific requirements [57].

6.2.3 Selection Criteria

For the initial detailed evaluation, the selection was based on three key criteria to ensure a comprehensive assessment of each model's capabilities:

1. **Accuracy of models**: Evaluations included comparisons based on the domcross analysis, which provided a benchmark of model performance against Standard German [51].
2. **Content of training data**: This criterion assessed the diversity and relevance of the training datasets, particularly focusing on the inclusion and representation of accented speech where word forms are generally kept the same but pronunciation varies.
3. **Testing**: The most important criterion, where models were tested using specific phrases from the datasets to observe their performance with the dialects in question directly.

As the Whisper models were not included in the domcross evaluation [51], we could only evaluate them based on our testing due to the lack of publicly available training data.

[30]The Word Error Rate (WER) is a common metric used to assess the performance of ASR systems. It calculates the percentage of words that were incorrectly transcribed by comparing the ASR output to a reference transcription.

6.2.4 Initial Evaluation

Accuracy: Based on the first criterion of accuracy, the Jaco-Assistant Scribosermo and NVIDIA FastConformer-Hybrid Large (de)'s predecessor models exhibited similar levels of performance for WER in the domcross evaluation [51] and were thus selected for further detailed testing. The Silero model, on the other hand, showed significantly lower performance and was not chosen for further evaluation.

Training Data: The second criterion assessed the quality of training data. Both the Jaco-Assistant Scribosermo and NVIDIA FastConformer-Hybrid Large (de) models include a diverse and largely overlapping set of training data, making them both well-suited for this criterion.

MCV [52] is an open-source project that provides and continues to build a vast and diverse dataset of human voices. It includes recordings from volunteers worldwide. This variety helps models trained on MCV to generalize better to different speech types and patterns.

MLS [53] is another extensive dataset that includes read speech from audiobooks in multiple languages.

Both Jaco-Assistant Scribosermo and NVIDIA FastConformer-Hybrid Large (de) have been trained on these two high-quality, diverse datasets. This enhances their performance in recognizing and processing speech from different sources and speakers.

Initial Tests The three remaining models were evaluated using sentences from the datasets containing the words 'Butter', 'Müll', and 'Herunterladen'. A selection of example transcriptions is presented in tabular format below to illustrate these findings. This visual comparison highlights the transcription accuracy and the models' ability to handle dialectal variations. As it is observable from all three examples, the Jaco-Assistant Scribosermo model performs by far the worst out of all models and is, therefore, quickly dismissed and not further analyzed. Table 19 and Table 20 show that both Whisper large-v2 and the NVIDIA model produce rather accurate transcripts and make the desired error: Transcribing the Bernese Swiss German word and heteronym 'Anke' as it is pronounced and not translating it to its Standard German form which would be 'Butter'. In Table 21 a further, intriguing example: Here NVIDIA produces a much worse general transcript in comparison to Whisper but makes the desired transcription error by transcribing the Bernese Swiss German word (compound heteronym and cognate) 'abelade' ('herunterladen' in Standard German) as 'Abalabe'. Whisper, on the other hand, makes a spot-on transcription and even translates the word 'abelade' to its Standard German form, 'herunterladen', which would be counterproductive to the initial goal of leveraging mistranscription of dialect-specific words as this example indicates that the model may sometimes not mistranscribe dialect-specific words but recognize and translate them to their Standard German form.

Dialect Region	Transcript	Model Used
Bern	zweitens anker elektrofahrrad	Jaco-Assistant Scribosermo
Bern	Zweitens, Anke ihre große Pfanne erwärme.	NVIDIA ASR
Bern	2. Anke ihre grosse Pfanne erwärme.	Whisper large-v2

Table 19: Transcripts for Sentence '2. Butter in einer grossen Pfanne erwärmen'

Dialect Region	Transcript	Model Used
Innerschweiz	dem sammelt sich erguesse	Jaco-Assistant Scribosermo
Innerschweiz	Indem sammelt sich der Güssel.	NVIDIA ASR
Innerschweiz	In dem sammelt sich der Güssel.	Whisper large-v2

Table 20: Transcripts for Sentence 'In diesen sammelt sich der Müll.'

Dialect Region	Transcript	Model Used
Bern	die aktuelle version wendeschalter aussentuer oberlaa	Jaco-Assistant Scribosermo
Bern	Die aktuelle Version nenne sich osterlosem I Tunes Store Abalabe.	NVIDIA ASR
Bern	Die aktuelle Version können Sie kostenlos aus dem iTunes Store herunterladen.	Whisper large-v2

Table 21: Transcripts for Sentence 'Die aktuelle Version können Sie kostenlos aus dem iTunes Store herunterladen.'

To summarize:

- **Jaco-Assistant Scribosermo (QuartzNet)**: This model made significant errors in the initial tests and was quickly dismissed from further consideration.
- **NVIDIA FastConformer-Hybrid Large (de)**: Performed reasonably well, delivering accurate transcriptions and showing good handling of the dialectical variations in alignment with our goals by transcribing pronounced words rather than translating them.
- **Whisper (large-v1, large-v2 and large-v3)**: Consistently produced high-quality, often spot-on transcriptions. Mostly handled dialectical variations in alignment with our goals while also translating words in some cases, which warranted a more thorough investigation in subsequent tests.

Following this initial evaluation, both NVIDIA and Whisper (large-v2) underwent a more thorough investigation to explore their respective strengths and limitations concerning their behavior with dialect-specific words. This deeper analysis aims to determine which model suits the project and how it can be used in the best way to gain interesting insights.

6.2.5 Advanced Evaluation

As the evaluation of the models progressed, the groundwork for subsequent experiments was established, aiming to measure the performance of both the NVIDIA and Whisper models. The initial goal was to identify a model that performed well with regular, non-dialect-specific words while failing to recognize Swiss German dialect-specific words, which are fundamentally different.

Groups: To compare the models' performance on dialect-specific versus regular words, we measured the RRs for these groups. A significant challenge was to define what constitutes a dialect-specific word. For preliminary testing of both models, we chose the following word lists:

1. **Dialektwörter.ch (dw.ch)**: A publicly available list of over 2'000 Swiss German words along with their Standard German translations. [58]
2. **Personal selection (selection)**: A curated list of eight Standard German words that exhibit significant variations in Swiss German, compiled using our expertise as Swiss German speakers and resources like Pons [59] and Auswanderluchs [60] (see Table 22). It is constituted of words with heteronym counterparts except 'Kuchen' and 'Mittagessen', which have cognate counterparts.

Standard German	Swiss German Examples
1. Müll	Abfall/Ghüdder
2. herunterladen	abelade
3. etwas	öppis
4. einmal	einisch
5. Kuchen	Chueche
6. Frühstück	Zmorge
7. Mittagessen	Zmittag
8. Idiot	Tubel

Table 22: Personal Selection of Standard German Words and their Swiss German Counterpart-Examples

The RR algorithm was executed multiple times, using a grid search approach[31], running each configuration with both the words list from Dialektwörter.ch and our selection. Additionally, the algorithm's parameters have been varied:

1. **Ignore Stopwords**: Alternating between 'true' for ignoring stopwords and 'false' for considering them in the analysis to analyze the impact of stopwords in the RR Computation.
2. **Match Exact and Similarity Threshold**: Alternating between 70% similarity, 80% similarity, and exact match, being equal to a 100% similarity, to analyze the impact of matching criteria on the RR Computation. For more information concerning the similarity threshold, refer to the similarity matching Chapter 6.1.1.
3. **Ignore Preterite Tense Sentences**: This was not used in this step of the analysis, as the preterite tense filtering pipeline was added in a later step of the project.

6.2.6 Results

NVIDIA FastConformer-Hybrid Large (de) During the evaluation of the NVIDIA FastConformer Model using both SDS-200 and STT4SG-350 datasets, low RRs were observed. We ran the RR-Analysis as a grid search over different parameters, which is shown in Table 23. The highest RR for non-dialect-specific, regular words was a mere 64.54%, even with a similarity threshold as low as

[31] Grid search is a systematic method for hyperparameter optimization that exhaustively searches through a manually specified subset of the hyperparameter space [61].

70%. This low performance is problematic as it hampers our ability to determine during later experiments whether words are dialect-specific or if the model's transcription quality is inherently poor. After reviewing these outcomes, we decided not to proceed with further analysis using the NVIDIA Model. The results of the grid search, sorted by non-dialect-specific RR first and then by dialect-specific RR, are shown below. The word set name 'all' represents all words from Dialektwörter.ch [58], and 'selection' refers to the curated selection described in Chapter 6.2.5.

Word Set	Ignore Stopwords	Match Exact	Similarity Threshold	Dialect-Specific RR (%)	Non-Dialect Specific RR (%)
all	True	True	N/A	41.39	39.37
selection	True	True	N/A	2.86	39.42
selection	False	True	N/A	13.62	43.41
all	False	True	N/A	34.90	43.65
selection	False	False	80	14.20	54.38
all	False	False	80	39.03	54.83
selection	True	False	80	8.57	56.67
all	True	False	80	51.01	56.79
selection	False	False	70	15.12	59.66
all	False	False	70	44.77	60.09
selection	True	False	70	8.57	64.35
all	True	False	70	55.70	64.54

Table 23: Nvidia FastConformer-Hybrid Large RR Comparisons

Whisper Large-V2 The Whisper Model exhibited more promising results: High RRs were achieved for both dialect-specific and non-dialect-specific word groups. Notably, most parameter combinations that ignored stopwords resulted in a desirable imbalance: the Non-Dialect-Specific RR consistently exceeded the Dialect-Specific RR. This indicates that non-dialect-specific words from the base sentence were found in a high number of transcripts, while dialect-specific words were less frequently found in the transcripts due to their different spoken forms.

The Retrieval Rate-Analysis for the NVIDIA FastConformer-Hybrid Large (de) was also conducted as a grid search over different parameters for Whisper, as shown in Table 24.

Word Set	Ignore Stopwords	Match Exact	Similarity Threshold	Dialect-Specific RR (%)	Non-Dialect Specific RR (%)
all	True	True	N/A	82.11	77.11
selection	True	True	N/A	74.29	77.22
all	False	True	N/A	86.32	79.92
selection	False	True	N/A	69.39	80.15
all	True	False	80	84.76	84.63
selection	True	False	80	74.29	84.63
all	False	False	80	87.82	84.88
selection	False	False	80	69.77	85.00
all	False	False	70	88.73	87.20
selection	False	False	70	70.21	87.28
selection	True	False	70	74.29	87.70
all	True	False	70	86.10	87.74

Table 24: Whisper Large-v2 RR Comparisons

6.3 Dialect Variability Analysis

This chapter explores various analyses conducted using the Whisper STT model. It offers insights into which dialects are closest to Standard German and which are the most divergent. Additionally, it outlines a method to compute dialect-specific words and compares these across different dialect regions.

6.3.1 Regional Retrieval Rate Analysis

This analysis aims to compare the dialect regions and understand which one is the furthest or most different from Standard German by measuring how well the Whisper STT Model understood each dialect region. This was achieved by calculating the average RR over all words for each region individually. Unlike the advanced analysis for model choice, this analysis did not involve grouping of words (for example, dialect-specific and non-dialect-specific); instead, it calculated a single RR for each dialect region.

Method: We ran the RR Algorithm for all words as one group on all records of both datasets, STT4SG-350 and SDS-200, filtered by region. To gain a deeper understanding and derive meaningful insights, the following parameters were varied: alternations between 'true' and 'false' on the stopwords removal to analyze their impact in the RR computation; usage of similarity threshold of 90% for a rather restrictive matching; and the sentences in preterite tense were ignored, with the issues surrounding the preterite tense described in Chapter 4.5.

Results: The main output, incorporating the preprocessing steps of ignoring stopwords and excluding sentences in the preterite tense, is shown below. Due to the vast distribution of speakers, median values were compared for a more accurate representation. The median RRs for each region are as follows:

- Innerschweiz: 85.88%
- Graubünden: 84.70%
- Zürich: 84.38%
- Ostschweiz: 84.37%
- Basel: 84.21%
- Bern: 82.35%
- Wallis: 76.43%

Innerschweiz has the highest median RR and is, therefore, closest to Standard German. Wallis has a median RR of around 5-9% lower than other regions, meaning that the Wallis Swiss German dialect is the hardest transcribe for the Whisper large-v2 model and bringing empirical evidence aligning with the stereotype that it is particularly difficult to understand [62]. The distribution for each region is illustrated in Figure 14, with each point representing a speaker.

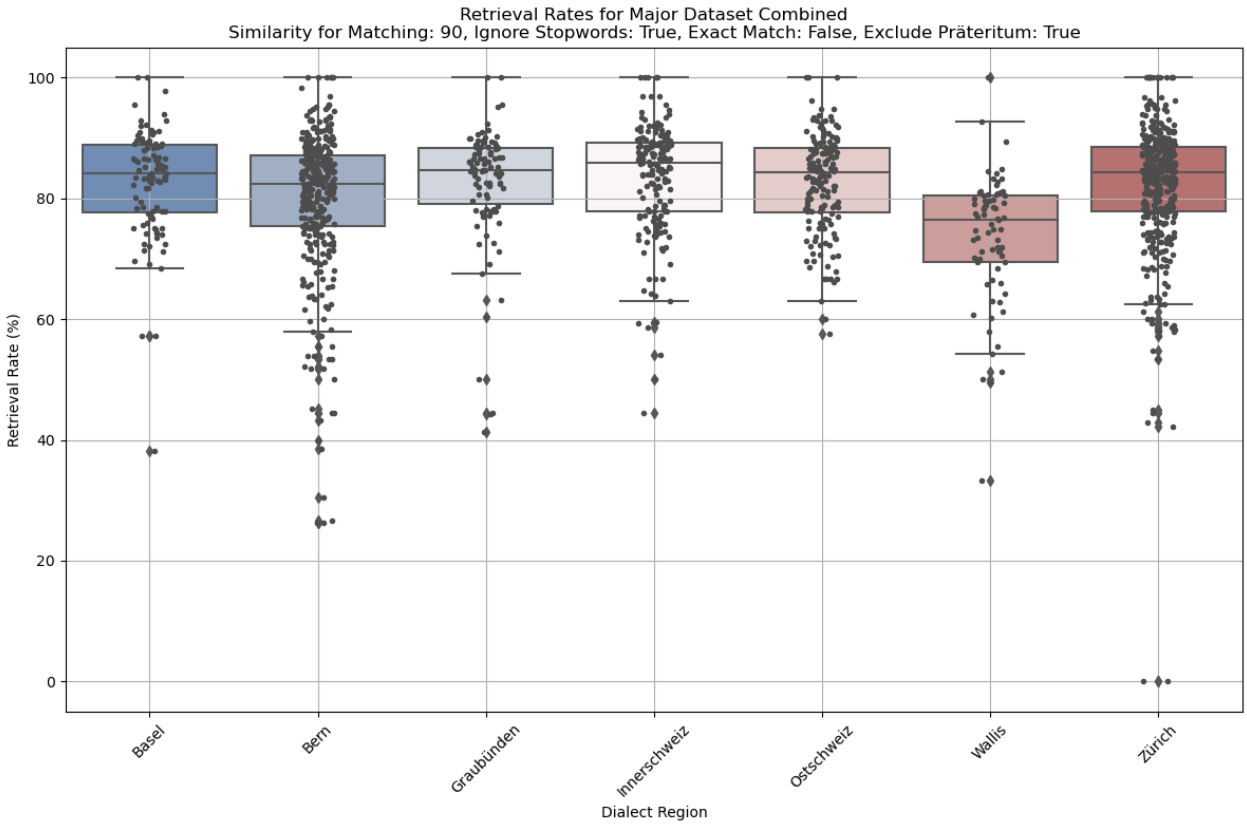

Figure 14: RRs per Region (Ignoring Stopwords and Preterite Tense Sentences)

Stopwords In this first example, a total of 2'213'814 words were processed, and 1'057'502 stopwords were excluded, resulting in a total of 1'156'312 words included in the RR calculations.

When running the analysis considering stopwords (not ignoring stopwords), the ranking of the regions remains the same, and the medians don't shift by more than 1%. The RRs can be seen below, and the distribution is illustrated in Figure 15.

- Innerschweiz: 86.04%
- Graubünden: 85.63%
- Zürich: 85.10%
- Basel: 84.90%

- Ostschweiz: 84.30%
- Bern: 82.61%
- Wallis: 75.90%

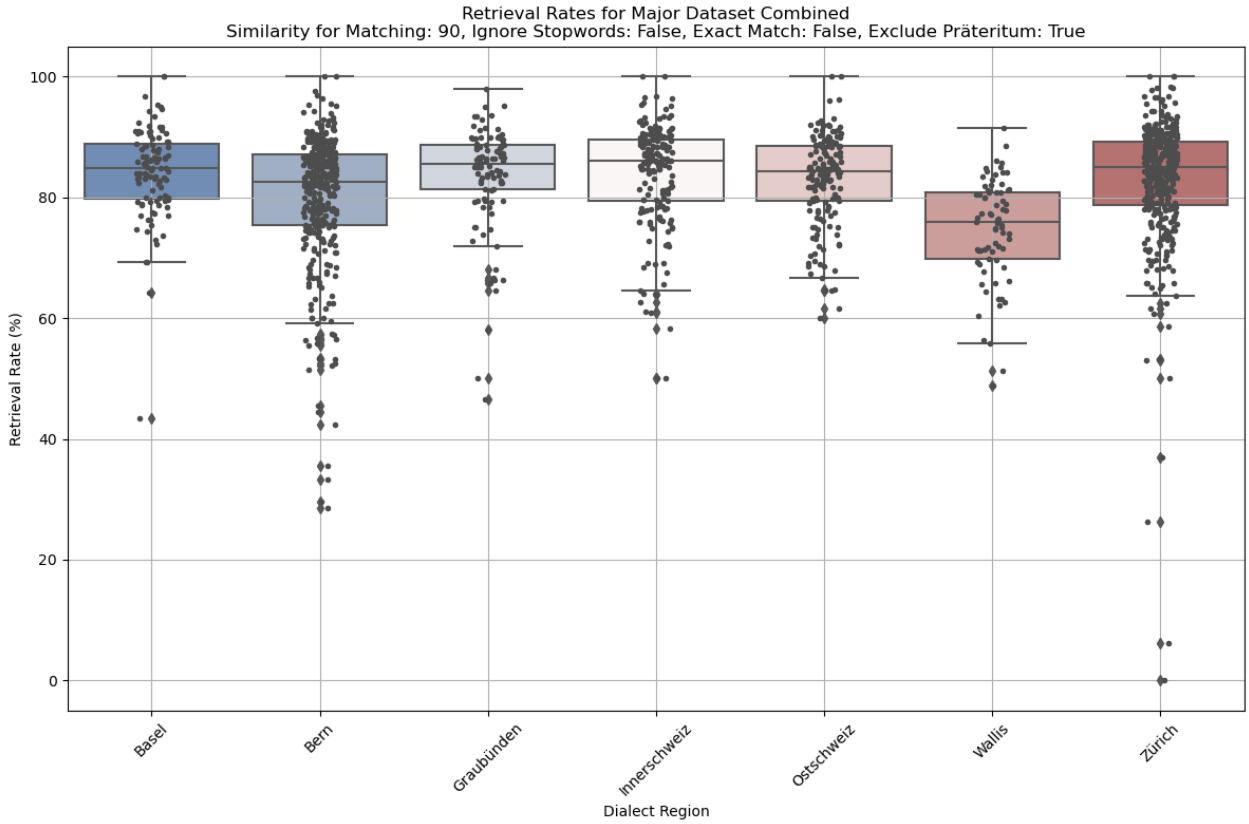

Figure 15: RRs per Region (Include Stopwords, Ignore Preterite Tense Sentences)

Special cases; RRs below 10% As shown in all figures, there are three obvious outliers in the Zürich region. Upon manually reviewing these records and listening to the audio files, it was evident that the recordings were not corrupt. Instead, the spoken sentences' content is entirely different from the base sentence, likely due to user error during dataset creation.

6.3.2 Dialect-specific Words Research

This analysis aims to extract dialect-specific words from the datasets by developing a data-driven method for recognizing and computing a list of dialect-specific words.

List Generation To generate a list of dialect-specific words, we computed the Word-Level RR (as seen in Chapter 6.1). In addition to the previously outlined parameters (ignore stopwords, similarity threshold, ignore preterite tense) for RR computation, we included the parameter ignore numbers (always set to 'true') to not populate our lists with numbers that will have a low RR, due to formatting variants, as outlined in Chapter 6.1.2. Additionally, the parameter minimal sentence count was used, which defines the minimal number of sentences a word must be present in to be included in the analysis. Two examples are shown to illustrate the lists that could be generated with the method.

Example 1: Similarity Threshold for Word Matching: 70%, Min sentence count: 20
For this example, a similarity threshold of 70% was chosen for a rather generous matching, allowing words to be matched even if they're written in a very different form. This leads to the consideration of various word forms as one by giving them a significant chance to be matched. A minimum sentence count of 20 was selected to catch more words that were used by multiple speakers from different regions. A lower minimum count would have led to more words with an extremely low RR, many of which would have an RR of 0.0%. While this method might have been more effective for recognizing dialect or region-specific words, it would have been less suitable for this initial example, as the aim was to inspect the low, mid, and high ends of the computed list.

In this first example, the results, which can be seen in Table 25a, showed only two words ('zirka' and 'gucken') with an RR of 0%, followed by 'bekanntgegeben' with 3.13%, and after that, a rapidly rising RR.

The words around the median that can be seen in Table 25b had roughly a 95% RR, which was surprisingly high, meaning that the Whisper large-v2 model very well recognized most words. The ignoring of stopwords and preterite tense sentences in this example, the low similarity threshold for matching (70%), and the limit to words with at least 20 occurrences contributed to these high RRs.

The third list slice Table 25c shows the words with the highest RR, all of which were transcribed correctly in all cases.

Table 25: Example 1 - List of Standard German Words with their Respective RR

Word	RR
zirka	0.00%
gucken	0.00%
bekanntgegeben	3.13%
lädt	3.85%
abs	4.76%
hierfür	6.85%
prozent	6.90%
weshalb	10.06%
begonnen	10.36%
circa	12.00%

(a) Lowest 10 RRs

Word	RR
streng	94.74%
jeans	94.74%
vernichtet	94.74%
abgehalten	94.74%
heftige	94.74%
aussenpolitik	94.74%
kommissionen	94.74%
brücken	94.74%
verkehrsmittel	94.74%
universitäten	94.74%

(b) Median RRs

Word	RR
wikipedia	100.00%
sitzplätze	100.00%
erschreckend	100.00%
grundlegenden	100.00%
mischt	100.00%
spekulation	100.00%
sparmassnahmen	100.00%
schlimmste	100.00%
epidemie	100.00%
unstimmigkeiten	100.00%

(c) Highest 10 RRs

The distribution of RRs can be seen in Figure 16, which illustrates the number of words (frequency) on a logarithmic scale for the entire RR span (from 0% to 100%).

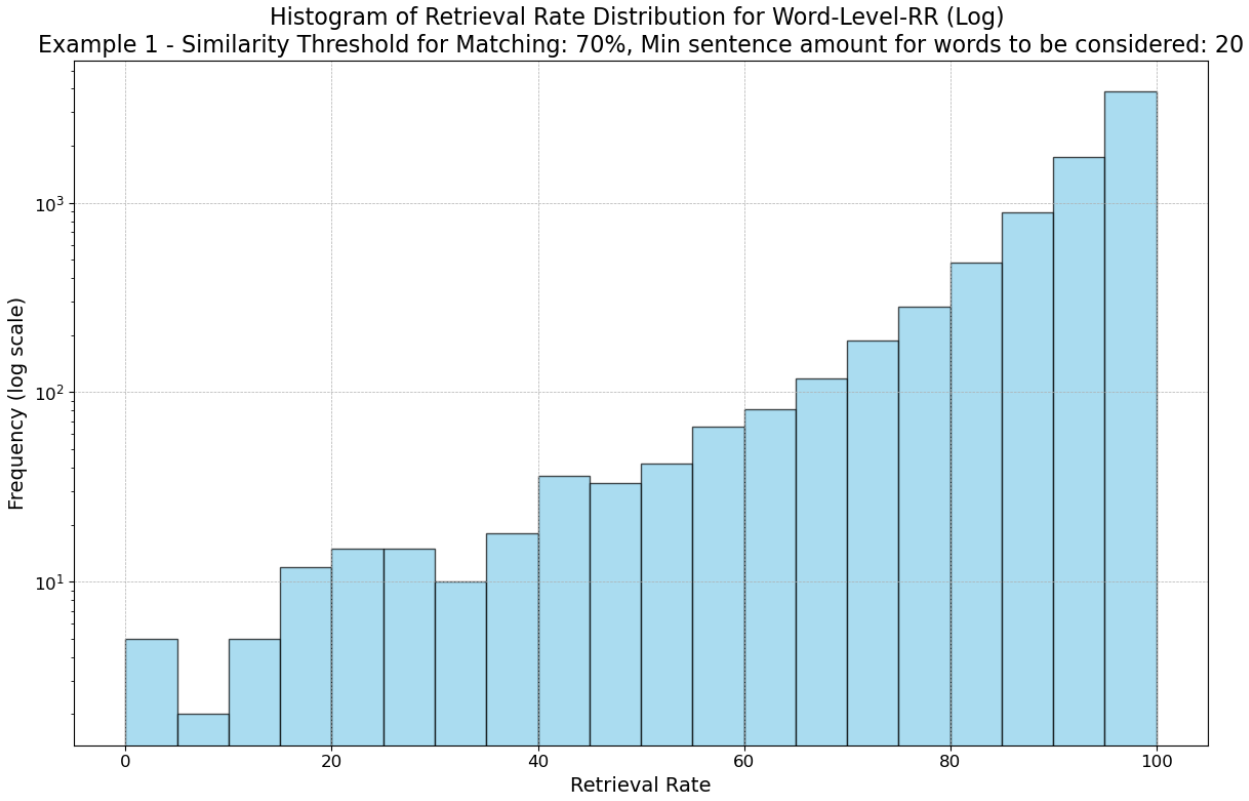

Figure 16: Histogram of RRs for Word-Level-RR Output of Example 1 (Log)

Example 2: Similarity. Threshold for Matching: 80%, Min sentence count: 20 For the second example, the similarity threshold was set higher at 80%, meaning a more rigorous matching with less chance of matching words in different forms. The minimum sentence count was kept the same to isolate and focus on the effect of the similarity threshold. As observed in Table 26a, there were more words with an RR of 0%, indicating that words were more singled out due to the increased similarity threshold. Words like 'verstarb' or 'hats' were treated as their own (two) different word forms in this example, whereas they might have been considered and matched with 'starb' or 'hat' in the first example, perhaps scoring a higher RR in those groups. Therefore, these words did not appear in the first example as they had a higher RR.

Still, a slightly lower can be observed, but very high, RR for words around the median, as shown in Table 26b. As expected, the top words all had a 100% RR, which is shown in Table 26c.

Table 26: Example 2 - List of Standard German Words with their Respective RR

Word	RR
hoeness	0.00%
nciht	0.00%
zirka	0.00%
tägi	0.00%
gucken	0.00%
eingebüsst	0.00%
hats	0.00%
verstarb	2.70%
bekanntgegeben	3.13%
lädt	3.85%

(a) Lowest 10 RRs

Word	RR
test	89.23%
notwendig	89.24%
häufiger	89.25%
transfers	89.25%
gewinner	89.25%
vielen	89.25%
zeigt	89.26%
frauen	89.26%
mittel	89.26%
geäussert	89.28%

(b) Median RRs

Word	RR
politikerinnen	100.00%
bulgarien	100.00%
präsidentin	100.00%
asiatischen	100.00%
wikipedia	100.00%
erschreckend	100.00%
grundlegenden	100.00%
spekulation	100.00%
schlimmste	100.00%
unstimmigkeiten	100.00%

(c) Highest 10 RRs

The distribution of RRs can be seen in Figure 17. A more balanced distribution was observed when comparing this example with the first one by using an 80% similarity threshold.

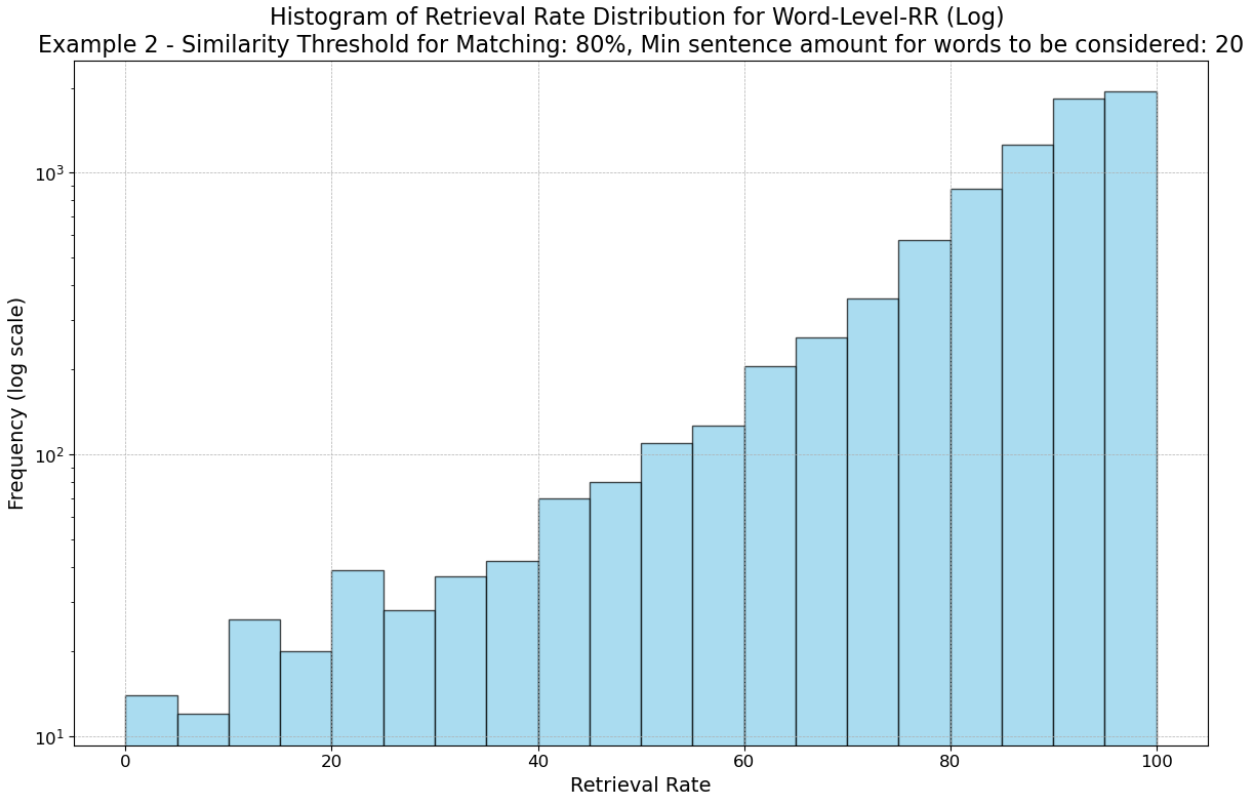

Figure 17: Histogram of RRs for Word-Level-RR Output of Example 2 (Log)

RR Threshold: After generating the list of all words' RRs, the challenge lay within the selection of the RR threshold, to consider any words below that threshold as dialect-specific. Due to the nature of this project and the expertise being in computer science and programming rather than

linguistics, the outputs were verified by comparing the computed lists with an existing dialect-word list.

List Comparison A method was developed to compare the generated lists of dialect-specific words to existing word lists to measure similarity. This comparison involves several steps and parameters. For this project, the output was compared to the list of single words found in Dialektwörter.ch [58], the same list used in the advanced model comparison in Chapter 6.2.5. The words were used in their written Standard German form and were not lemmatized.

A grid search was conducted over the parameters of the Word-Level-RR algorithm and RR-Threshold to fine-tune the computation of the dialect-specific word list:

1. **Ignore Stopwords, Numbers, and Preterite Tense**: Always ignoring as these would impact and flood the dialect words list with non-relevant words.
2. **Similarity Thresholds**: Ranging from 70 to 100, where 100 denotes an exact match.
3. **Minimal Sentence Count**: Alternating between five as this allows for more region dialect-specific words to be uncovered and 20, which can highlight more general dialect-specific words.
4. **RR Threshold**: Varying between 0% and 100%.

All combinations of these parameters were iterated over to compute dialect-specific word lists, and each was compared to the extracted list from Dialektwörter.ch.

Filtering Words : Before comparison, all words provided by Dialektwörter.ch were filtered to include only those present in the dataset using the following process:

1. Extracted all unique words from the sentences in both datasets (SDS-200 and STT4SG-350).
2. Applied exact or similarity matching based on the similarity threshold to filter the list of dialect-specific words from Dialektwörter.ch. Any word that could not be matched with a word from the datasets was dropped for this comparison.

Comparison Process : The DW.ch list was treated as a set of true dialect-specific words to compare and evaluate the generated lists. This approach allowed Precision and Recall to be computed for each list generated from different parameter combinations. It is important to note that while the DW.ch list does not claim to be exhaustive, it was one of the largest and most complete lists available at the time. This list could be replaced with any other list treated as the true set of dialect-specific words for this analysis.

1. Precision: The ratio of correctly predicted dialect-specific words to the total number of predicted dialect-specific words. It measures the accuracy of our algorithms' dialect-specific word list when using the specific parameter set [63].
2. Recall: The ratio of correctly predicted dialect-specific words to the total number of actual dialect-specific words. It measures the algorithms' ability to identify all relevant words when using the specific parameter set [63].
3. F1 Score: Combination of Precision and Recall, often described as their harmonic mean. It balances both the precision and recall of our algorithms, allowing us to find the optimal combination of parameters by comparing only one metric [63].

More detailed information about these metrics can be found in Chapter 8.4.

Comparison Results In this analysis, the goal was to identify a combination of parameters for the Word-Level RR algorithm that would yield acceptable precision and recall when compared with the filtered Dialektwörter.ch list, ultimately resulting in a satisfactory F1 score. It was anticipated that an RR threshold of 20-30% combined with a similarity threshold between 80% and 100% would achieve this goal. The expectation was that precision would initially be high and then decrease beyond this threshold, while recall would exhibit a substantial increase within the 20-30% RR threshold and then level off. This pattern would suggest the discovery of numerous dialect-specific words at lower RRs, with diminishing returns at higher RRs.

However, the results were unexpected and did not conform to the hypotheses.

Figure 18 illustrates the development of precision and recall under the scenario where the minimum number of sentences is set to five. The lines in the figure represent various similarity thresholds, the X-axis shows different RR threshold values, and the Y-axis depicts the precision or recall scores. Contrary to the expectations, precision exhibited a slight but noticeable increase initially, stabilizing at a low range of 6-8% to then decreasing again for the similarity threshold of 80%. Conversely, recall increased steadily but more significantly at higher RR values, which is the opposite of the expected behavior. The similarity threshold of 70% was notably different from the other thresholds, displaying completely different patterns and indicating that it was not restrictive enough and resulted in matching words that were too dissimilar from each other.

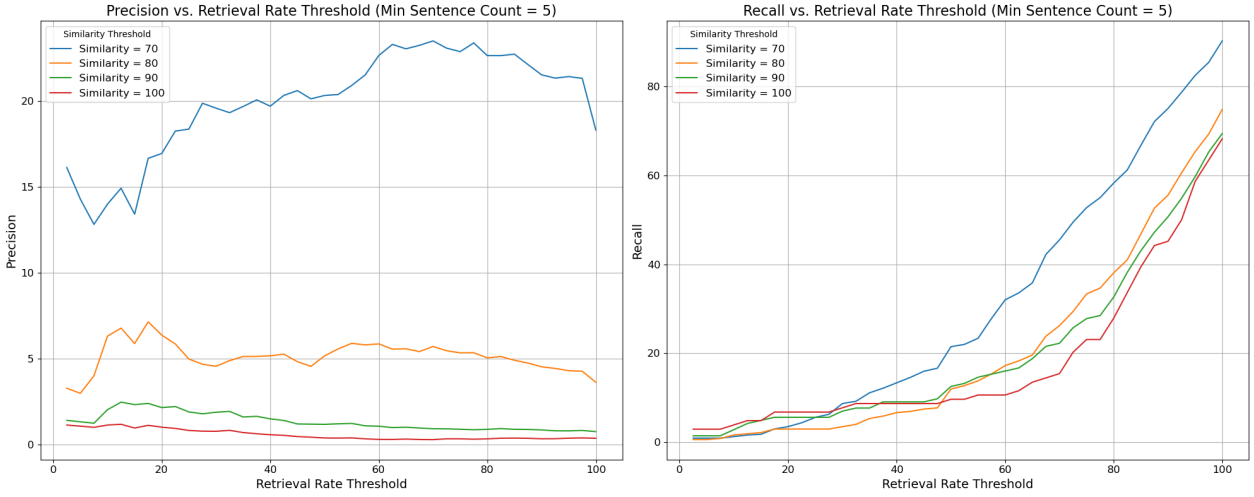

Figure 18: Dialect-Specific Words List Comparison - Evolution of Precision and Recall for Min. Sent.: 5

Figure 19 presents the evolution of the F1 score. The F1 score showed a steady increase up to the high 90s RR thresholds, contrary to the anticipated leveling off.

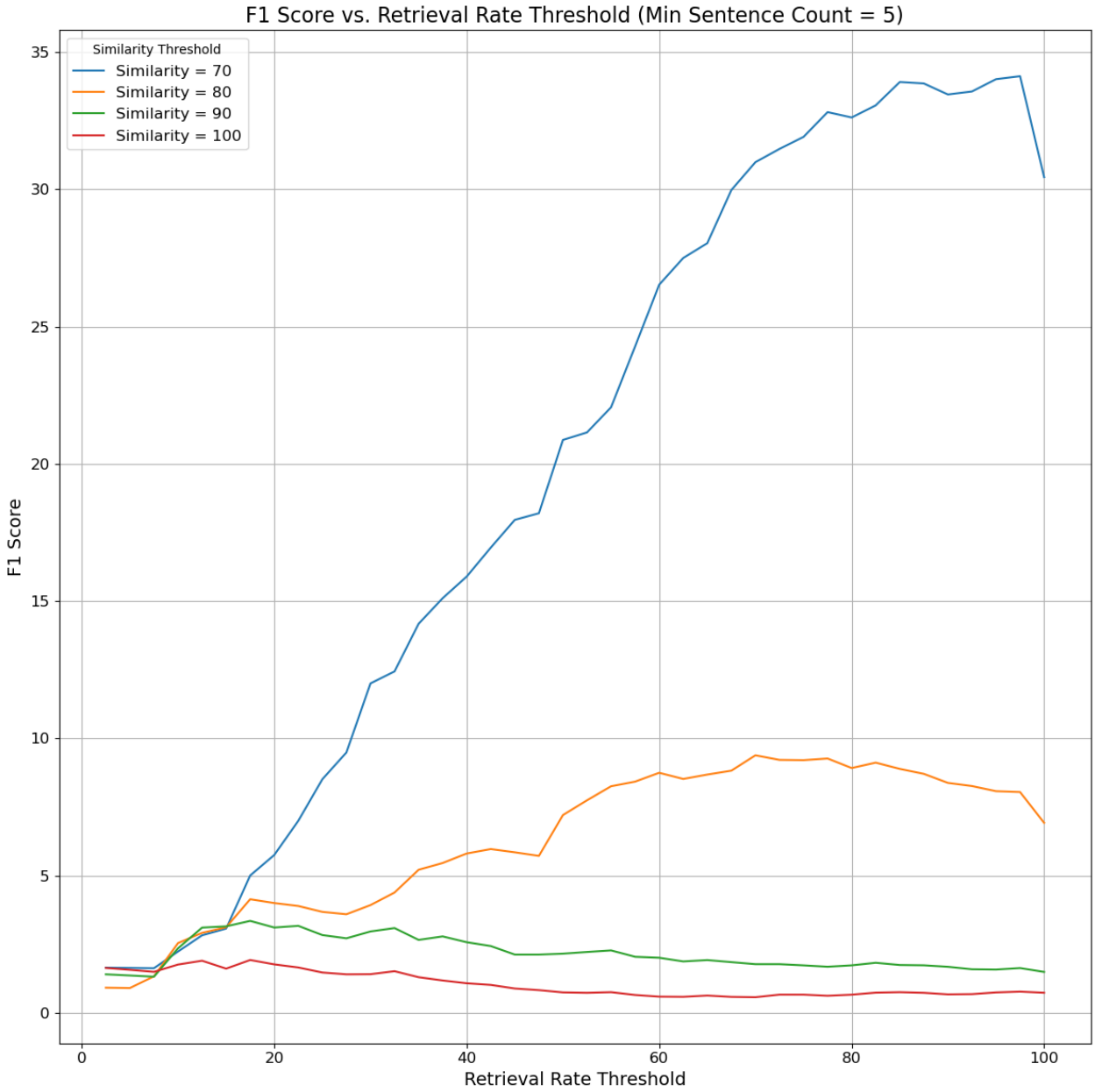

Figure 19: Dialect-specific Words List Comparison - Evolution of F1-Score for Min. Sent.: 5

These unexpected results suggest that the grid search fails to identify a viable combination of parameters that could reliably pinpoint dialect-specific words. Furthermore, it suggests that the method fundamentally lacks a valid combination capable of achieving this goal. This outcome may suggest several possibilities: the method used for generating and discovering dialect-specific word lists might be inherently flawed, the comparison method might not be suitably adapted for this task, or the existing word list, regarded as containing 'true' dialect words, might be incomplete or otherwise inadequate for this purpose. If the used list is outdated and does not represent modern dialect usage, it could also compromise this analysis. Further investigation and refinement of both the methodology and the word lists are required to improve the reliability and accuracy of dialect-specific word identification.

Manual List Verification: As the data-driven list comparison did not work out as expected, a different verification process was needed to measure the quality of the computed lists. The new focus lay on manual verification by Swiss-German-speaking individuals. A pipeline was developed in which the user could select the previously outlined parameters to compute the Word-Level RR list, and an RR threshold could be set as an upper bound to consider only words below that threshold. Alternatively, the number of words to be verified could be specified, or the entire list could be processed without selecting a threshold or a maximum amount. In the new method of verifying, the user was then shown every word within the threshold, along with the results of the word replacement algorithm (explained in Chapter 6.1.3) for that specific word. The information shown to the user is illustrated as an example UI in Figure 20. Within the scope of this project, the verification was done with a terminal application. The user could then decide if the word was to be considered dialect-specific based on their knowledge and the replacement words shown to them. These potential replacement words were intended to provide more context as to why the word had a low RR and what potential dialect-specific forms appeared.

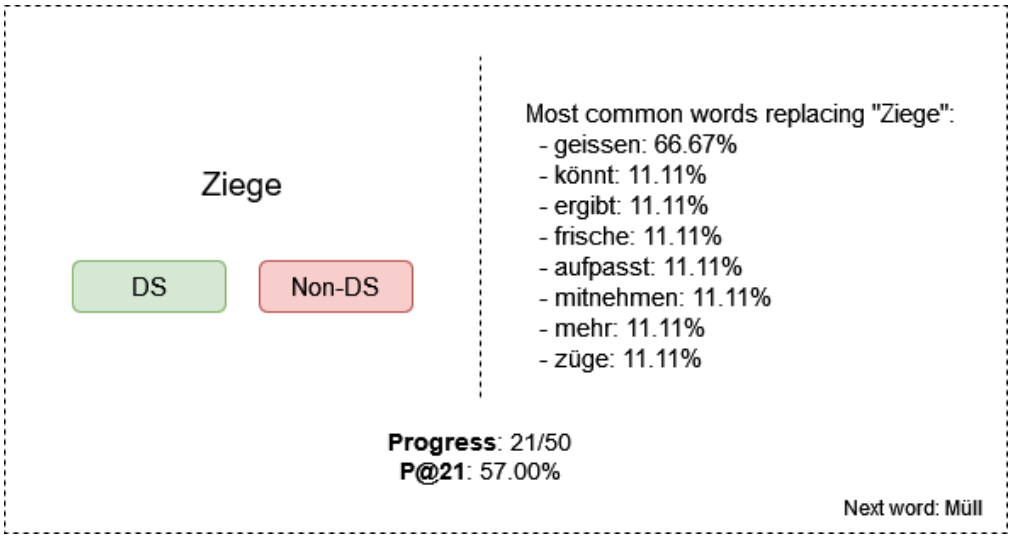

Figure 20: Dialect-Specific Words Manual Verification Interface

Throughout this process, the precision was steadily updated with the users' inputs. While determining more and more words and increasing the words included in the calculation, the precision was dynamically shown to the user to give an idea of the current precision of the dialect-specific word list. This can be seen at the bottom of Figure 20 with 'P@21' meaning the precision at 21 words. At the end of the procedure, the word list up to the specified threshold was saved, containing the sorted words with their RR and the new information on whether they were determined to be dialect-specific. In the previous verification, recall and F1 scores were also computed, as a list of considered true positives was used for comparison. However, this was not possible in this example, as the user only evaluated outputted words and did not create a list of dialect-specific words that could've been considered true positives.

Manual Verification Results: Within the scope of this project, one example evaluation was performed to demonstrate the potential outcomes of this verification process. As the initial goal was to discover parameters for list generation that produce a plausible list of dialect-specific words, multiple combinations of these parameters should be examined by a large number of evaluators. However, the necessary time and resources were not available for such extensive evaluation. By

averaging the different accuracy scores resulting from different evaluators, each set of parameters would have a score and could be compared to find the most effective one. For this example, the Word-Level RR was generated with a similarity threshold of 80% and a minimal sentence count of 20, as used in the examples in Chapter 6.3.2. The first 50 words with the lowest RR, ranging from 0% to around 20%, were manually reviewed to determine their dialect-specificity based on existing knowledge. The issue of defining a dialect-specific word arises again. Therefore, only words that have a heteronym or compound heteronym-cognate counterpart in any known dialect region known to us were marked as dialect-specific. The lowest ten words are shown in Table 27, along with their RR and dialect-specific labels.

Word	**Considered Dialect Specific**	**RR (%)**
hoeness	No	0.00
zirka	No	0.00
tägi	No	0.00
gucken	Yes	0.00
bekanntgegeben	No	3.12
lädt	No	3.85
verstarb	Yes	4.05
abs	No	4.76
liessen	No	6.60
hierfür	Yes	6.62

Table 27: Dialect Specific Word List Verification

The precision for ten analyzed words was 30%, and the precision after analyzing all 50 words, representing the final result of this example evaluation, was 40%. Many words such as 'zirka' and 'abs' were observed that have a low RR due to their transcription being fundamentally different, as they can be written in different forms and are sometimes shortened (bezüglich -> bezgl. or zirka -> ca), whether it's in the base sentence or the transcript. No meaningful insights can be derived from these candidates.

The second observation is the large number of words replaced with a heteronym, which is also very common in Standard German. Examples include 'hierzu' -> 'zudem' or 'beginnen' -> 'anfangen'.

This process could be repeated for multiple evaluators to gain an average precision score for this specific set of parameters. Those evaluators could then analyze multiple other sets of parameters using the same procedure, which would establish a baseline for determining the most effective combination of parameters.

6.3.3 Regional Words Comparison

The goal of this analysis was to compute region-specific dialect words, provide a general overview, and compare the regions based on these words. This included examining replacement words or misinterpretations that arise due to dialectal differences. By identifying and analyzing these regional variations, deeper insights into the distinct linguistic characteristics of each dialect region were gained. This analysis was performed on the two datasets, including all sentences from SDS-200 and STT4SG-350 that were not detected as the preterite tense.

Method: The analysis consisted of multiple steps. First, general and regional dialect-specific words were collected using the Word-Level RR algorithm. These lists were then merged, ensuring that only words used in all dialect regions were included. Finally, the RRs per word and region were visualized comprehensively, along with the replacement words for each combination. This approach allowed for a systematic comparison of the regions based on their specific dialect words and the identification of any replacement words or misinterpretations by the STT model.

Word List Computation: The list of words to be analyzed consisted of two parts:

1. **Global Dialect-specific Words**: This part included words that are pronounced in dialect-specific ways across all regions. Similar to the Dialect Words Research Chapter, the Word-Level RR Algorithm was used to compute the RR of all words with a similarity threshold of 70% (for more open matching and grouped results), ignoring stopwords, numbers, and preterite tense sentences, as well as using a minimum sentence count of 100 (to discover words used in a large number of sentences and therefore in multiple regions). The lowest 50 words on this list were chosen as candidates for further investigation.
2. **Regional Dialect-specific Words**: This part aims to include and discover words that are pronounced in a specific way in only one or a few regions. To find these region-specific words, the Word-Level RR Algorithm was run on all records of one region at a time with the same parameters but a lower minimum sentence count of five to include words that are less frequently used and, therefore, have a higher chance of being regional. The 50 words with the lowest RR for each region were chosen as candidates for further investigation.

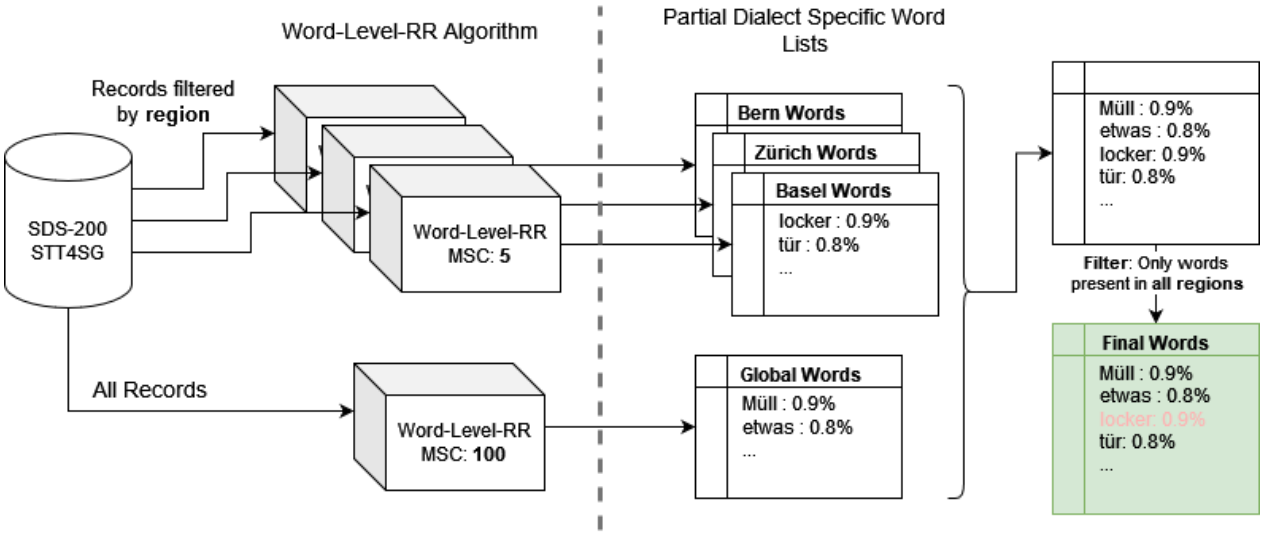

Figure 21: Regional Comparison - Word List Computation

All lists of words were then combined, resulting in a total of 199 word candidates. To compare the words across the regions, it was necessary to filter these words to include only those present in all regions, meaning in their respective, uncapped (not only the 50 lowest) Word-Level RR list. This filtering process removed words that were only recorded and transcribed in certain regions, thus eliminating their chance of being discovered and depriving the analysis of their inclusion. However, this filtering was necessary and beneficial because, without comparison across all regions, it would be impossible to conduct any analysis in a data-driven way. After filtering, the final list consisted

of 81 unique words with a low RR in either one region or the dataset overall. These were the final words selected for deeper investigation. The list computation is visualized in Figure 21.

Word-Region-RR Heatmap: This list was represented as a heatmap, correlating words with dialect regions and displaying the RR for each word-region combination. The structure of the heatmap is shown in Figure 22.

The Y-axis lists all Standard German words included in the final analysis, while the X-axis lists all dialect regions. Each cell shows the RR of the word within that specific region and the number of sentences on which this RR value is based. The RR indicates how well the word was transcribed in records from that specific region. The number of sentences can vary for a given word across different regions, as the SDS-200 dataset includes sentences recorded only in certain regions. Additionally, each cell contains two words identified through the replacement word frequency analysis, explained in detail in Chapter 6.1.3. This analysis identifies other words that were most likely transcribed instead of the word currently being analyzed. If the RR is 100%, the word under analysis was always found in the transcription of the records it figured in. The list of replacement words would, therefore, be empty, displaying 'No Data'.

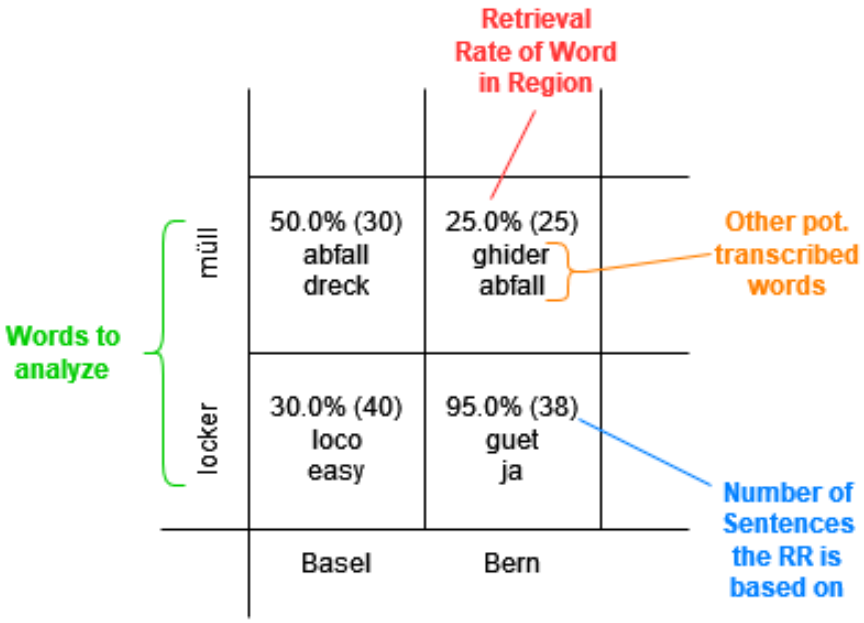

Figure 22: Regional Comparison - Heatmap Structure

Intending to differentiate between words specific to regional dialects and those common across all Swiss German dialects, the list was sorted of words to be analyzed by their global RR. Words at the higher end of the RR sorting were expected to have a low RR in only a few regions, while words at the lower end of the RR sorting were expected to have a low RR in most regions.

Replacement Words: To understand why the RR for a specific word/region pair is low and to identify words that were likely mistranscribed for that specific word, the replacement word frequency algorithm was used, which is explained in Chapter 6.1.3. These steps and visualizations ensure a comprehensive analysis of transcription inaccuracies, providing insights into why certain words have low RRs in specific regions.

Results: For the words on the lower end of the global RR sorting shown in Figure 23. As expected, words that are poorly transcribed with a low RR across all regions were found. The most notable observations of this selection of words are as follows:

- **Prozent, Zentimeter**: These words are often transcribed as symbols or units. They are not dialect-specific but are expected to appear in this result due to their extremely low RR, independent of the dialect region. If preterite tense sentences had been included, this result would be full of preterite tense verbs.
- **Weshalb, Begonnen, Geschieht, Getan**: These words are generally expressed differently (heteronym) in most regions. 'Weshalb' is often replaced by 'Warum' or 'Wieso', 'Begonnen' by 'Angefangen', 'Geschieht' by 'Passiert', and 'Getan' by 'Gemacht' across all regions.
- **Hierfür**: 'Hierfür' is rarely pronounced as 'hierfür'; it is mostly replaced by 'dafür' (compound heteronym/cognate).
- **hierzulande**: This compound word is seldom used as-is in any dialect region. It is unclear which words apart from 'Land' contribute to its composition (of the compound word) from this analysis.

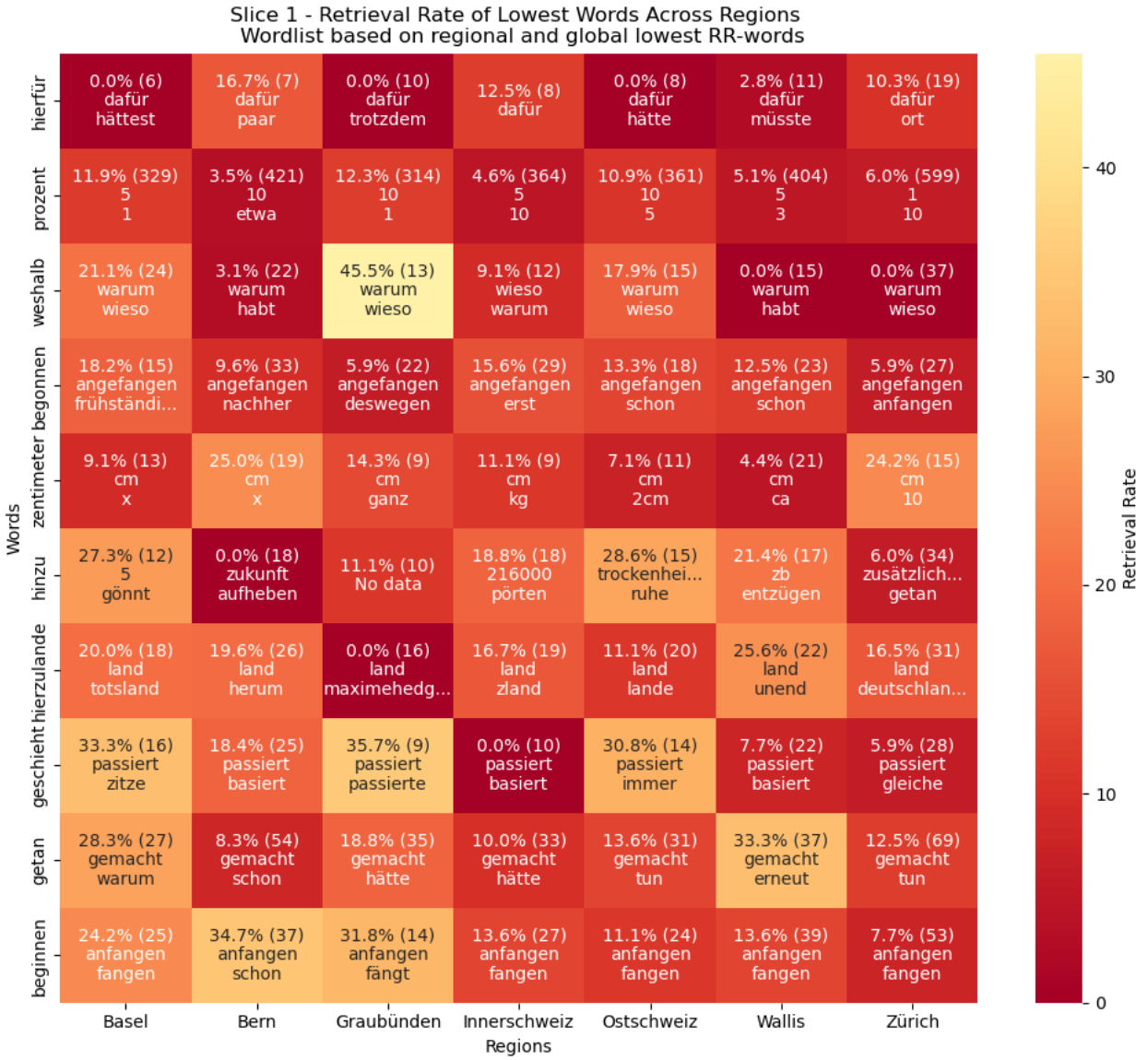

Figure 23: RR across Regions - Lowest global-RR (out of analyzed word list)

For the words at the higher end of the global RR sorting shown in Figures 24 and 25. Figure 24 represents the 10/21 lowest RR words. A mix of words was obtained, either with low RRs across all regions or words with only one region showing a significantly low RR.

- **Rede, Gebe, Daraufhin, Womöglich**: These words are generally used in a different form (heteronym) across all dialect regions.
- **Uhr**: This is another example of specific transcription, where the STT model transcribes a time indication as only digits (e.g., 07:00) or where it is pronounced without saying 'Uhr', as in Standard German ('Halb Fünf' instead of 'Vier Uhr Dreissig').
- **Mag**: Most regions pronounce the word 'mag' as written in Standard German. However, there is a notable difference in the Bern region, where the word 'gerne' is transcribed, perhaps in the form of 'gern haben' (heteronym).

- **Dauer**: In cases of mistranscriptions, this word is mostly transcribed as 'Tour' (cognate), reflecting the common pronunciation of 'au' as 'u' or 'ou', or 'ü' in the Wallis region. This is a challenge not only for the STT model but also for people learning Swiss German [64].
- **Rat**: This word is mostly transcribed correctly except in the Basel Region, where it is often recognized as 'rot', indicating that the 'a' might be pronounced as 'o' in this context.
- **Möchten**: This word has a very low RR score in the Wallis region, suggesting a pronunciation similar to 'werdet'.
- **Bekommen**: 'Bekommen' was mostly transcribed as 'kriegen' in Graubünden, indicating this might be the local replacement (heteronym).

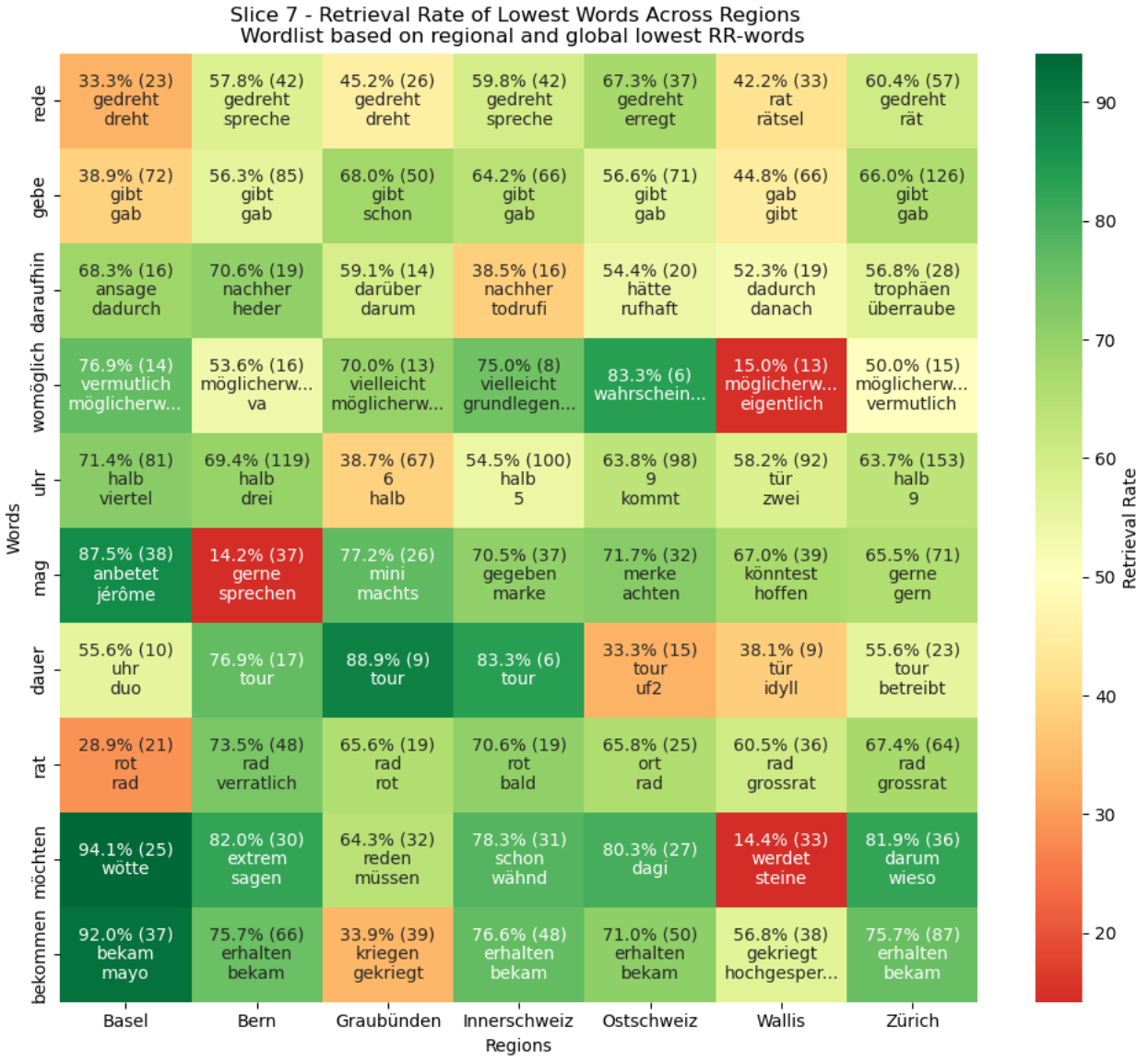

Figure 24: RR across Regions - Highest (1/2) Global-RR (out of Analyzed Word List)

For Figure 25, representing the eleven lowest RRs, it was found that these words are generally well transcribed with a high RR, except in the Basel, Graubünden, and Wallis regions. Notably, most of these words are specific to the Wallis dialect region, highlighting its well-known complexity and difficulty in comprehension [62]. A low RR in one region suggests that the region has a specific way of pronouncing the word. All heat maps covering the selected words can be found in Chapter 9.8.

Interpretation of Low RR Words in Wallis:

- **Voraus**: The compound word 'vorausgehen' can be pronounced as 'virrgaa' [62], which could explain the replacement word 'fit' for instances where the base sentences contained both 'voraus' and 'gehen' separately.
- **Beruf**: In the Wallis region, the 'u' is pronounced as 'ü', leading to the transcription of 'Prüf' or 'Prüfen' (cognate) instead of 'Beruf'.
- **Tür, Türen**: The word 'Tür' can be pronounced as 'Poort' (heteronym) [62], resulting in transcriptions as 'porte' or 'Bord'.
- **Spürbar**: 'Spürbar' was often transcribed as 'Spüren' in Wallis. The similarity between these two words is below 70% according to TheFuzz Similarity Calculation [46], so they are not counted as the same word.
- **Freien**: 'Freien' was mostly transcribed as 'frühen'. No specific pronunciation could be found upon manual research, but the 'ei' can be pronounced as 'ii' [62], which might have influenced this transcription.

Interpretation of Low RR Words in Graubünden:

- **Ziehen**: This word can be pronounced as 'Ziechen' (cognate) in various dialect regions, including Graubünden, as seen in the Schweizerisches Idiotikon [65]. 'Ziechen' was transcribed as 'suchten' or 'züchten'.
- **Bekommt**: 'Bekommt' was mostly transcribed as 'kriegen' (heteronym), as seen in the previous heat map 25.

Interpretation of Low RR Words in the Basel Region:

- **Locker**: Often transcribed as 'loco', indicating the 'er' sound could be pronounced as 'o' in some contexts in the Basel Region.

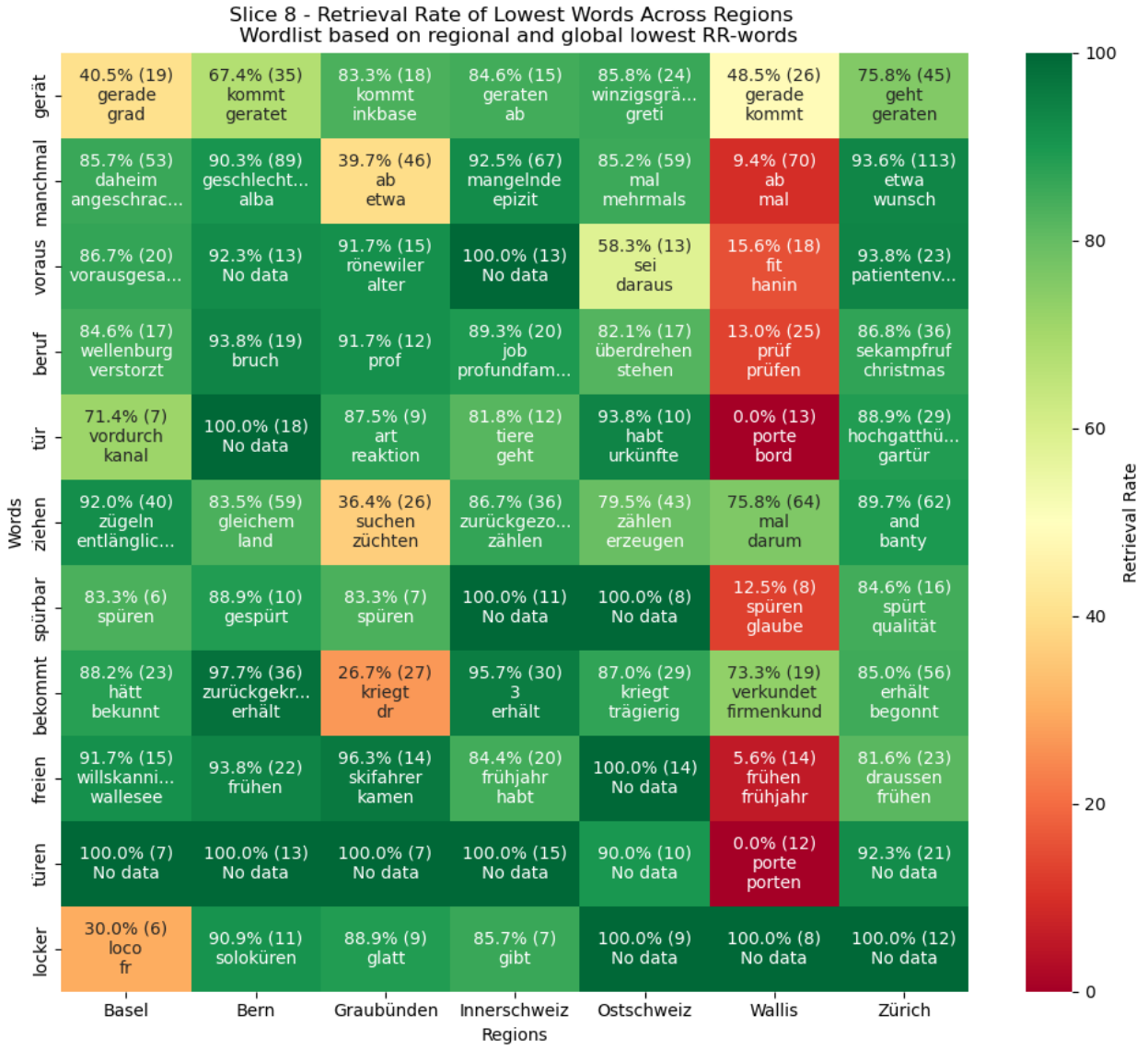

Figure 25: RR across Regions - Highest (2/2) Global-RR (out of Analyzed Word List)

Without the initial filtering out of preterite tense sentences during Chapter 4 as well as the ignoring of numbers during Word-Level RR computation (explained in Chapter 6.1.2), the words considered would be filled with preterite tense verbs and numbers. The rest of the produced heatmaps figure in Appendix A (Chapter 9.8 and represent the words in-between the ten highest and 20 lowest shown in this chapter.

7 Conclusion and Outlook

This chapter discusses the key findings from the analysis of speech speeds and dialectal variations through transcriptions across Swiss German dialects. In previous chapters, we discussed the need for an investigation with refined methodologies and expanded datasets and their implications for understanding regional linguistic variations. The following sections provide a comprehensive evaluation of the results, propose avenues for continued exploration in the field of linguistic research, and consider the implications of transcription practices on the interpretation of dialectal differences.

7.1 Speed Analysis

The analysis of speech speeds across various Swiss German dialects has yielded insightful yet diverse outcomes. Despite the meticulous methodologies applied, the results did not align uniformly with traditional beliefs about regional speech tempos, particularly the notion that Bernese dialects are notably slower than those of Zürich.

Dialect Region	Avg. WPS for each File		WPS with Mean[1]		WPS with Median[2]	
	STT4SG-350	SDS-200	STT4SG-350	SDS-200	STT4SG-350	SDS-200
Fastest Region	2.00	2.20	2.06	2.20	1.98	2.17
Slowest Region	1.94	1.94	1.94	1.83	1.90	1.81
Difference	0.06	0.26	0.12	0.37	0.08	0.36

[1] An average WPS value was calculated for each speaker, followed by the computation of the mean WPS across each dialect region.

[2] An average WPS value was calculated for each speaker, followed by the computation of the median WPS across each dialect region.

Table 28: Summary of the Variance in the Speed Analysis for Both Datasets

STT4SG-350 Dataset: In Table 28, it is evident that the STT4SG-350 dataset shows only minor speed differences across dialect regions. As described in Chapter 5.4, the reasons for this could be a low density of speakers per region, an error in the analysis, or that the differences are due to reading speed. If the latter is the case, the dataset may not be suitable for such an analysis. Another approach that could yield more accurate results would include more speakers from each region and additional recordings in the dataset. If minor differences in a speed analysis persist, it is more likely that these differences are related to reading speed.

SDS-200: When examining these results more closely in relation to the SDS-200 dataset, a pattern emerges: the results dataset shows greater variation than those from the STT4SG-350 dataset. This suggests that the SDS-200 dataset requires closer inspection for detailed analysis. The analysis indicates that Wallis speaks the fastest, followed by Zürich and Bern at the end, as seen in Chapter 5.4. The difference between Bern and Graubünden is minimal, whereas Innerschweiz shows a slightly larger gap. One possible explanation is that Bern, having many speakers in the SDS-200 dataset, achieves an accurate WPS measurement, whereas regions like Graubünden have too few records or speaker variance to provide a precise result. One possible solution would be to balance the dataset so that approximately the same number of speakers is available for each region. As seen in Chapter 5.1, accurate values can be obtained with around 100 speakers per region. How many recordings a single speaker contributes is irrelevant if the average speaking speed is first calculated per speaker

and then averaged for the dialect region. This approach could likely yield better results for the other dialect regions (Innerschweiz, Basel, Graubünden, and Ostschweiz) as well.

Measurement Methodology: The methodology for measuring speech speed primarily utilized the words per second metric, which is standard but might not capture the full complexity of speech patterns. Alternative metrics such as characters per second, syllable analysis, and analysis of pauses between words were identified as potential methods that could provide deeper insights into the nuances of dialectal speech tempo. These methods might help uncover finer distinctions that the basic WPS metric overlooks. But in hindsight, the different metrics most likely correlate with each other (including WPS) and would likely not differ significantly.
Given these considerations, the study suggests that further research with expanded datasets and alternative speech metrics could provide more definitive conclusions about speech speeds across Swiss German dialects. This future work would help validate the initial findings and refine the understanding of dialectal variations in Switzerland.

7.2 Linguistic Analysis

The linguistic analysis undertaken in this project has demonstrated significant versatility, accommodating a wide range of analytical approaches. This flexibility presents numerous possibilities, albeit with the challenge of maintaining focus amidst the expansive analytical scope. The results of our analysis reveal distinct variations between Standard German and the dialects of different regions. We introduced a method for identifying dialect-specific words and uncovering their unique pronunciations or alternative expressions. This approach enabled a comprehensive comparison of linguistic characteristics across the regions, providing a detailed understanding of regional linguistic diversity.

7.2.1 STT Model

An extensive search was conducted to identify a suitable STT model that aligned with the project goals, given the idea of using an existing Standard German model and the constraints on time that precluded developing our own model. The selected model, 'Whisper large-v2', is of high quality but introduces certain translation issues, which are detailed in Chapter 7.2.3. In the future, training a custom model or enhancing an existing one could ensure optimal behavior and transcription in contexts where the Whisper STT model translates rather than only transcribes, which is not aligned with the goals.

7.2.2 Dialect Variability Analysis

The analyses performed provide insights into dialect variability and constitute multiple approaches for further experiments in dialect research.

Regional Retrieval Rate Analysis The analysis of general RR resulted in comparing dialect regions and their differences relative to Standard German, using the RR algorithm and measurement developed in this project. Small differences were observed as most regions present a similar RR overall, except for the Wallis region, which showed a notably lower RR score, confirming the difficulty in understanding this dialect [62].

Dialect-Specific Words Research To identify dialect-specific words with a data-driven approach, we combined the Word-Level RR computation with an RR threshold. By altering parameters, we were able to generate multiple lists of dialect-specific words, which we attempted to

verify by comparing them with an existing dialect-specific word list. As the results were unsatisfying, a secondary verification approach was developed, setting the base for a larger manual verification process to find the ideal parameter set.

Refining the algorithm for identifying dialect-specific words, possibly by using a different metric than the current RR, and enhancing the algorithm for replacement word frequency could be explored. This improved algorithm could be used in a broader application to identify words with the most diverse replacements. As discussed in Chapter 2.2.1, words exist on a continuum from dialect to accent to standard language. Setting a hard threshold on a measurement like RR to produce a finite list of dialect-specific words might not be appropriate.

Additionally, the list of dialect-specific words from Dialektwörter.ch, which we used for comparison, has not been updated since 2007 [58]. Therefore, it can be assumed that the list is outdated and does not reflect current dialect word usage. A more recent word list could be used to verify the dialect-specific word list computation.

Lastly, the manual verification of the dialect-specific word lists could be done on a larger scale, examining output lists from various parameter combinations with a large number of evaluators to find the best parameter set.

Regional Words Comparison The Regional Words Comparison was conducted by utilizing the Word-Level RR on the regions individually to identify regional dialect-specific words. The approach of representing these words crossed with all regions as a heatmap allows for a detailed overview of the specific regional occurrences and alterations, along with deeper insight with the displaying of substitute words found in the transcriptions. This method could be applied more broadly to different selections to obtain insight into specific words of interest.

Sentence structure Analysis An analysis that we did not conduct due to time limitations is the comparison of the positional changes of words from Standard German across different dialect regions, potentially calculating a metric to determine which dialect region exhibits the most variation and also identifying the patterns of the variation.

7.2.3 Limitations

Whisper STT - Translations We observed that the Whisper large-v2 model occasionally translates words instead of merely transcribing them. This can be seen in 29 for the words 'Herunterladen' and 'Tönen'.

Word in base sentence	Spoken word	Transcribed word
Herunterladen	Abelade (Bernese dialect)	Herunterladen
Tönen	Tönen	Klingen

Table 29: Whisper Transcripts for Words 'Herunterladen' and 'Tönen'

Upon manual investigation, other examples were found where spoken words like 'Tönen' were still transcribed as 'Tönen' and not translated into 'Klingen', suggesting the presence of inconsistent translation behavior. Our subsequent analysis employs a word-by-word Levenshtein distance similarity to find matches and determine RRs and other measurements. This translation behavior can impact our results, highlighting a limitation of this approach.

We chose the Whisper model because a high RR is needed to have a solid base for the analysis, but this translation issue is something to watch out for and possibly a problem to tackle to get a

higher accuracy of analysis for further projects. It is important to note that no automatic detection of these translations can be performed, as the very base for the analysis is the mistranscription of certain words.

Future projects should implement custom model training focused on eliminating translations in the Whisper large-v2 model to address the issue of translations. Fine-tuning the model with datasets that emphasize transcription accuracy without translation and using domain-specific data where original language preservation is critical can help achieve this goal. This approach ensures the model consistently transcribes speech accurately, improving the reliability of our analyses.

Spacy Model - Morphology and Tense Identification The spacy model provides robust morphological analysis for text, accurately identifying features such as case, number, gender, and tense for tokens in sentences. However, we have observed that the model can occasionally misidentify the tense of verbs, particularly in specific contexts. An example of this issue is seen with the verb 'besassen', which, despite being in the preterite tense, is sometimes misidentified as present tense by the model.

Consider the following sentences:

- **Sentence 1**: 'Die anderen besassen alles was sie wollten.'
- **Sentence 2**: 'Die besten ihrer wenigen Chancen besassen diese Eigenschaft.'

When analyzed with the spacy model, the verb 'besassen' was detected as:

- **Sentence 1**: 'besassen' (Tense=Present)
- **Sentence 2**: 'besassen' (Tense=Present)

In Sentence 1, the verb 'wollten' is correctly identified as past tense, which causes the function to return `'true'` for Sentence 1. However, in Sentence 2, 'besassen' should have been identified as preterite tense, but the model tagged it as present tense, causing the function to return `'false'`. This inconsistency highlights the model's accuracy limitation, particularly with tense identification in specific contexts.

Enhancements to Address Tense Misidentification To improve the model's accuracy in tense identification, particularly for the German language, the following enhancements can be considered:

- **Context-Aware Training**: If the model can be trained specifically, enhancing the training data with more examples where verbs in preterite tense appear in various syntactic contexts will be beneficial. Including a diverse set of sentences where the distinction between past and present tense is clear will help the model learn these nuances better.
- **User Feedback and Detection**: Incorporating a system where users can provide feedback on incorrect tense identifications will also be crucial. This feedback can be used to refine and retrain the model iteratively, ensuring continuous improvement based on real-world usage.

Implementing these enhancements can reduce the occurrence of tense misidentifications and improve the overall reliability of the spacy model for morphological analysis in various linguistic contexts. This will ensure more accurate sentence parsing and better support for linguistic research and applications.

8 Directories

8.1 References

[1] S. Chopra and B. Duignan, *dialectology*, [Accessed June 6, 2024], Dec. 2011. [Online]. Available: https://www.britannica.com/science/dialectology.

[2] Cornelsen Verlag GmbH, *Heteronym*, [Accessed June 7, 2024], 2024. [Online]. Available: https://www.duden.de/rechtschreibung/Heteronym.

[3] Merriam-Webster, Inc., *Cognate*, [Accessed June 7, 2024], 2024. [Online]. Available: https://www.merriam-webster.com/dictionary/cognate.

[4] Q. D. Atkinson and R. D. Gray, "Curious Parallels and Curious Connections—Phylogenetic Thinking in Biology and Historical Linguistics," *Systematic Biology*, vol. 54, no. 4, pp. 513–526, Aug. 2005, [Accessed June 7, 2024], ISSN: 1063-5157. DOI: 10.1080/10635150590950317. eprint: https://academic.oup.com/sysbio/article-pdf/54/4/513/24198511/54-4-513.pdf. [Online]. Available: https://doi.org/10.1080/10635150590950317.

[5] H. Fraser, "A Framework for Deciding How to Create and Evaluate Transcripts for Forensic and Other Purposes," *Frontiers in Communication*, vol. 7, 2022, [Accessed June 7, 2024], ISSN: 2297-900X. DOI: 10.3389/fcomm.2022.898410. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fcomm.2022.898410.

[6] Q. D. Atkinson and R. D. Gray, "Curious Parallels and Curious Connections—Phylogenetic Thinking in Biology and Historical Linguistics," *Systematic Biology*, vol. 54, no. 4, pp. 513–526, Aug. 2005, [Accessed June 7, 2024], ISSN: 1063-5157. DOI: 10.1080/10635150590950317. eprint: https://academic.oup.com/sysbio/article-pdf/54/4/513/24198511/54-4-513.pdf. [Online]. Available: https://doi.org/10.1080/10635150590950317.

[7] K. Pfister and M. Zuber, "NLP- Project 3, Swiss german," Zurich University of Applied Sciences, Tech. Rep., 2023.

[8] A. Drigatti and L. Keller, "Quantitative Analyse des Sprechtempos verschiedener schweizerdeutscher Dialekte," Zurich University of Applied Sciences, Tech. Rep., 2023.

[9] "Heteronym," Karadeniz Technical University, Tech. Rep., Jan. 2021, [Accessed June 7, 2024]. [Online]. Available: https://www.coursehero.com/file/99167669/VERNACULAR-AND-STANDARD-LANGUAGEdocx/.

[10] C. Paonessa, Y. Schraner, J. Deriu, M. Hürlimann, M. Vogel, and M. Cieliebak, "Dialect Transfer for Swiss German Speech Translation," in *Findings of the Association for Computational Linguistics: EMNLP 2023*, H. Bouamor, J. Pino, and K. Bali, Eds., [Accessed June 4, 2024], Singapore: Association for Computational Linguistics, Dec. 2023, pp. 15 240–15 254. DOI: 10.18653/v1/2023.findings-emnlp.1018. [Online]. Available: https://aclanthology.org/2023.findings-emnlp.1018.

[11] LAWMEDIA AG, *Bevölkerung, Landessprachen und Religionen*, [Accessed June 3, 2024], 2020. [Online]. Available: https://www.ansiedlung-schweiz.ch/die-schweiz/bevoelkerung-landessprachen-und-religionen/.

[12] Prof. Dr. Adrian Leemann, *Dialäkt Äpps - Schweiz forscht*, [Accessed May 26, 2024], Mar. 2021. [Online]. Available: https://www.schweizforscht.ch/projekte/projektarchiv/dialaekt-aepps.

[13] Schweizerisches Idiotikon, [Accessed June 6, 2024], 2023. [Online]. Available: https://www.idiotikon.ch/woerterbuch/idiotikon-digital.

[14] Dictionary.com, *Language vs. Dialect Vs. Accent: Letting The Di erences Speak For Themselves*, [Accessed June 3, 2024], 2023. [Online]. Available: https://www.dictionary.com/e/language-vs-dialect-vs-accent/.

[15] D. Sharma and B. Rampton, "Lectal Focusing in Interaction: A New Methodology for the Study of Style Variation," *Journal of English Linguistics*, vol. 43, pp. 3–35, Jan. 2014, [Accessed June 6, 2024]. DOI: 10.1177/0075424214552131.

[16] R. Berthele, "The Extraordinary Ordinary: Re-engineering Multilingualism as a Natural Category," *Language Learning*, vol. 71, no. S1, pp. 80–120, 2021, [Accessed June 6, 2024]. DOI: https://doi.org/10.1111/lang.12407. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/lang.12407. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/lang.12407.

[17] Lingolia Team, *Perfekt – perfect tense in german grammar*, [Accessed May 24, 2024]. [Online]. Available: https://deutsch.lingolia.com/en/grammar/tenses/present-perfect.

[18] Cornelsen Verlag GmbH, *Perfekt*, [Accessed May 24, 2024], 2024. [Online]. Available: https://www.duden.de/rechtschreibung/Perfekt.

[19] W. Abraham and C. J. Conradie. Wien: Akademie Verlag, 2002, ISBN: 9789027297969 [Accessed June 3, 2024]. DOI: 10.1524/9783050084336.241. [Online]. Available: https://doi.org/10.1524/9783050084336.241.

[20] H. Fischer, "Präteritumschwund im Deutschen: Neue Erkenntnisse zu einem alten Rätsel," *Beiträge zur Geschichte der deutschen Sprache und Literatur*, vol. 143, no. 3, pp. 331–363, Sep. 1, 2021, ISSN: 1865-9373. DOI: 10.1515/bgsl-2021-0027. (visited on 06/01/2024).

[21] Dialog Sprachschule Wien, *Das Präteritum*, [Accessed June 4, 2024], Mar. 2023. [Online]. Available: https://www.dialog-wien.at/ueber-uns/blog/praeteritum/.

[22] A. Roelofs, "A spreading-activation theory of lemma retrieval in speaking," *Cognition*, vol. 42, no. 1, pp. 107–142, 1992, [Accessed June 4, 2024], ISSN: 0010-0277. DOI: https://doi.org/10.1016/0010-0277(92)90041-F. [Online]. Available: https://www.sciencedirect.com/science/article/pii/001002779290041F.

[23] R. Nordquist, *Lemmas Explained*, [Accessed June 4, 2024], Nov. 2019. [Online]. Available: https://www.thoughtco.com/what-is-a-lemma-1691108.

[24] J. M. Ph.D. and E. Kavlakoglu, *What are stemming and lemmatization?* [Accessed June 6, 2024], Dec. 2023. [Online]. Available: https://www.ibm.com/topics/stemming-lemmatization.

[25] M. Plüss, M. Hürlimann, M. Cuny, *et al.*, "Sds-200 : A swiss german speech to standard german text corpus," en, in *Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022)*, 13th Language Resources and Evaluation Conference (LREC), Marseille, France, 20-25 June 2022, European Language Resources Association, Jun. 2022, pp. 3250–3256. DOI: 10.21256/zhaw-26131. [Online]. Available: https://aclanthology.org/2022.lrec-1.347.

[26] M. Plüss, J. M. Deriu, Y. Schraner, *et al.*, "Stt4sg-350 : A speech corpus for all swiss german dialect regions," en, in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, [Accessed May 21, 2024], Association for Computational Linguistics, 2023, pp. 1763–1772. DOI: 10.18653/v1/2023.acl-short.150. [Online]. Available: https://digitalcollection.zhaw.ch/handle/11475/29062.

[27] H. Bredin, *pyannote.audio: Neural building blocks for speaker diarization: PyTorch based*, [Accessed May 5, 2024], 2020. [Online]. Available: https://github.com/pyannote/pyannote-audio.

[28] F. Wickham, *WAV, MP3, AAC, Or FLAC: Which Is Better?* [Accessed June 4, 2024]. [Online]. Available: https://www.funktasy.com/music-gear-tech/mp3-wav-flac-aac/.

[29] J. Robert, *Pydub*, [Accessed May 25, 2024], 2022. [Online]. Available: https://github.com/jiaaro/pydub/.

[30] Schweizerische Eidgenossenschaft, *Grenzwerte für Lärm*, [Accessed May 26, 2024], Oct. 2023. [Online]. Available: https://www.bafu.admin.ch/bafu/de/home/themen/laerm/fachinformationen/laermbelastung/grenzwerte-fuer-laerm.html.

[31] P. Schneider and C. Frei, "Automatic Identification of Swiss German Dialects using Large Language Models," [Accessed May 26, 2024], Zurich University of Applied Sciences, Jun. 2023. [Online]. Available: https://www.zhaw.ch/storage/engineering/institute-zentren/cai/studentische_arbeiten/Spring_2023/Spring23_BA_ciel_Dialect_Recognition_Swiss_German_Schneider_Frei.pdf.

[32] Dr. Amsel, F. W. Kaeding, "Zur Statistik des deutschen Wortschatzes," Statistisches Bundesamt, Wiesbaden, Tech. Rep., 2007, [Accessed May 19, 2024]. [Online]. Available: https://www.destatis.de/DE/Methoden/WISTA-Wirtschaft-und-Statistik/2007/08/statistik--wortschatz-082007.pdf?__blob=publicationFile.

[33] Prof. Dr. Anja Steinbeck, *Sprechgeschwindigkeit*, [Accessed May 19, 2024]. [Online]. Available: https://framenet-constructicon.hhu.de/diskurslinguistik/index.php?title=Sprechgeschwindigkeit.

[34] A. Kohler, *«Berner sind langsam»*, [Accessed May 22, 2024], Mar. 2014. [Online]. Available: https://www.nzz.ch/panorama/montagsklischee/berner-sind-langsam-ld.648074.

[35] NVIDIA, *NVIDIA FastConformer-Hybrid Large (de)*, [Accessed April 21, 2024], 2023. [Online]. Available: https://huggingface.co/nvidia/stt_de_fastconformer_hybrid_large_pc.

[36] OpenAI, *Whisper: General-purpose Speech Recognition*, [Accessed April 21, 2024], 2023. [Online]. Available: https://github.com/openai/whisper.

[37] doublex, *Dataset bias (" Translated by Amara.org Community")*, [Accessed June 4, 2024], Mar. 2023. [Online]. Available: https://github.com/openai/whisper/discussions/928.

[38] KaiserChr, *A possible solution to Whisper hallucination*, [Accessed June 4, 2024], Dec. 2022. [Online]. Available: https://github.com/openai/whisper/discussions/679.

[39] M. Honnibal and I. Montani, *spaCy: Industrial-strength Natural Language Processing in Python*, [Accessed May 5, 2024], 2020. [Online]. Available: https://spacy.io.

[40] B. McFee and S. Balke, *Librosa Library*, [Accessed April 16, 2024], 2023. [Online]. Available: https://librosa.org/.

[41] L. Vecchio, *Walliser haben in der Deutschschweiz das schnellste Mundwerk*, [Accessed May 26, 2024], Mar. 2024. [Online]. Available: https://www.20min.ch/story/dialektforschung-walliser-haben-in-der-deutschschweiz-das-schnellste-mundwerk-103062930.

[42] H. R. P. - Universität München, *Phonetische Analyse der Sprechgeschwindigkeit* (Forschungsberichte). Inst. für Phonetik und Sprachliche Kommunikation, Mar. 2001, [Accessed May 26, 2024]. [Online]. Available: https://www.phonetik.uni-muenchen.de/forschung/FIPKM/vol38/f38_hp_1.pdf.

[43] C. A. Gebhard, "Sprechtempo im Sprachvergleich: Eine Untersuchung phonologischer und kultureller Aspekte anhand von Nachrichtensendungen," [Accessed May 26, 2024], Ph.D. dissertation, Humboldt-Universität zu Berlin, Philosophische Fakultät II, Berlin, Jul. 2012. DOI: 10.18452/16567. [Online]. Available: http://dx.doi.org/10.18452/16567.

[44] B. Jakob, *Die Walliser schwatzen die Berner unter den Tisch*, [Accessed April 21, 2024], Jul. 2007. [Online]. Available: https://www.uniaktuell.unibe.ch/2007/die_walliser_schwatzen_die_berner_unter_den_tisch/index_ger.html.

[45] S. Grashchenko, "Levenshtein distance computation," *Baeldung*, 2023. [Online]. Available: https://www.baeldung.com/cs/levenshtein-distance-computation.

[46] SeatGeek, *Fuzzy string matching like a boss.* [Accessed April 21, 2024], 2014. [Online]. Available: https://github.com/seatgeek/thefuzz.

[47] J. M. Ph.D. and E. Kavlakoglu, *What is bag of words?* [Accessed June 6, 2024], Jan. 2024. [Online]. Available: https://www.ibm.com/topics/bag-of-words.

[48] NLTK Team, *Natural Language Toolkit*, [Accessed June 4, 2024], 2023. [Online]. Available: https://www.nltk.org/.

[49] Groupe Allo-Media, *text2num is a python package that provides functions and parser classes*, [Accessed June 6, 2024], 2018. [Online]. Available: https://github.com/allo-media/text2num.

[50] G. R. Sahani, "Understanding TF-IDF in NLP.," *Analytics Vidhya*, 2020, [Accessed June 6, 2024]. [Online]. Available: https://medium.com/analytics-vidhya/understanding-tf-idf-in-nlp-4a28eebdee6a.

[51] domcross on GitHub, *German STT Evaluation*, [Accessed April 21, 2024], 2022. [Online]. Available: https://github.com/domcross/german-stt-evaluation.

[52] Mozilla Corporation, *Mozilla Common Voice*, [Accessed June 6, 2024], 2018. [Online]. Available: https://commonvoice.mozilla.org/.

[53] V. Pratap, Q. Xu, A. Sriram, G. Synnaeve, and R. Collobert, "Mls: A large-scale multilingual dataset for speech research," in *Interspeech 2020*, ser. interspeech$_2$020, ISCA, Oct. 2020. DOI: 10.21437/interspeech.2020-2826. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2020-2826.

[54] C. Wang, M. Rivière, A. Lee, *et al.*, *VoxPopuli: A Large-Scale Multilingual Speech Corpus for Representation Learning, Semi-Supervised Learning and Interpretation*, [Accessed June 7, 2024], 2021. arXiv: 2101.00390 [cs.CL].

[55] VoxForge, *VoxForge Dataset*, [Accessed June 6, 2024], 2006. [Online]. Available: https://www.voxforge.org/.

[56] D. Bermuth, A. Poeppel, and W. Reif, "Scribosermo: Fast Speech-to-Text models for German and other Languages," 2021, [Accessed April 21, 2024]. [Online]. Available: https://gitlab.com/Jaco-Assistant/Scribosermo.

[57] Silero AI Team, *Silero Speech-To-Text Models*, [Accessed April 21, 2024], 2023. [Online]. Available: https://pytorch.org/hub/snakers4_silero-models_stt/.

[58] B. Nussbaumer, *Sammlung Schweizerdeutscher Dialektwörter und -begri e*, [Accessed April 21, 2024], 2000. [Online]. Available: https://dialektwoerter.ch/.

[59] Pons, *Schweizerdeutsch – Unsere Top-10-Wortliste*, [Accessed April 21, 2024]. [Online]. Available: https://de.pons.com/p/wissensecke/wortschatz-to-go/schweizerdeutsch.

[60] C.-P. Pohl, *Schweizerdeutsch – Diese Begri e musst du als Deutscher kennen*, [Accessed April 21, 2024], Nov. 2021. [Online]. Available: https://auswanderluchs.ch/schweizer-begriffe/.

[61] R. Joseph, "Grid Search for model tuning," *Towards Data Science*, Dec. 2018, [Accessed June 6, 2024]. [Online]. Available: https://towardsdatascience.com/grid-search-for-model-tuning-3319b259367e.

[62] V. Schmid, *Deutsche mundarten*, [Accessed May 24, 2024], 2019. [Online]. Available: https://www.walliserdialekt.ch/walliser-dialekt.

[63] T. Kanstrén, "A Look at Precision, Recall, and F1-Score," *Towards Data Science*, 2020, [Accessed June 6, 2024]. [Online]. Available: https://towardsdatascience.com/a-look-at-precision-recall-and-f1-score-36b5fd0dd3ec.

[64] SRF Schweizer Radio und Fernsehen, *Das Schweizerdeutsche «AU» ist für Deutsche Expats ein Rätsel*, [Accessed May 27, 2024], 2024. [Online]. Available: https://www.srf.ch/audio/schwiiz-und-duetlich/das-schweizerdeutsche-au-ist-fuer-deutsche-expats-ein-raetsel?id=12565193.

[65] Schweizerisches Idiotikon, [Accessed May 26, 2024], 2023. [Online]. Available: https://digital.idiotikon.ch/idtkn/id17.htm#!page/170875/mode/1up.

[66] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu (ACL '02). Philadelphia, Pennsylvania: Association for Computational Linguistics, 2002, pp. 311–318. DOI: 10.3115/1073083.1073135. [Online]. Available: https://doi.org/10.3115/1073083.1073135.

[67] J. Holdsworth, *What is NLP?* [Accessed June 7, 2024], 2024. [Online]. Available: https://www.ibm.com/topics/natural-language-processing.

8.2 List of Figures

8.3 List of Tables

8.4 Glossary

BLEU Score BLEU (Bilingual Evaluation Understudy) [66] Score is an algorithm for evaluating the quality of text that has been machine-translated from one natural language to another. Quality is considered to be the correspondence between a machine's output and that of a human: 'the closer a machine translation is to a professional human translation, the better it is' – this is the central idea behind BLEU. BLEU was one of the first metrics to claim a high correlation with human judgments of quality and remains one of the most popular automated and inexpensive metrics. This thesis evaluates the effectiveness of automatic speech recognition systems developed to recognize and transcribe Swiss German dialects.

Bag of Words An unstructured representation of a text or a document that describes the occurrence of words within it, ignoring context and word order [47].

Levenshtein Distance The Levenshtein distance between two strings is defined as the minimum number of single-character edit operations required to transform one string into the other. These operations include insertions, deletions, or substitutions. As explained in the Baeldung article [45], the Levenshtein distance between two strings a and b is given by $\text{lev}_{a,b}(|a|, |b|)$, where:

$$\text{lev}_{a,b}(i,j) = \begin{cases} \max(i,j) & \text{if } \min(i,j) = 0, \\ \min \begin{cases} \text{lev}_{a,b}(i-1,j) + 1 \\ \text{lev}_{a,b}(i,j-1) + 1 \\ \text{lev}_{a,b}(i-1,j-1) + 1_{(a_i \neq b_j)} \end{cases} & \text{otherwise.} \end{cases}$$

In this equation:

- The indicator function 1 is equal to 0 if $a_i = b_j$ and 1 otherwise.
- The length of the string a is denoted by $|a|$.
- $\text{lev}_{a,b}(i,j)$ represents the distance between the first i characters of a and the first j characters of b.

The formula has two main parts:

- The first part accounts for the number of insertion or deletion steps needed to transform a prefix into an empty string or vice versa.
- The second part is a recursive expression of deletions, insertions and substitutions.

For more information, refer to the Baeldung article about Levenshtein Distance Computation [45].

Stopwords Stopwords are common words that usually carry less semantic weight and are often filtered out to improve the efficiency and relevance of text-processing tasks. Examples of stopwords in English include words like 'and', 'the', and 'is'. In German, stopwords include words such as 'und' (and), 'der' (the - masculine), 'die' (the - feminine), 'das' (the - neuter), 'ist' (is), and 'zu' (to). By ignoring these words, the analysis can concentrate on more significant words that contribute more meaningfully to the sentence structure and content. This technique helps reduce noise and improve the RR calculations' precision.

Grid Search Grid Search is a hyperparameter optimization technique used to improve the performance of various analytical models, not limited to machine learning. It involves systematically searching through a specified subset of parameters to find the combination that yields the best performance based on a given evaluation metric [61]. In our analysis, we use Grid Search to identify the optimal parameters, ensuring that the settings provide the most accurate and effective results.

Precision Precision is a measure of how many positive predictions made by a model are correct. It is calculated as the number of true positive results divided by the total number of positive results predicted by the model [63]. The formula for Precision is:

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}} \tag{4}$$

Recall Recall, also called Sensitivity, is a measure of how many positive cases were correctly found made by a model. It is calculated as the number of true positive cases found divided by the total number of actual positives [63]. The formula for Recall is:

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}} \tag{5}$$

F1 Score Combination of Precision and Recall, often described as their harmonic mean. It balances both the precision and recall by requiring both to have a higher value for the F1 value to rise [63]. The formula for the F1 score is:

$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{6}$$

TF-IDF TF-IDF stands for Term Frequency-Inverse Document Frequency. It is a numerical statistic used to reflect how important a word is in a document relative to a collection of documents (corpus). The term frequency (TF) measures how frequently a word appears in a document, increasing the score, while the inverse document frequency (IDF) measures how important a word is by considering how common or rare it is across all documents in the corpus, decreasing the score if the word is common [50]. In our analysis, TF-IDF can help weigh words that are uncommon when looking for replacement words, enhancing the focus on distinctive terms that are more likely to be dialect-specific.

Natural Language Processing (NLP) Natural Language Processing (NLP) is a multidisciplinary field that intersects computer science, artificial intelligence, and linguistics, focusing on the interaction between computers and human language. NLP involves programming computers to process and analyze large volumes of natural language data. Key tasks within NLP include text translation, sentiment analysis, speech recognition, and chatbot functionality. These techniques enable the understanding, interpretation, and generation of human language, making them valuable for applications ranging from customer service automation to aiding in complex decision-making processes. By leveraging algorithms and machine learning, NLP understands the content in text and speech and grasps the nuances, intent, and sentiment behind the language used [67].

8.5 List of Abbreviations

ASR	Automatic Speech Recognition
BLEU	Bilingual Evaluation Understudy
CAI	Central Artificial Intelligence
CER	
CPS	Characters per Second
EMNLP	Empirical Methods in Natural Language Processing
FLAC	Free Lossless Audio Codec
LPS	Laute pro Sekunde
MCV	Mozilla's Common Voice
MLS	Multilingual LibriSpeech
MP3	MPEG Audio Layer 3
NLP	Natural Language Processing
NLTK	Natural Language Toolkit
NZZ	Neue Züricher Zeitung
RR	Retrieval Rate
STT	Speech-To-Text
TF-IDF	term frequency-inverse document frequency
TSV	Tab-separated Values
TTS	Text-To-Speech
VAD	Voice Activity Detection
WAV	Waveform Audio File Forma
WER	Word Error Rate
WPS	Words per Second
ZHAW	Züricher Hochschule für angewandte Wissenschaften

9 Appendix A: Detailed Analysis Results and Figures

9.1 Data Management Overview

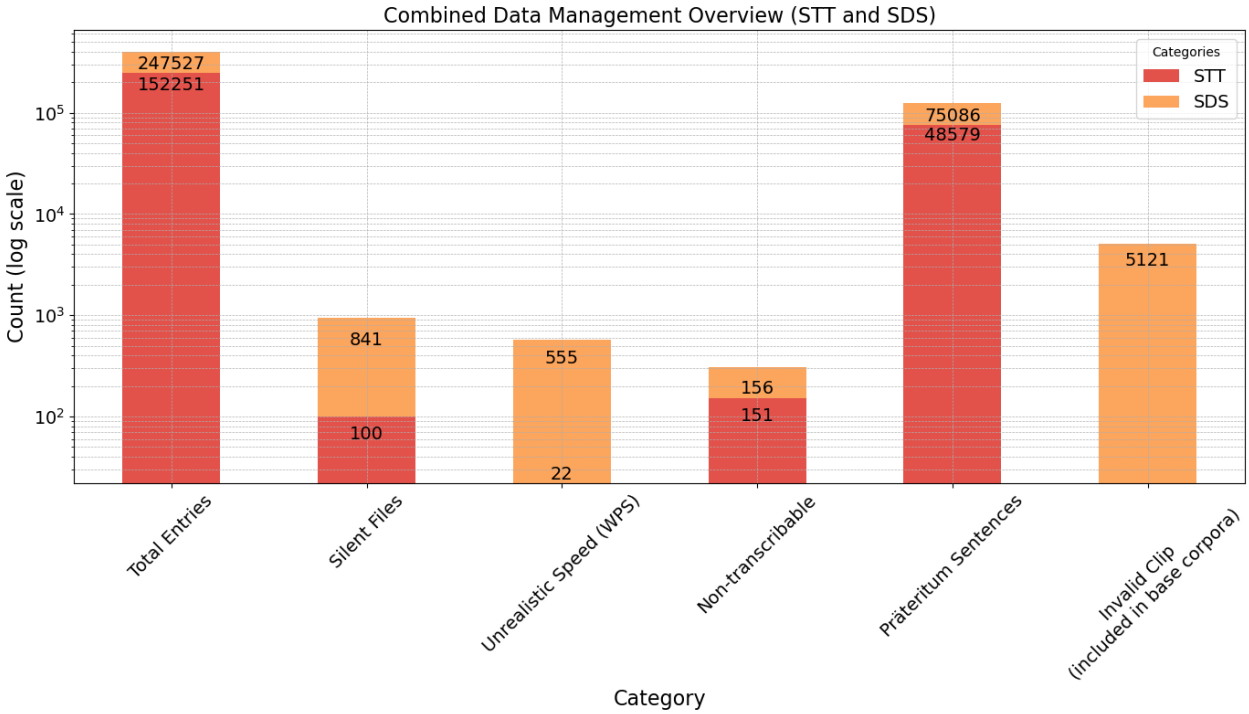

Figure 26: Data Management Overview Combined

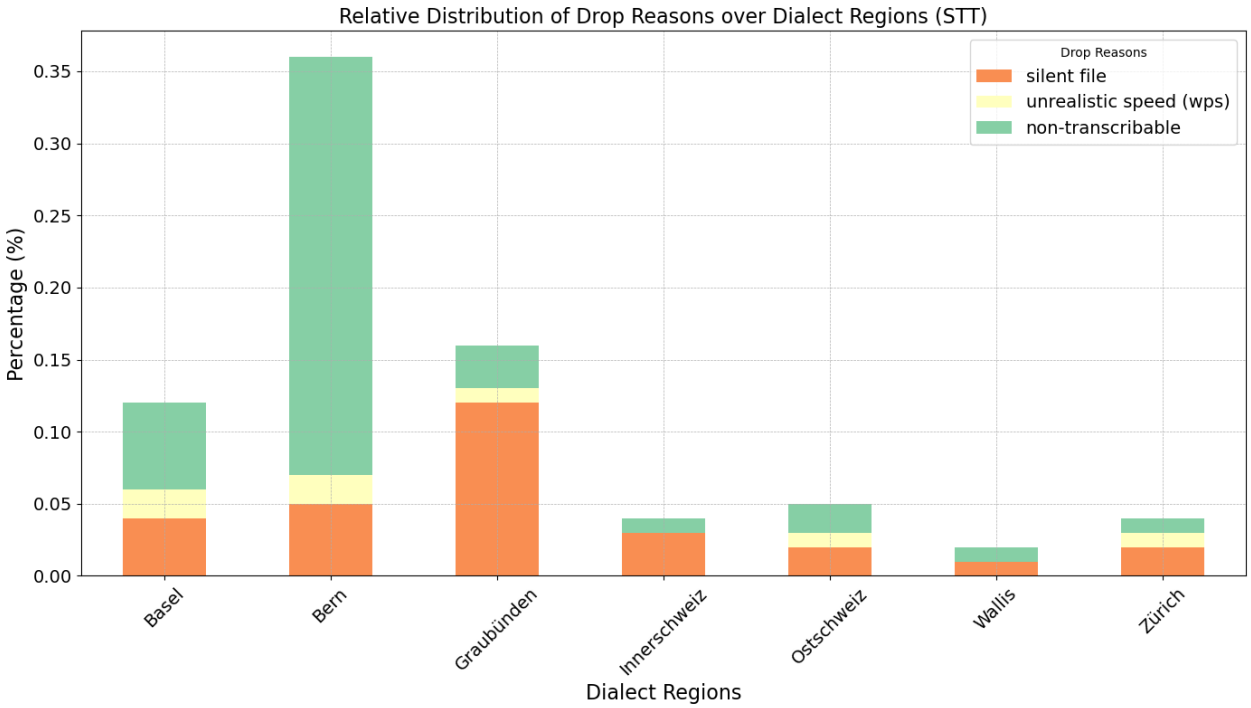

Figure 27: Data Management relative Drop Reasons Regions - STT

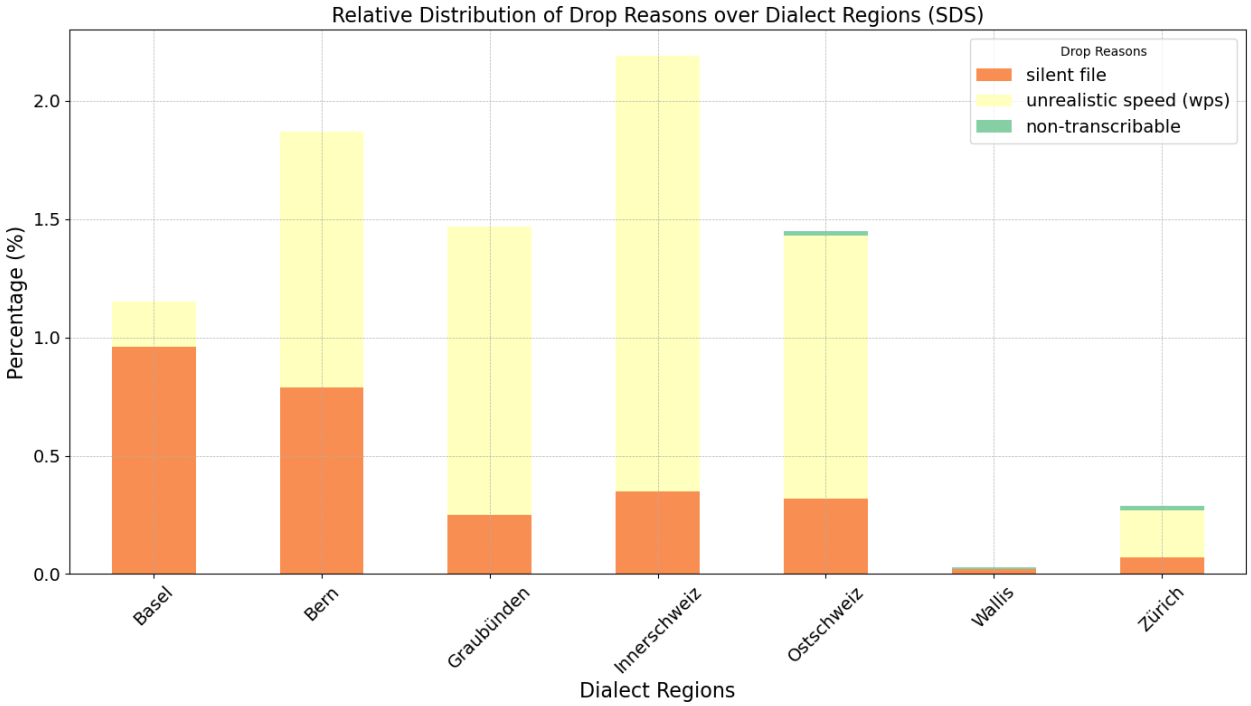

Figure 28: Data Management relative Drop Reasons Regions - SDS

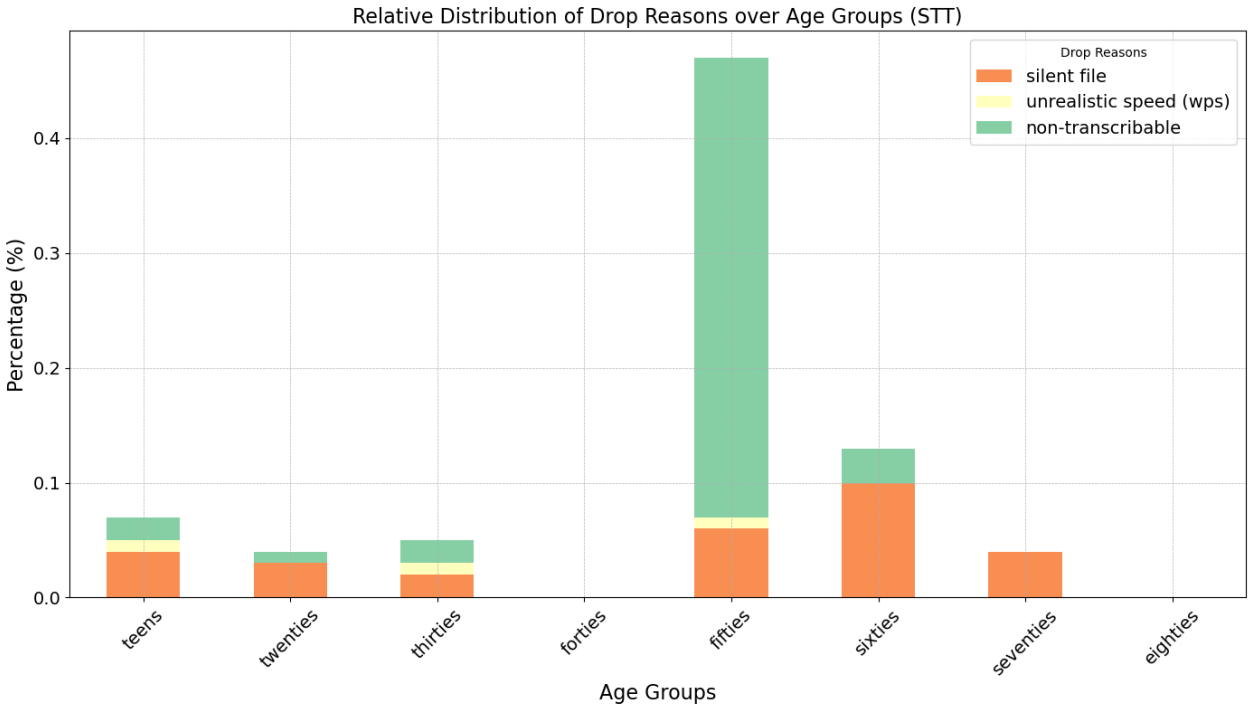

Figure 29: Data Management relative Drop Reasons Age - STT

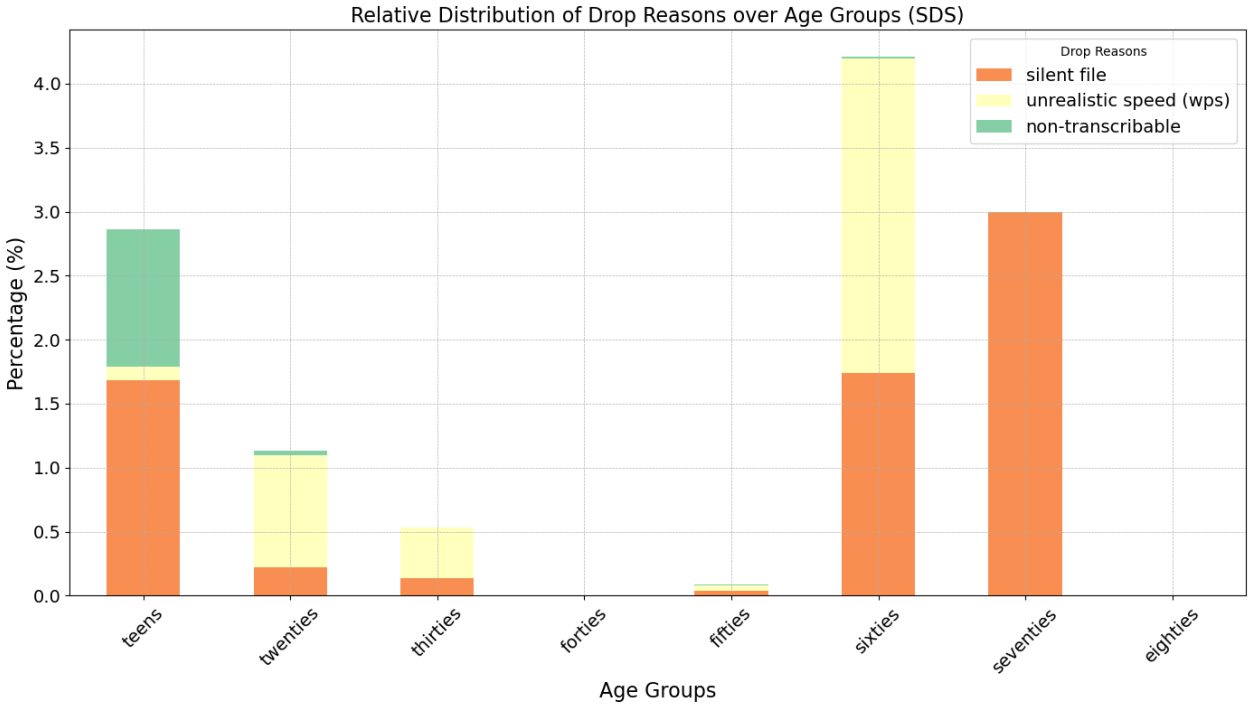

Figure 30: Data Management relative Drop Reasons Age - SDS

9.2 Speaker Distributions

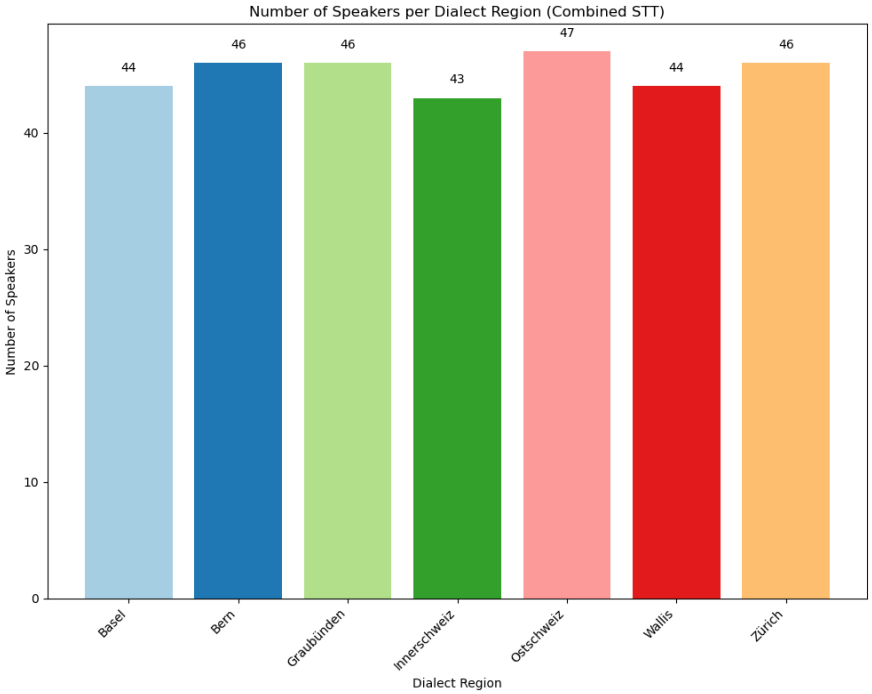

Figure 31: Speaker Distribution STT4SG-350 Dataset

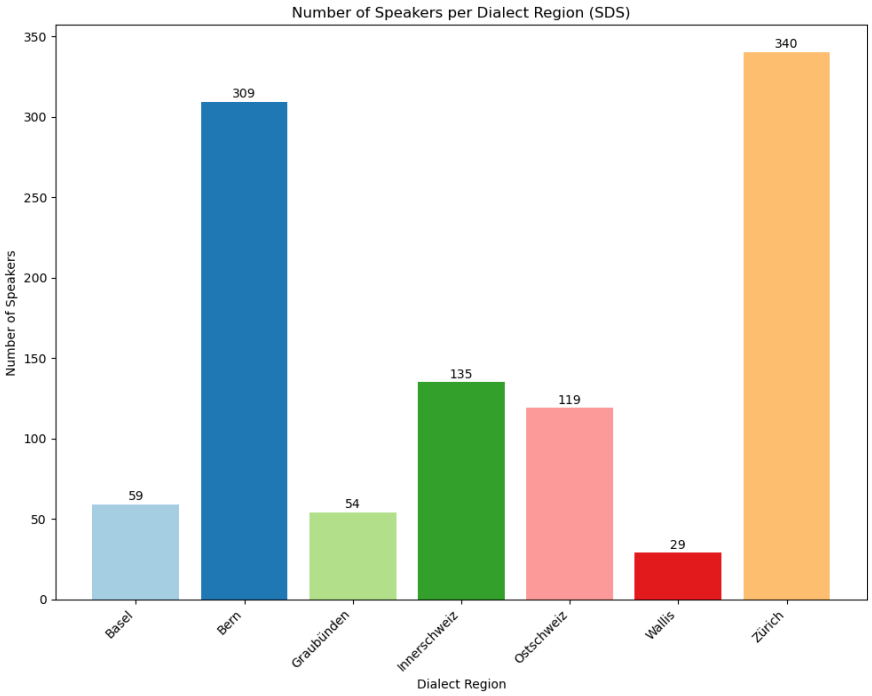

Figure 32: Speaker Distribution SDS-200 Dataset

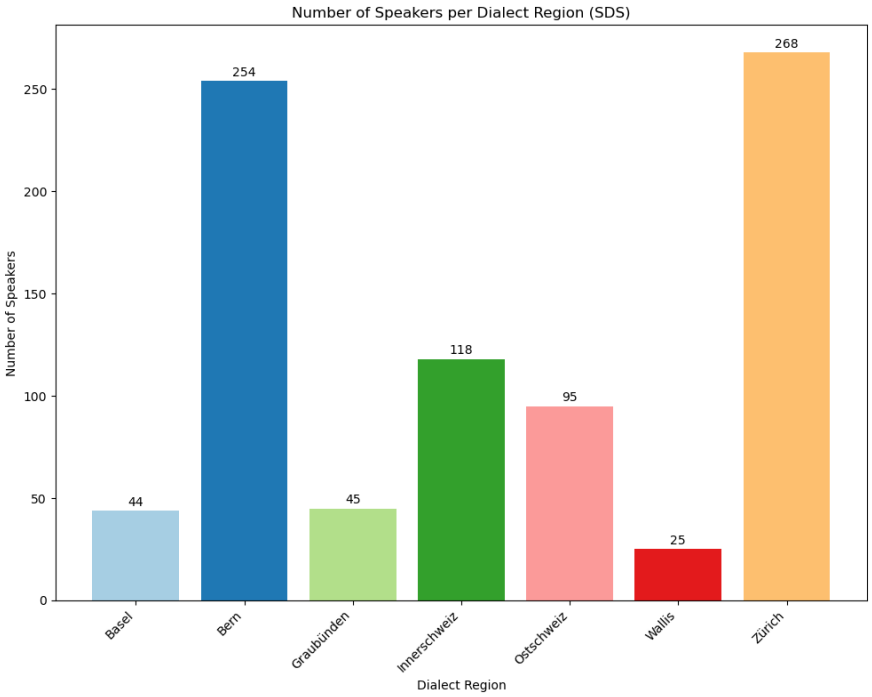

Figure 33: Speaker Distribution after SDS-200 Sampling

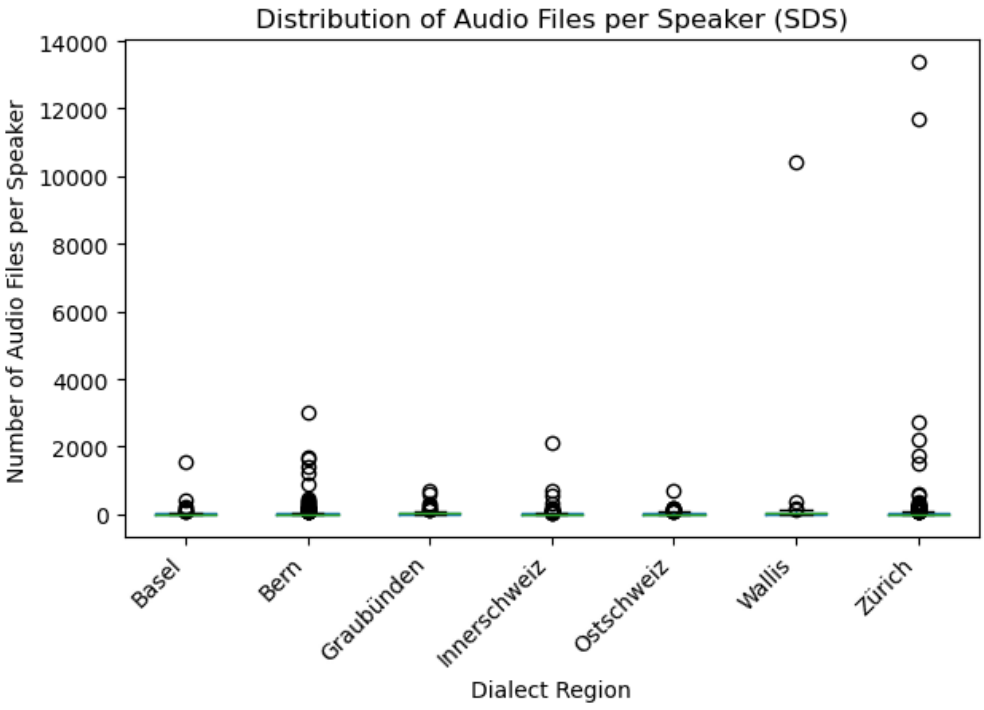

Figure 34: Audio File Distribution of the Dataset SDS-200

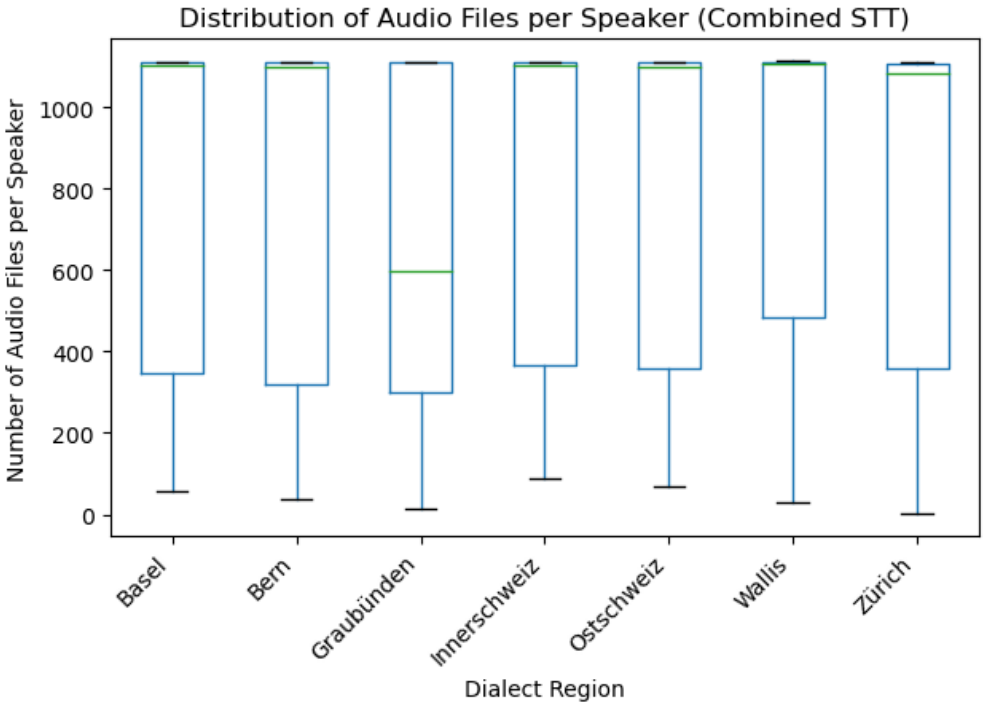

Figure 35: Audio File Distribution of the Dataset STT4SG-350

9.3 Dialect-Specific Distribution of Speech Speeds

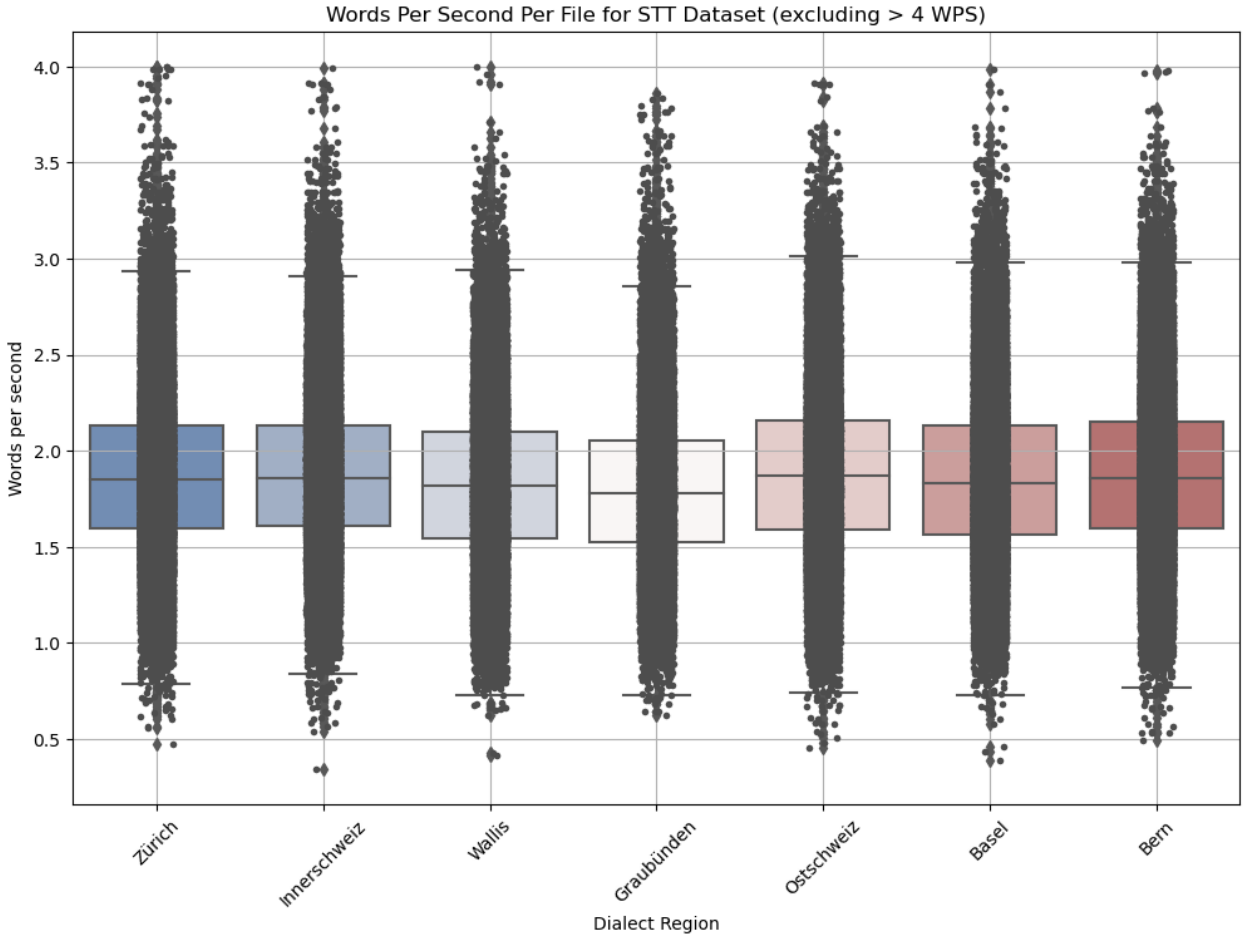

Figure 36: WPS per File for each Dialect Region of the STT4SG-350 Dataset

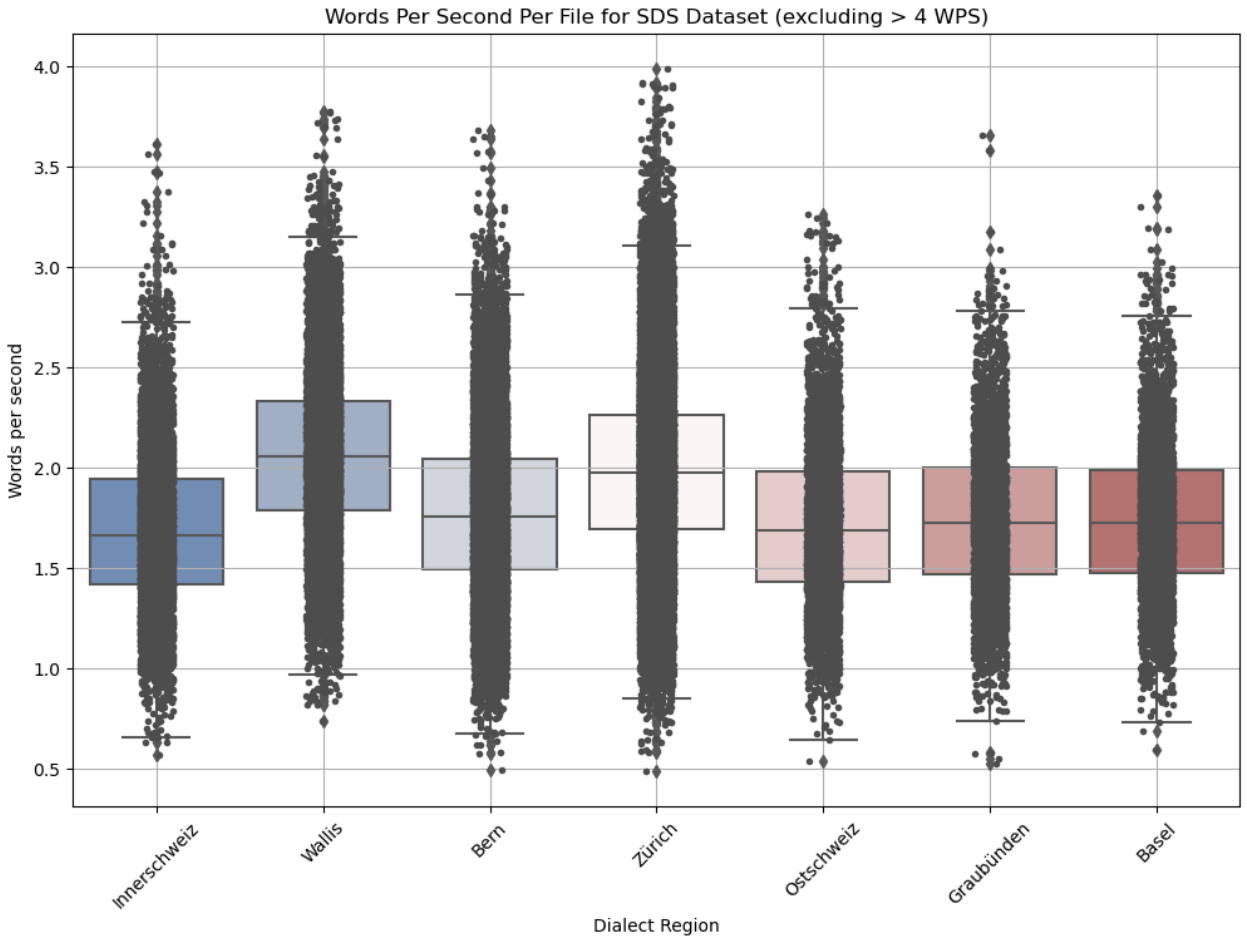

Figure 37: WPS per File for each Dialect Region of the SDS-200 Dataset

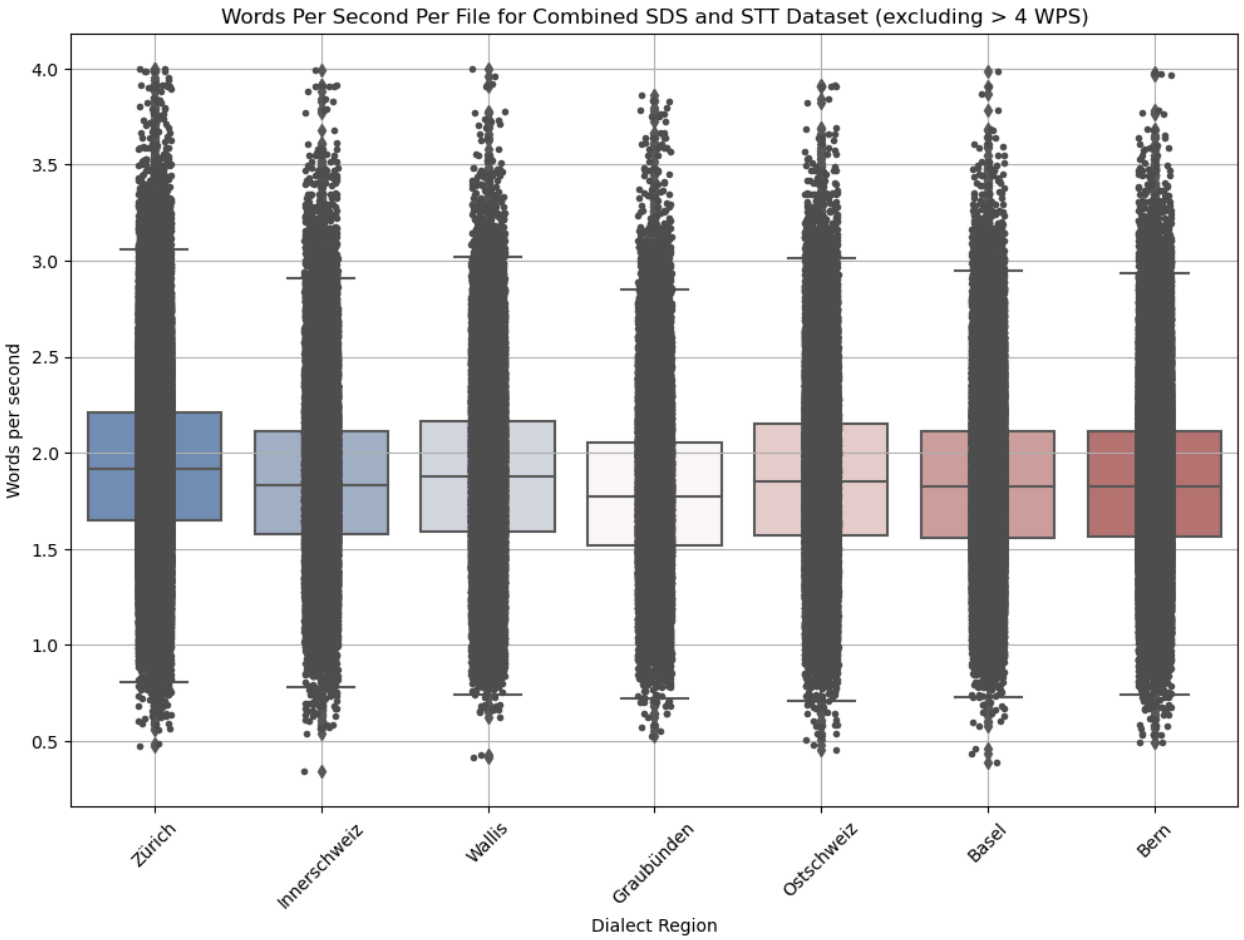

Figure 38: WPS per File for each Dialect Region of Both Datasets

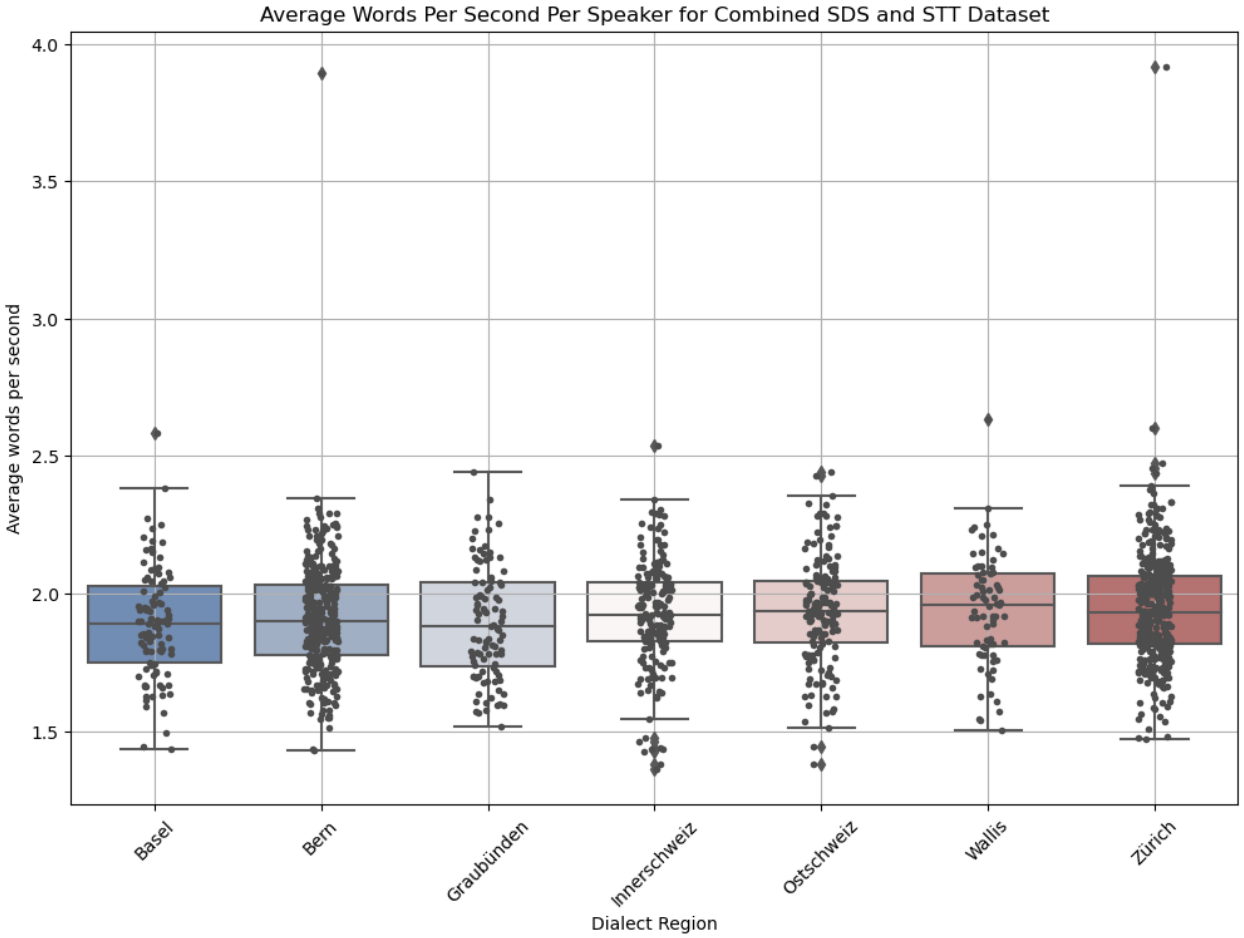

Figure 39: Average WPS per Speaker for Both Datasets

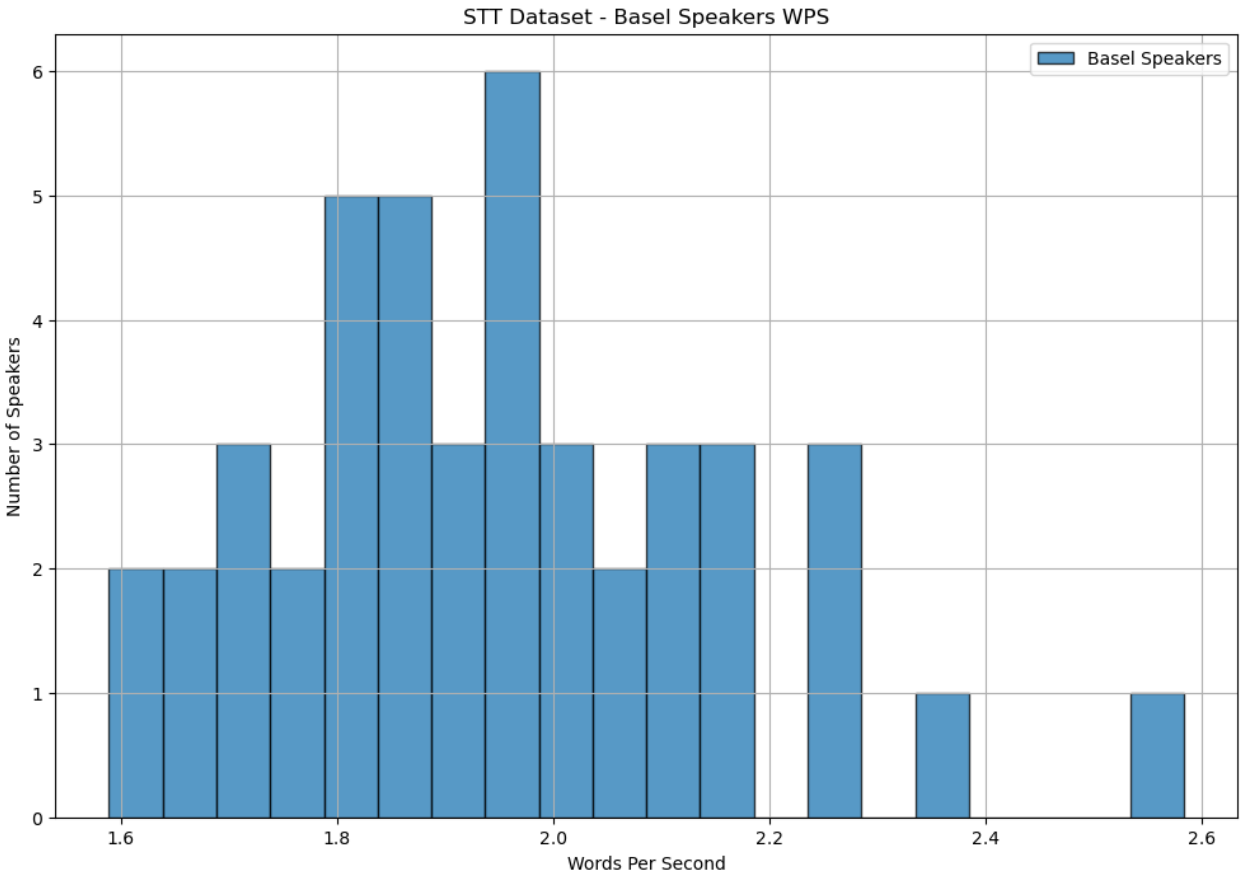

Figure 40: Histogram STT4SG-350 Region Basel

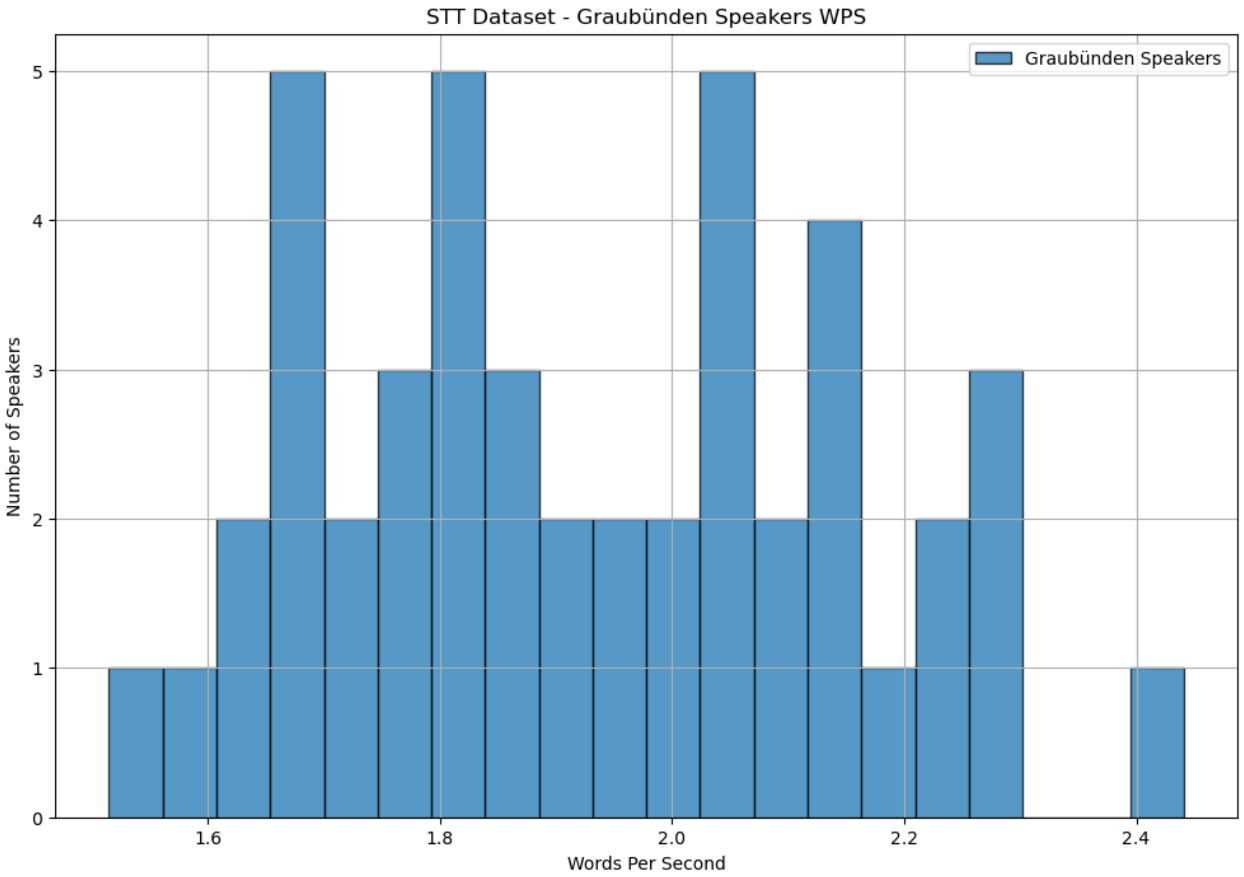

Figure 41: Histogram STT4SG-350 Region Graubünden

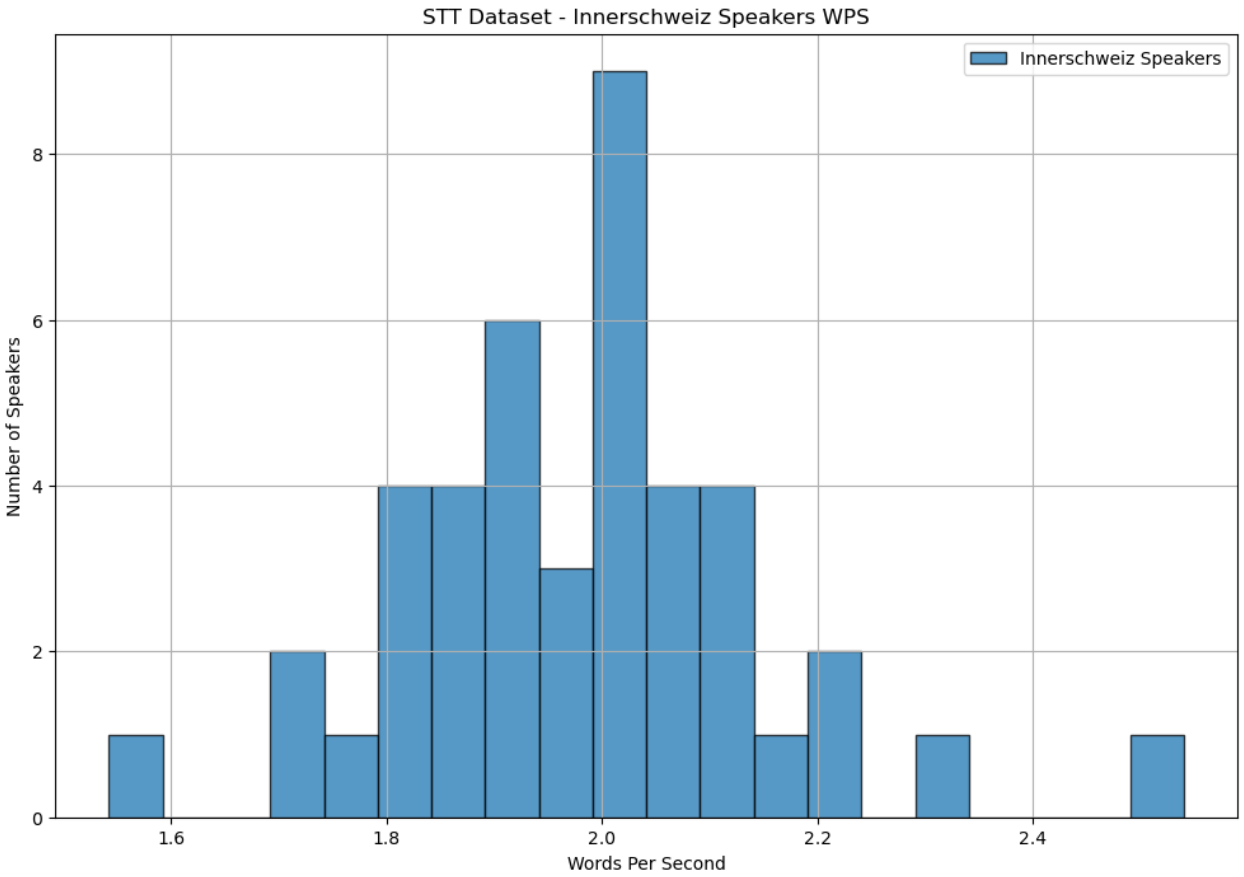

Figure 42: Histogram STT4SG-350 Region Innerschweiz

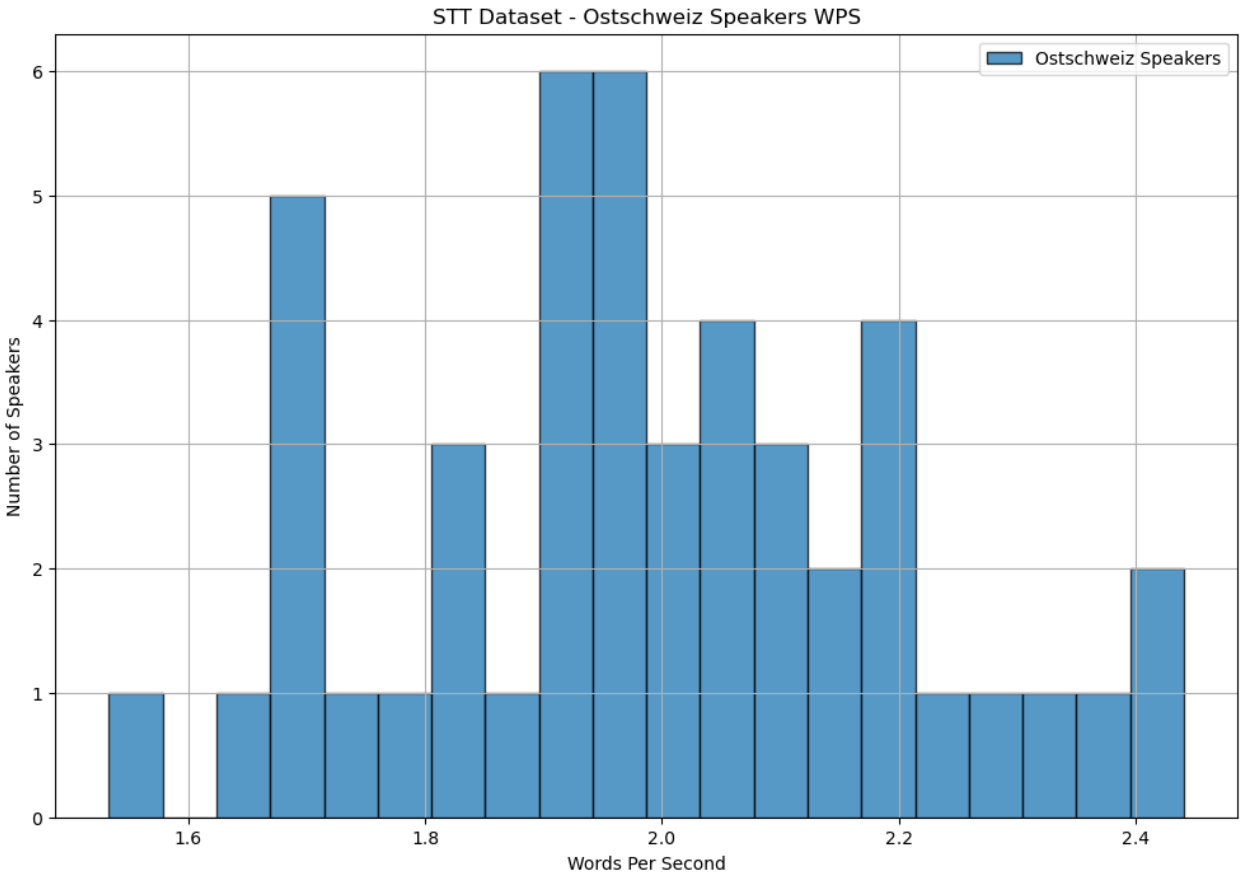

Figure 43: Histogram STT4SG-350 Region Ostschweiz

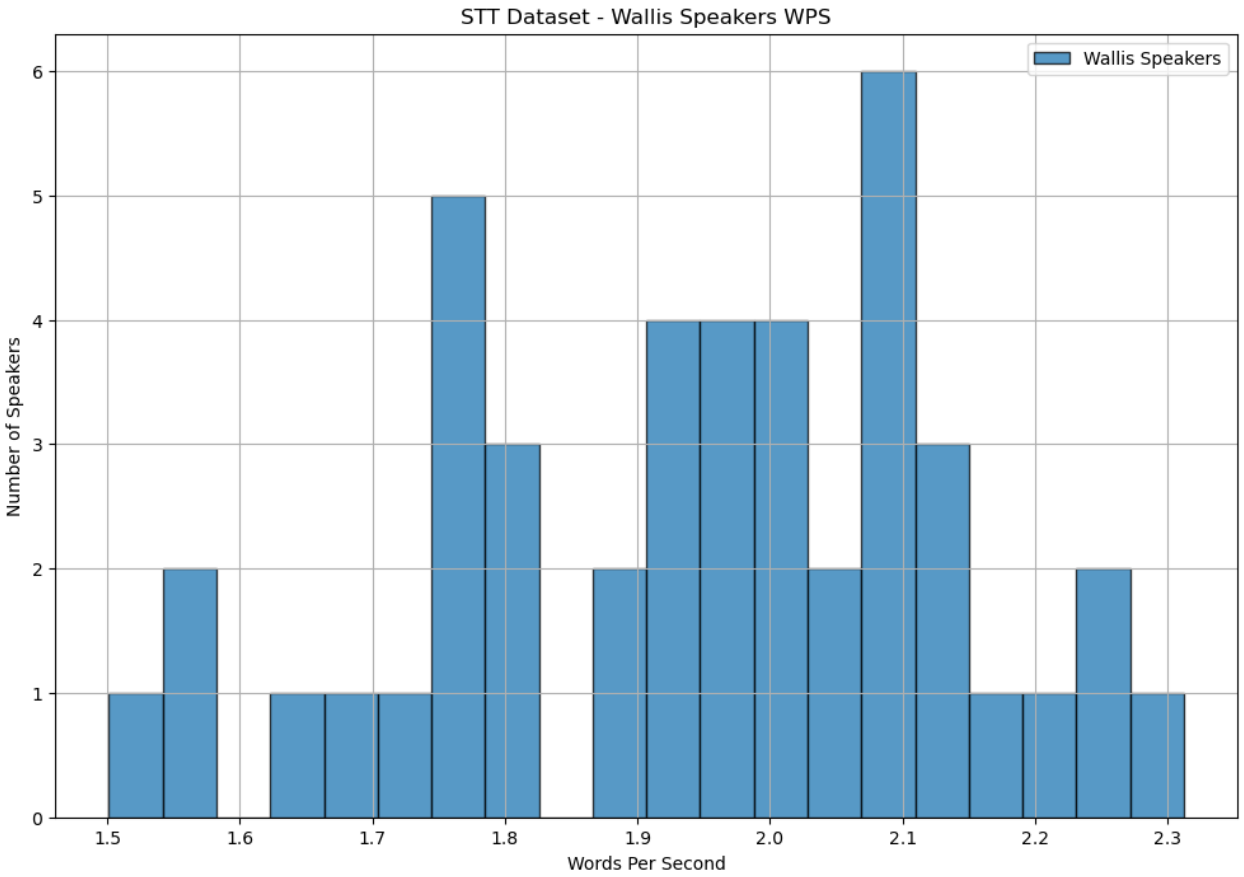

Figure 44: Histogram STT4SG-350 Region Wallis

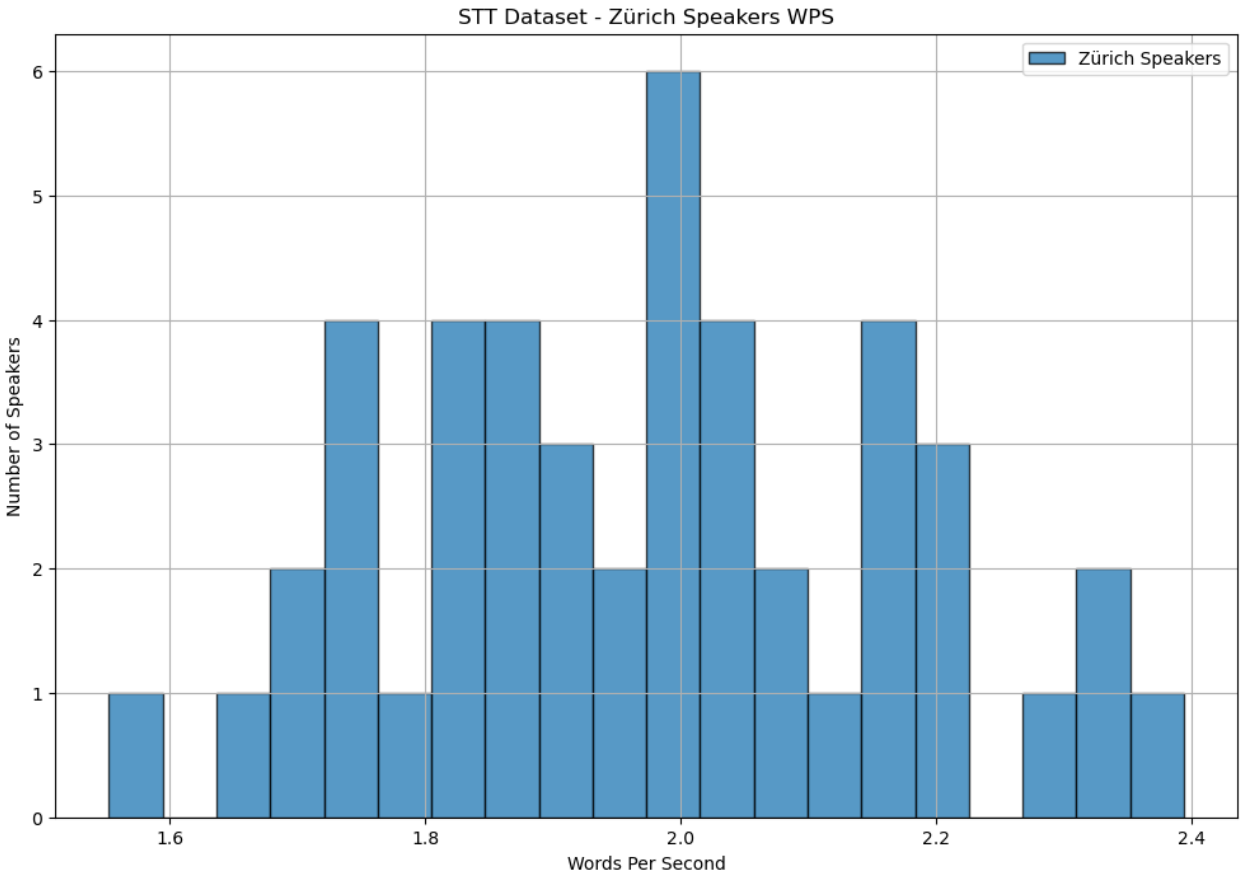

Figure 45: Histogram STT4SG-350 Region Zürich

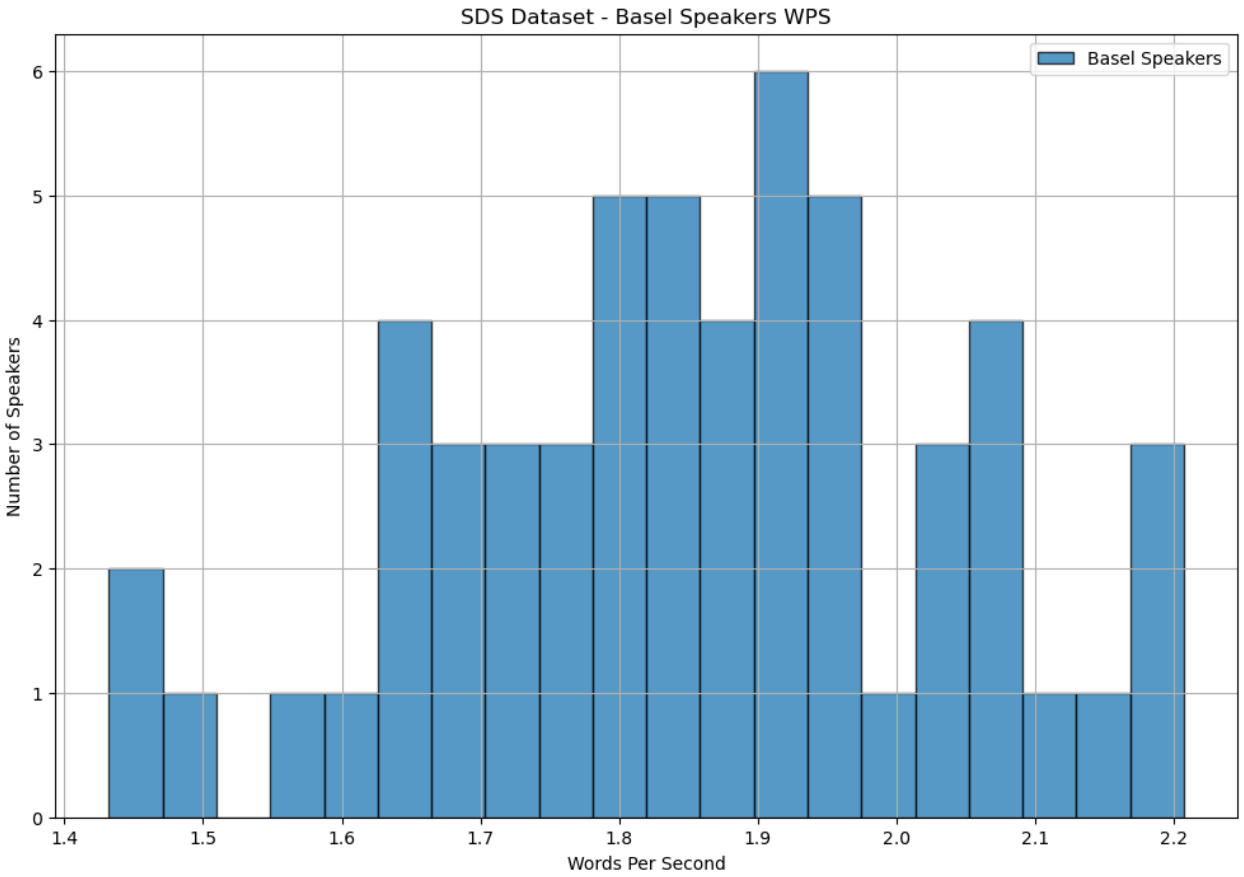

Figure 46: Histogram SDS-200 Region Basel

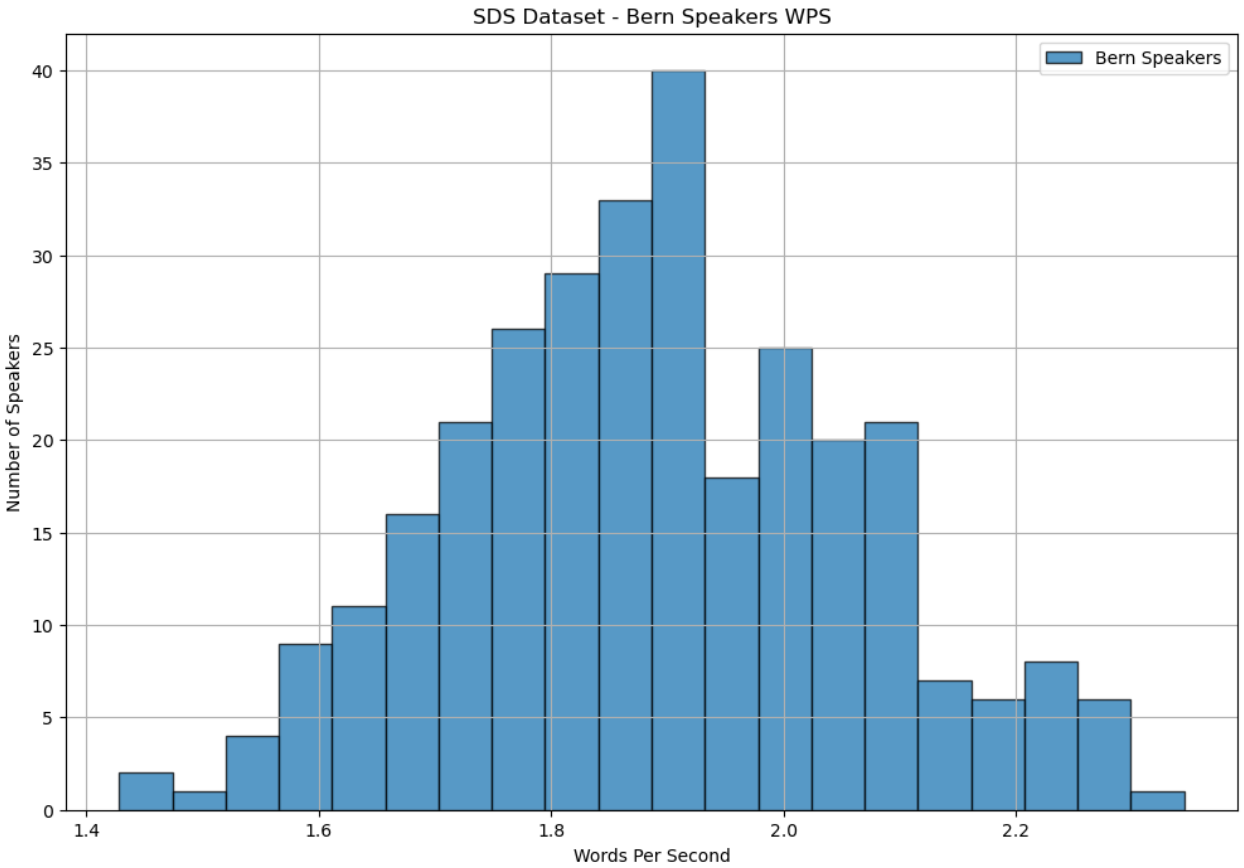

Figure 47: Histogram SDS-200 Region Bern

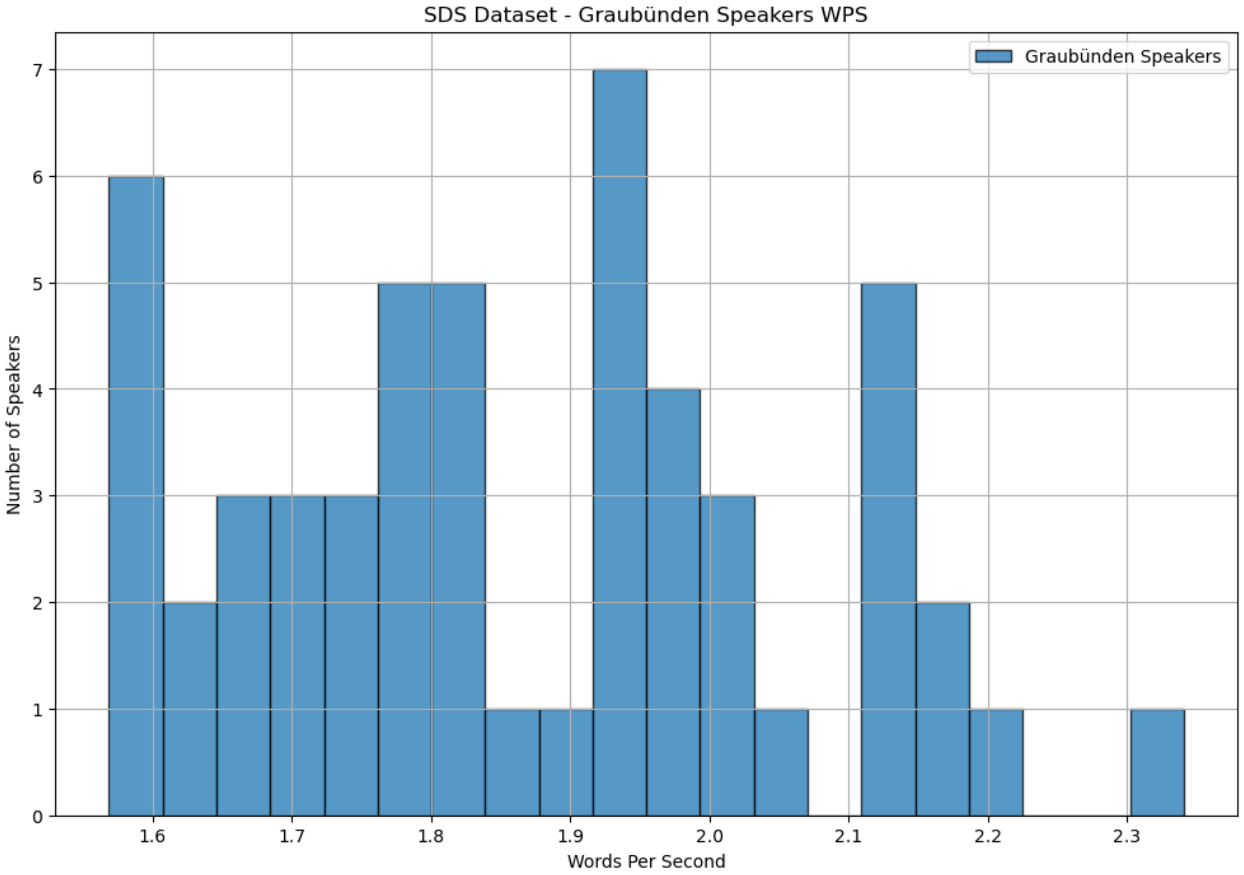

Figure 48: Histogram SDS-200 Region Graubünden

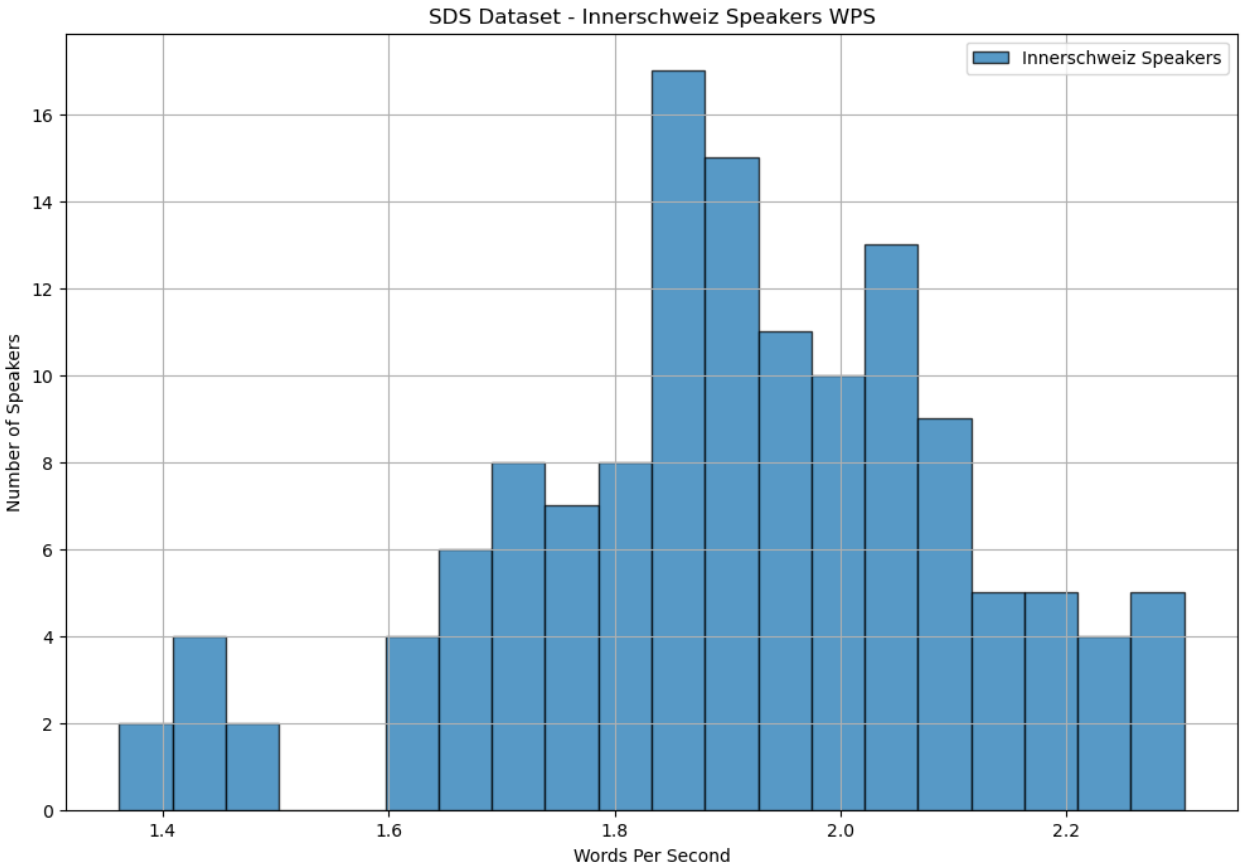

Figure 49: Histogram SDS-200 Region Innerschweiz

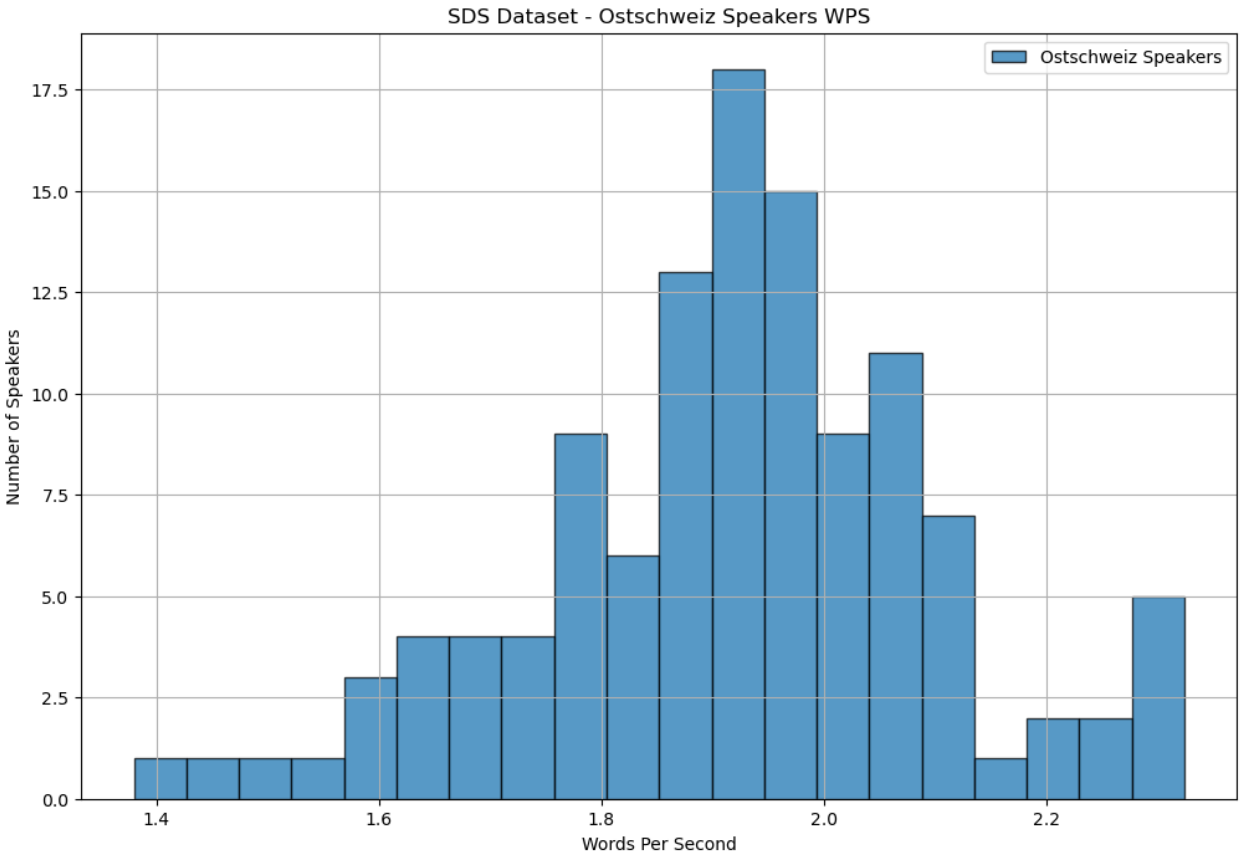

Figure 50: Histogram SDS-200 Region Ostschweiz

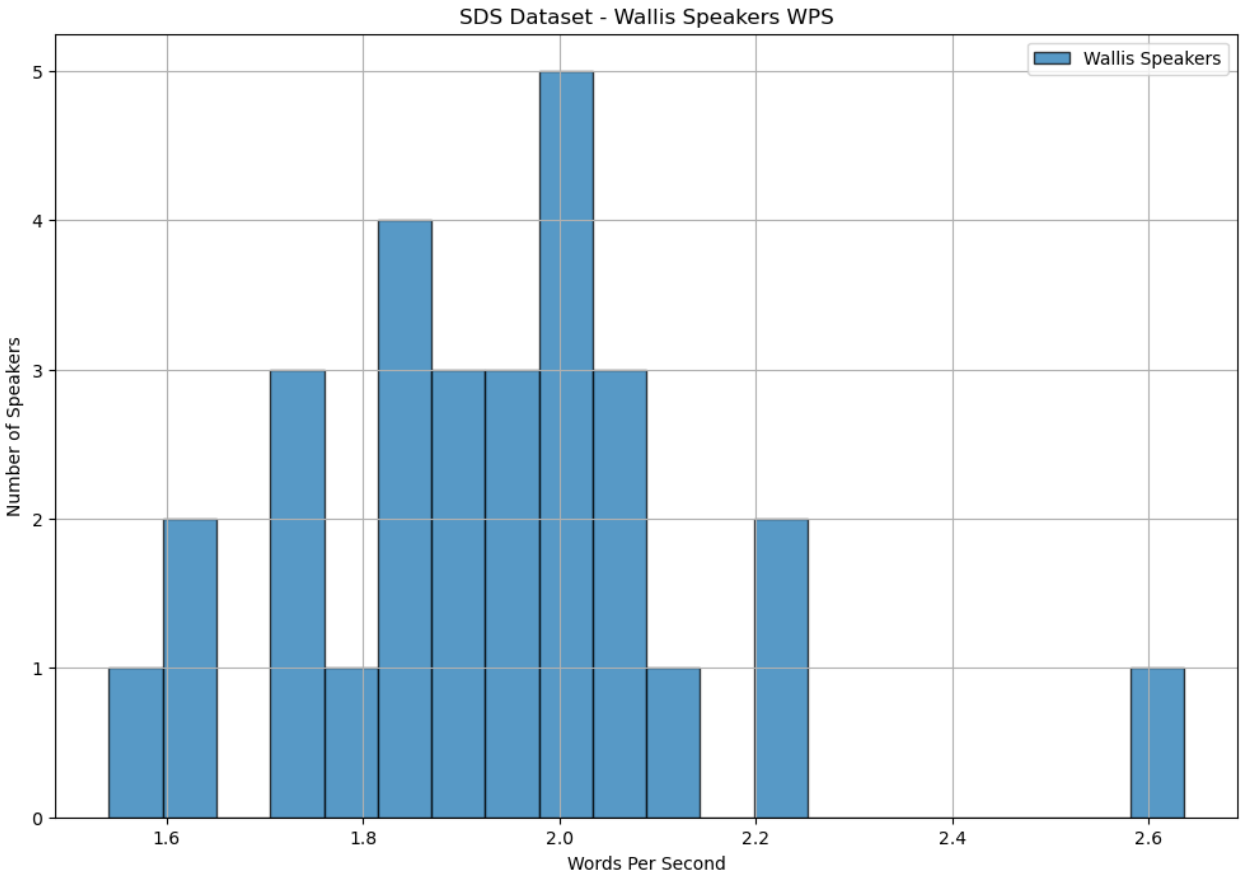

Figure 51: Histogram SDS-200 Region Wallis

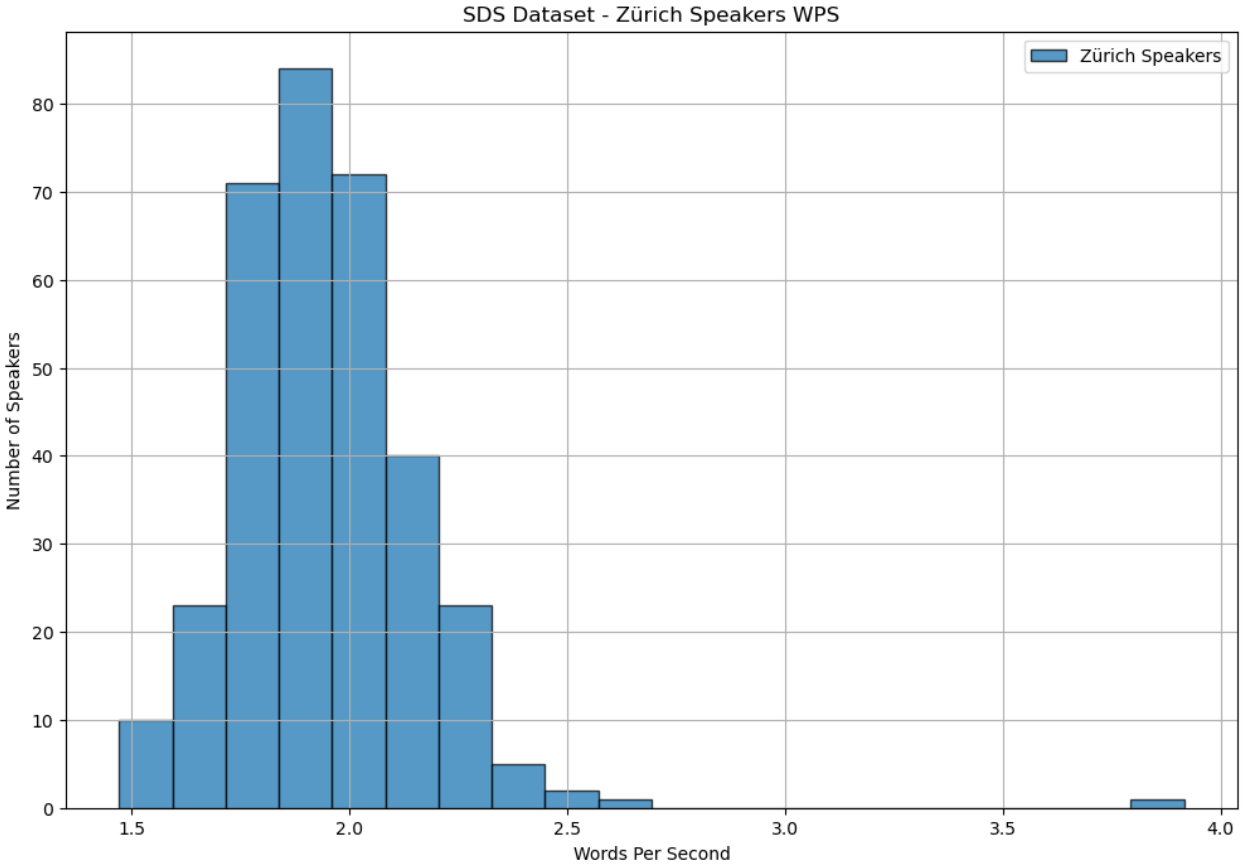

Figure 52: Histogram SDS-200 Region Zürich

9.4 Trimmed Time of Both Datasets

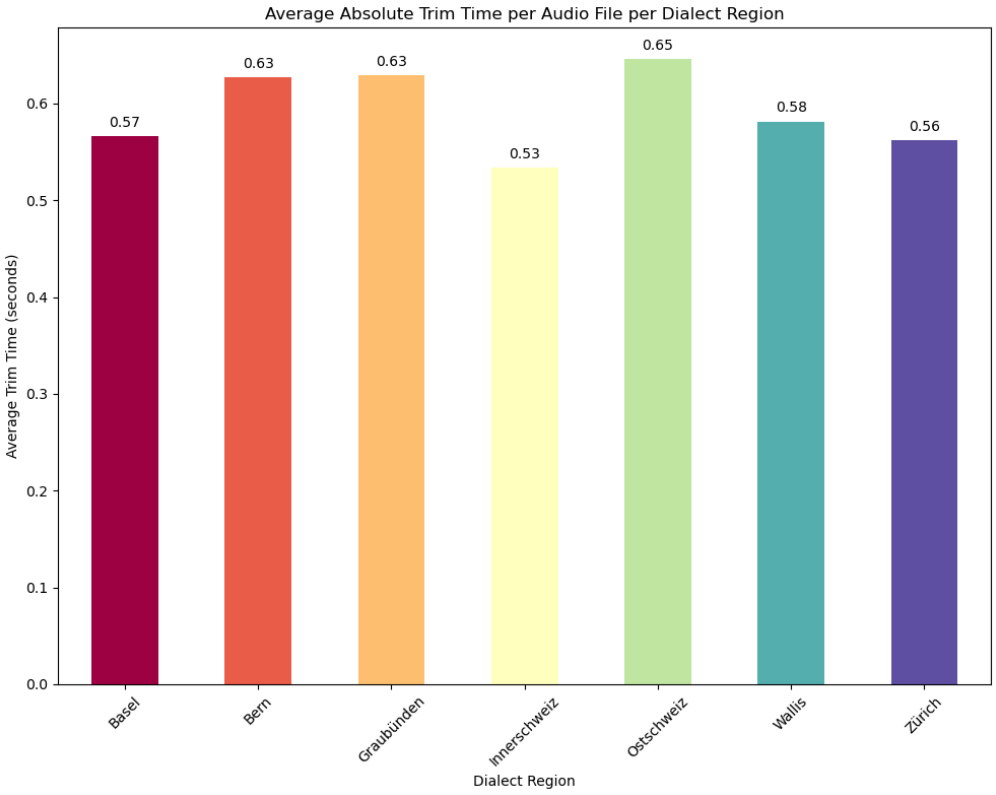

Figure 53: Absolute Trimmed Time of Both Datasets

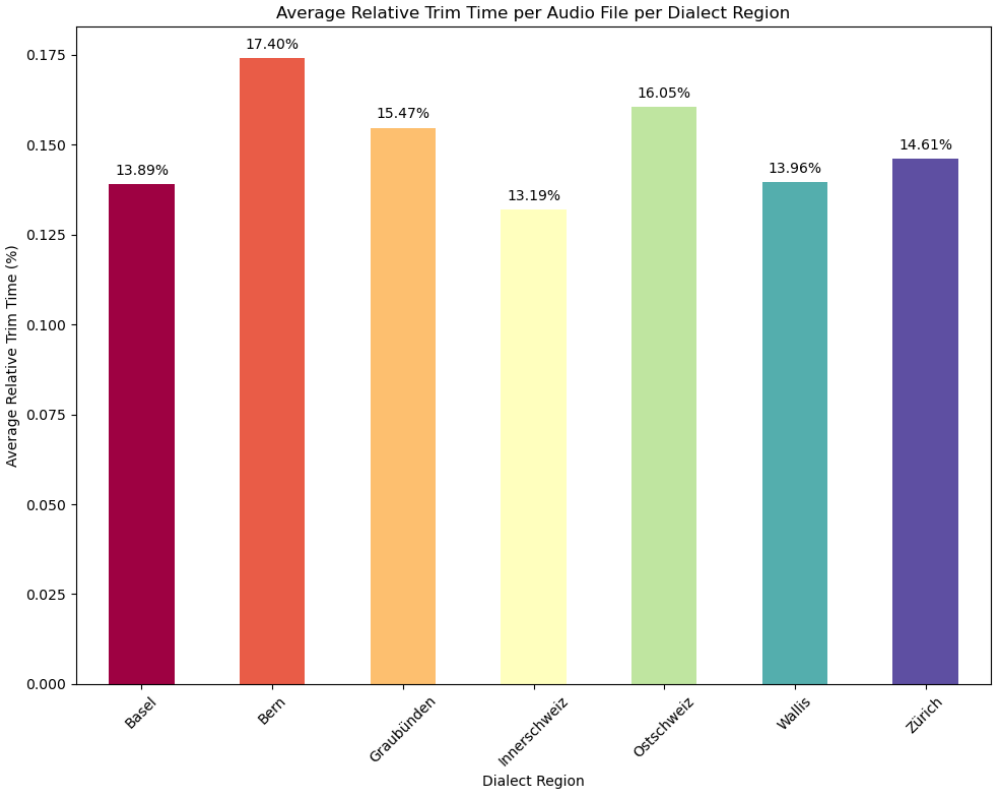

Figure 54: Relative Trimmed Time of Both Datasets

9.5 Dialect Variability Analysis: Regional Retrieval Rate Analysis

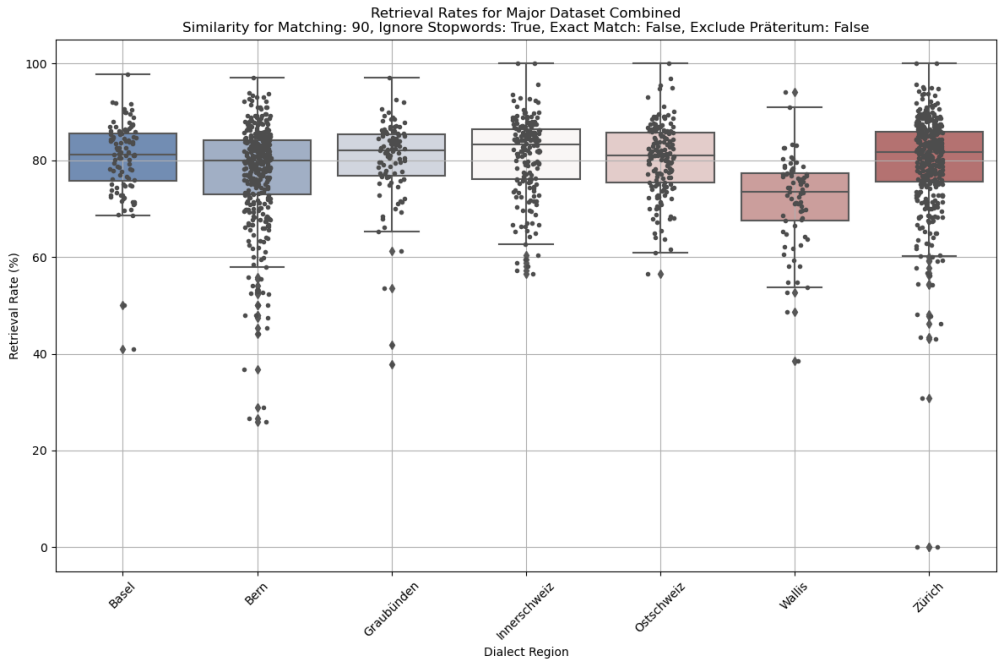

Figure 55: RRs per Region (Ignoring Stopwords, Considering Preterite Tense Sentences, Similarity Threshold: 90%)

Dialect Region	Mean	Median	Q1	Q3
Basel	80.39%	81.26%	75.74%	85.59%
Bern	77.43%	80.00%	72.99%	84.20%
Graubünden	79.98%	81.99%	76.82%	85.39%
Innerschweiz	80.92%	83.24%	76.19%	86.45%
Ostschweiz	80.24%	80.99%	75.41%	85.67%
Wallis	71.59%	73.42%	67.58%	77.29%
Zürich	79.33%	81.63%	75.63%	85.98%

Table 30: Statistical Summary of RR by Dialect Region - (Ignoring Stopwords, Considering Preterite Tense Sentences, Similarity Threshold: 90%)

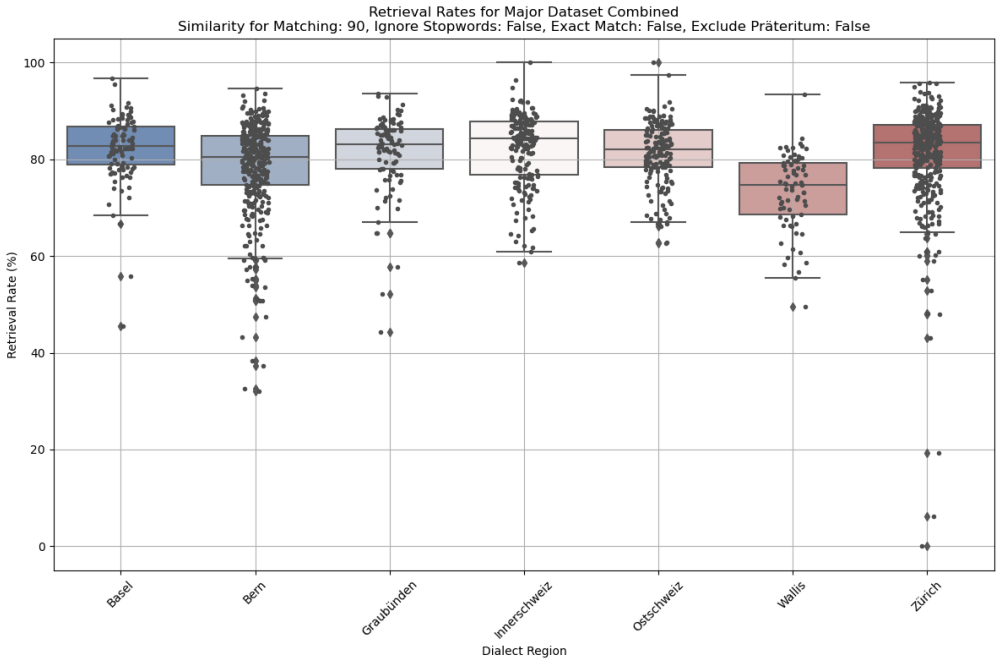

Figure 56: RRs per Region (Considering Stopwords, Considering Preterite Tense Sentences, Similarity Threshold: 90%)

Dialect Region	Mean	Median	Q1	Q3
Basel	82.32%	82.72%	78.92%	86.85%
Bern	78.35%	80.52%	74.65%	84.93%
Graubünden	81.40%	83.09%	78.00%	86.27%
Innerschweiz	82.14%	84.29%	76.86%	87.85%
Ostschweiz	81.20%	82.11%	78.31%	85.99%
Wallis	73.35%	74.77%	68.56%	79.19%
Zürich	81.12%	83.48%	78.21%	87.13%

Table 31: Statistical Summary of RR by Dialect Region - (Considering Stopwords, Considering Preterite Tense Sentences, Similarity Threshold: 90%)

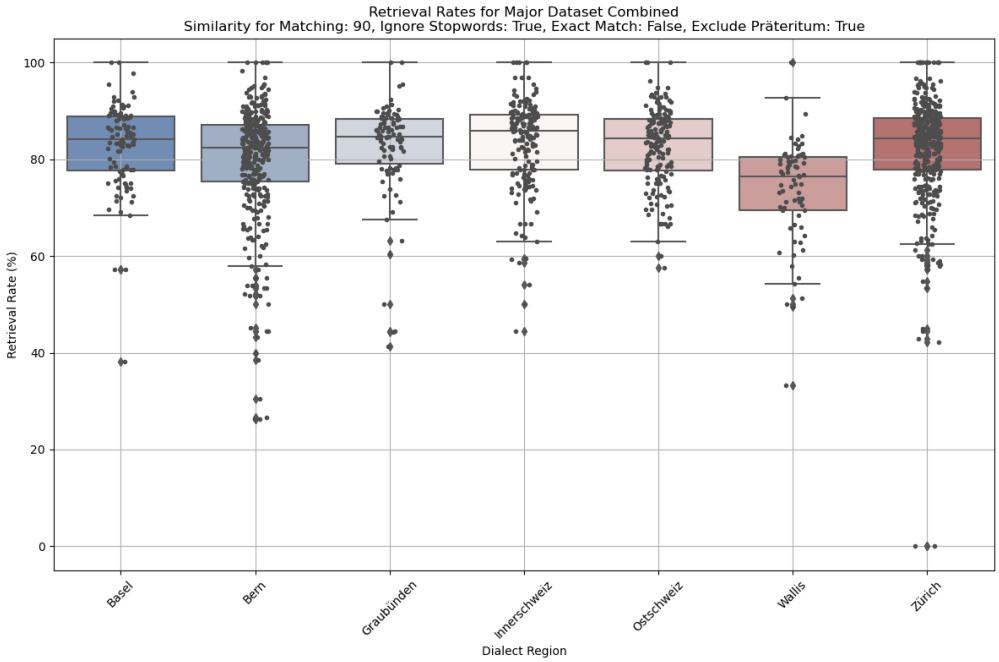

Figure 57: RRs per Region (Ignoring Stopwords and Preterite Tense Sentences, Similarity Threshold: 90%)

Dialect Region	Mean	Median	Q1	Q3
Basel	82.64%	84.21%	77.70%	88.91%
Bern	79.68%	82.35%	75.33%	87.10%
Graubünden	82.41%	84.70%	79.15%	88.41%
Innerschweiz	83.28%	85.88%	77.78%	89.19%
Ostschweiz	82.64%	84.37%	77.71%	88.37%
Wallis	73.54%	76.43%	69.44%	80.46%
Zürich	81.97%	84.38%	77.81%	88.49%

Table 32: Statistical Summary of RR by Dialect Region - (Ignoring Stopwords and Preterite Tense Sentences, Similarity Threshold: 90%)

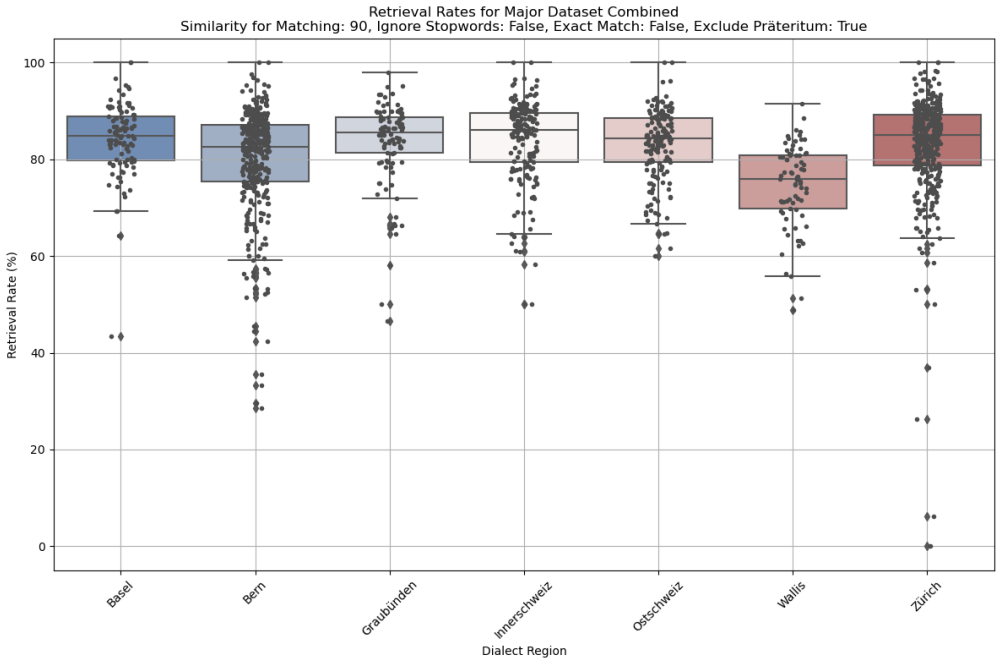

Figure 58: RRs per Region (Considering Stopwords, Ignoring Preterite Tense Sentences, Similarity Threshold: 90%)

Dialect Region	Mean	Median	Q1	Q3
Basel	84.19%	84.90%	79.81%	88.90%
Bern	79.90%	82.61%	75.44%	87.08%
Graubünden	83.40%	85.63%	81.31%	88.75%
Innerschweiz	83.86%	86.04%	79.43%	89.65%
Ostschweiz	82.97%	84.30%	79.44%	88.50%
Wallis	74.62%	75.90%	69.80%	80.90%
Zürich	82.92%	85.10%	78.78%	89.16%

Table 33: Statistical Summary of RR by Dialect Region - (Considering Stopwords, Ignoring Preterite Tense Sentences, Similarity Threshold: 90%)

9.6 Dialect-specific Words Research: List Generation Distribution

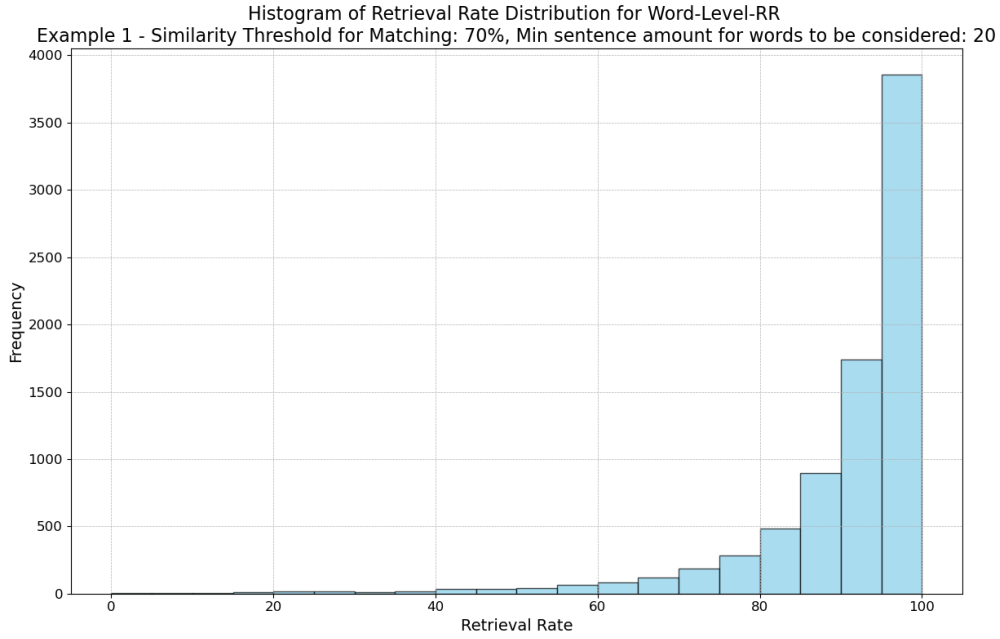

Figure 59: Histogram of RRs for Word-Level-RR output of Example 1

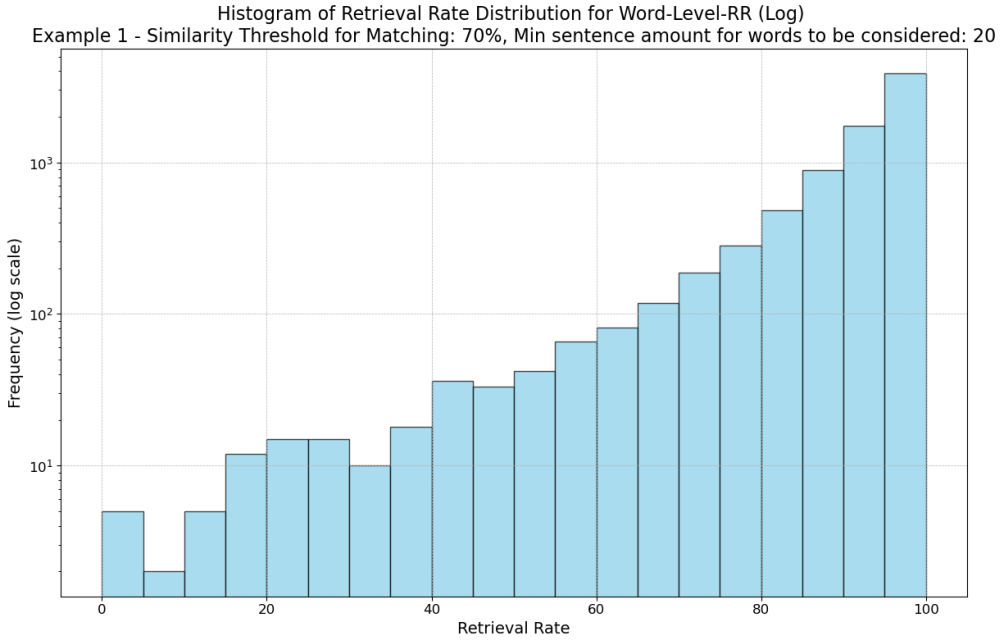

Figure 60: Histogram of RRs for Word-Level-RR output of Example 1 (Log)

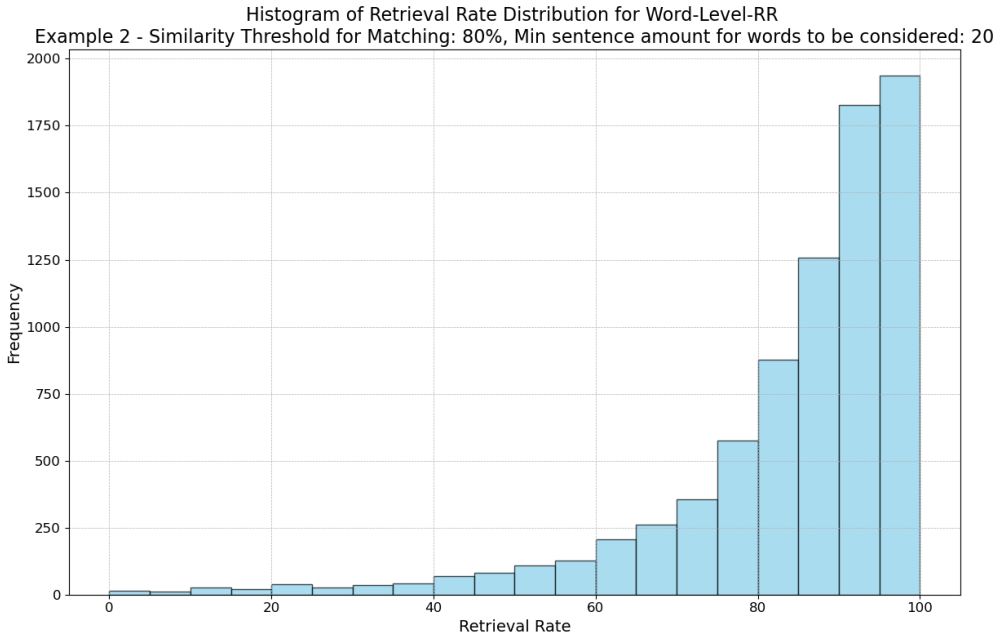

Figure 61: Histogram of RRs for Word-Level-RR output of Example 2

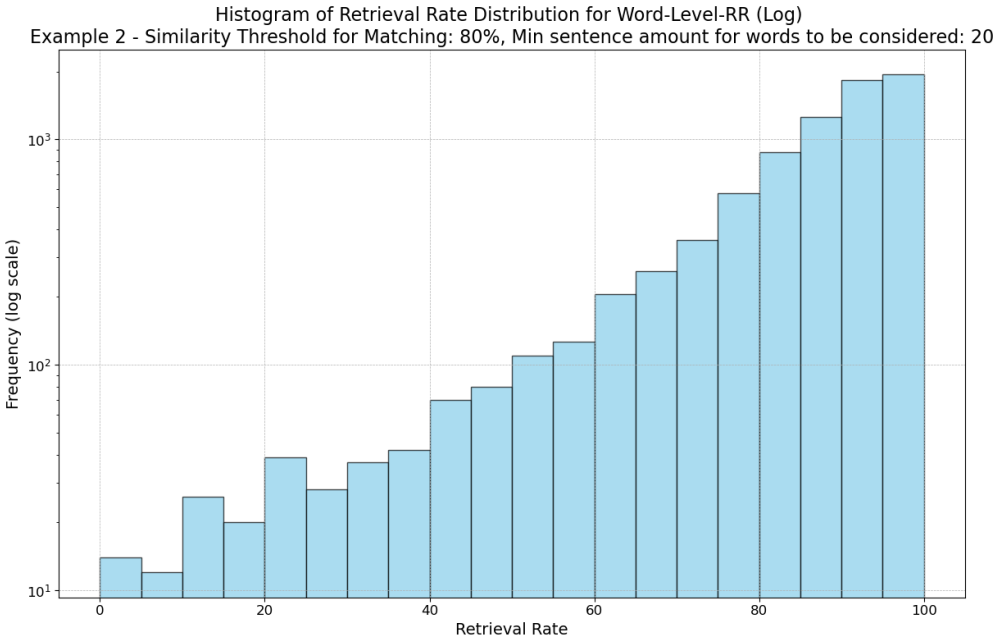

Figure 62: Histogram of RRs for Word-Level-RR output of Example 2 (Log)

9.7 Dialect-specific Words Research: List Comparison Gridsearch

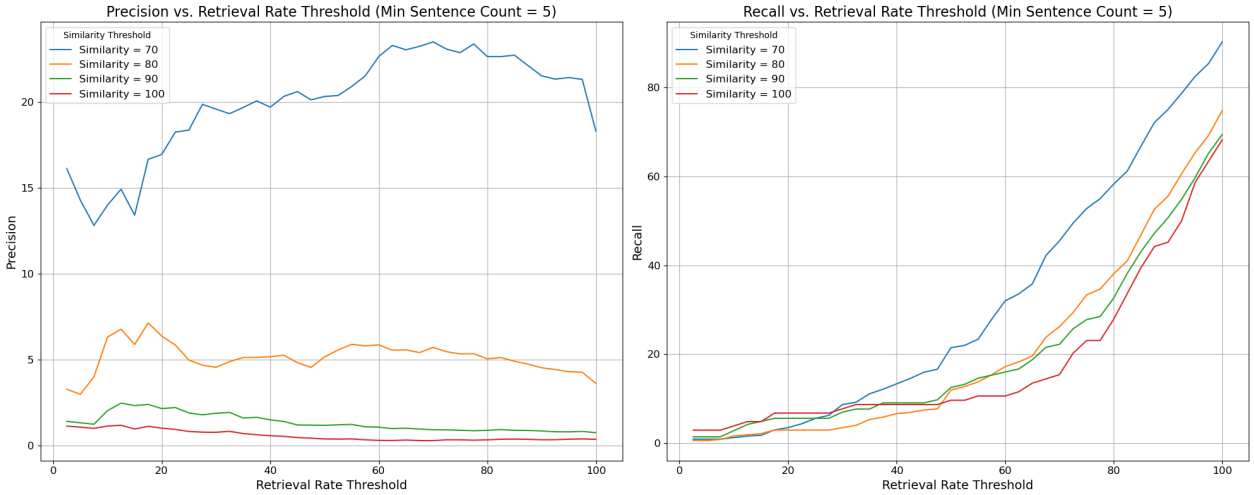

Figure 63: Dialect-specific Words List Comparison - Evolution of Precision and Recall for Min. Sent.: 5

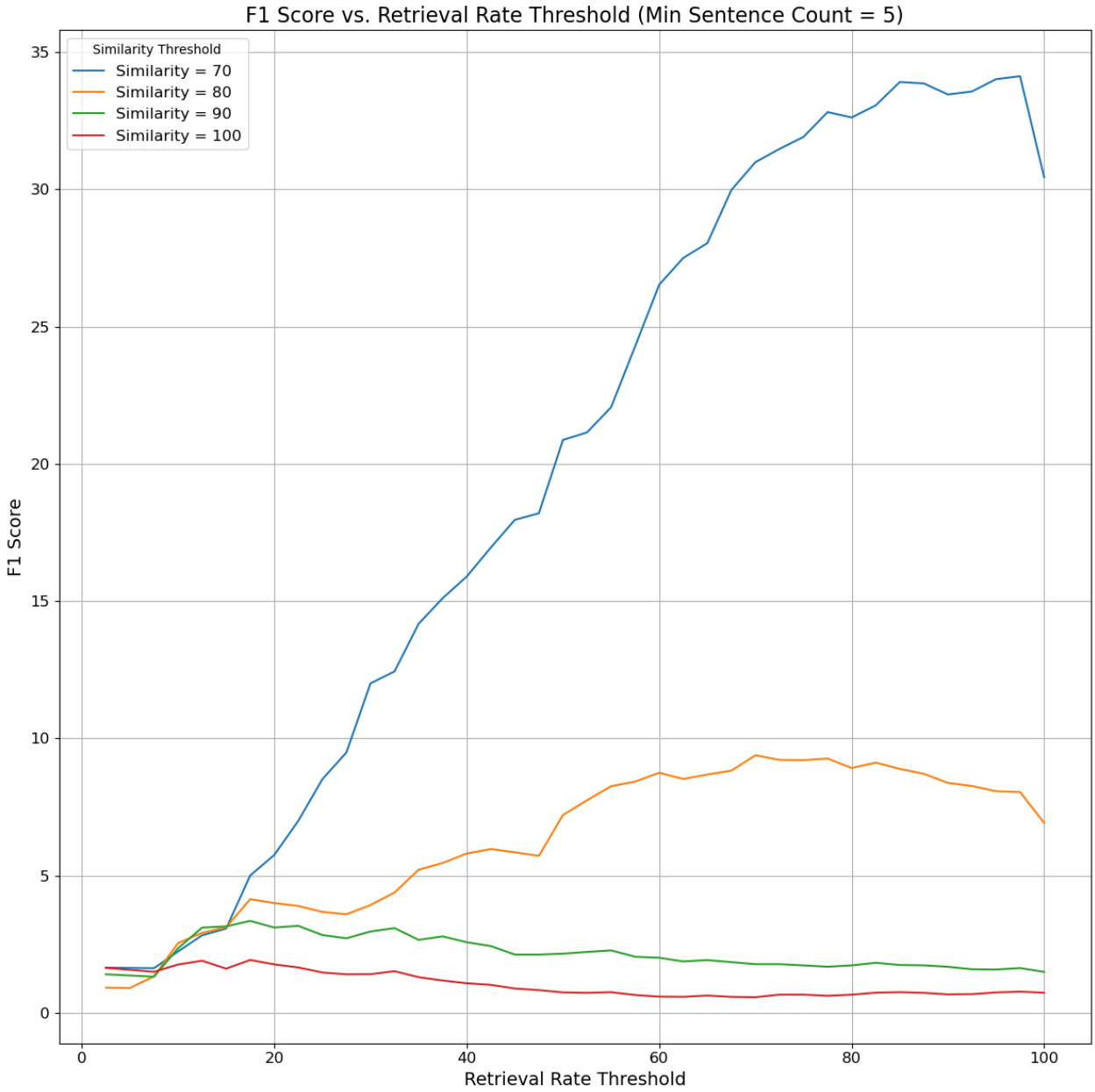

Figure 64: Dialect-specific Words List Comparison - Evolution of F1-Score for Min. Sent.: 5

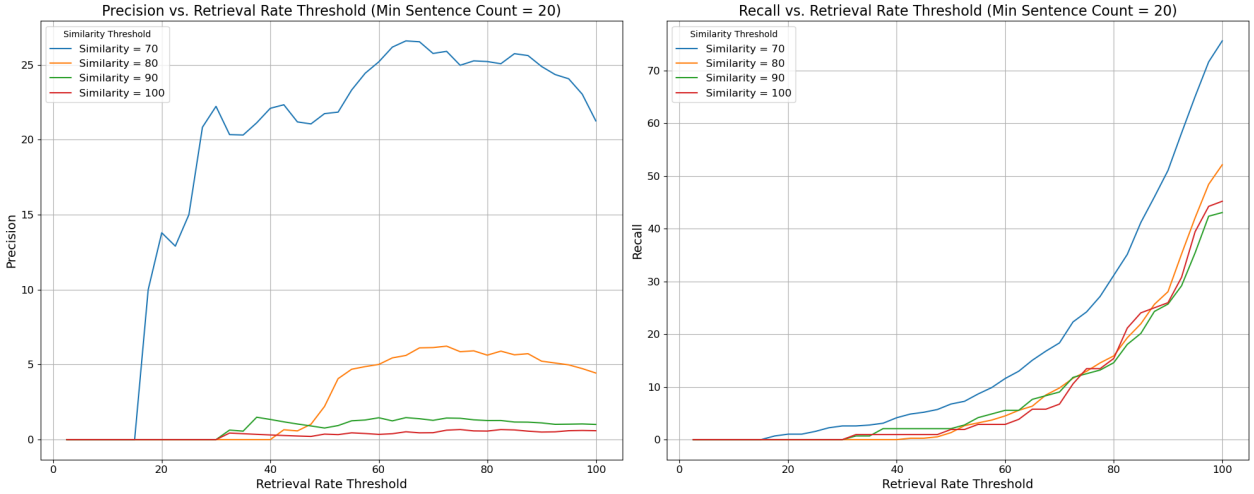

Figure 65: Dialect-specific Words List Comparison - Evolution of Precision and Recall for Min. Sent.: 20

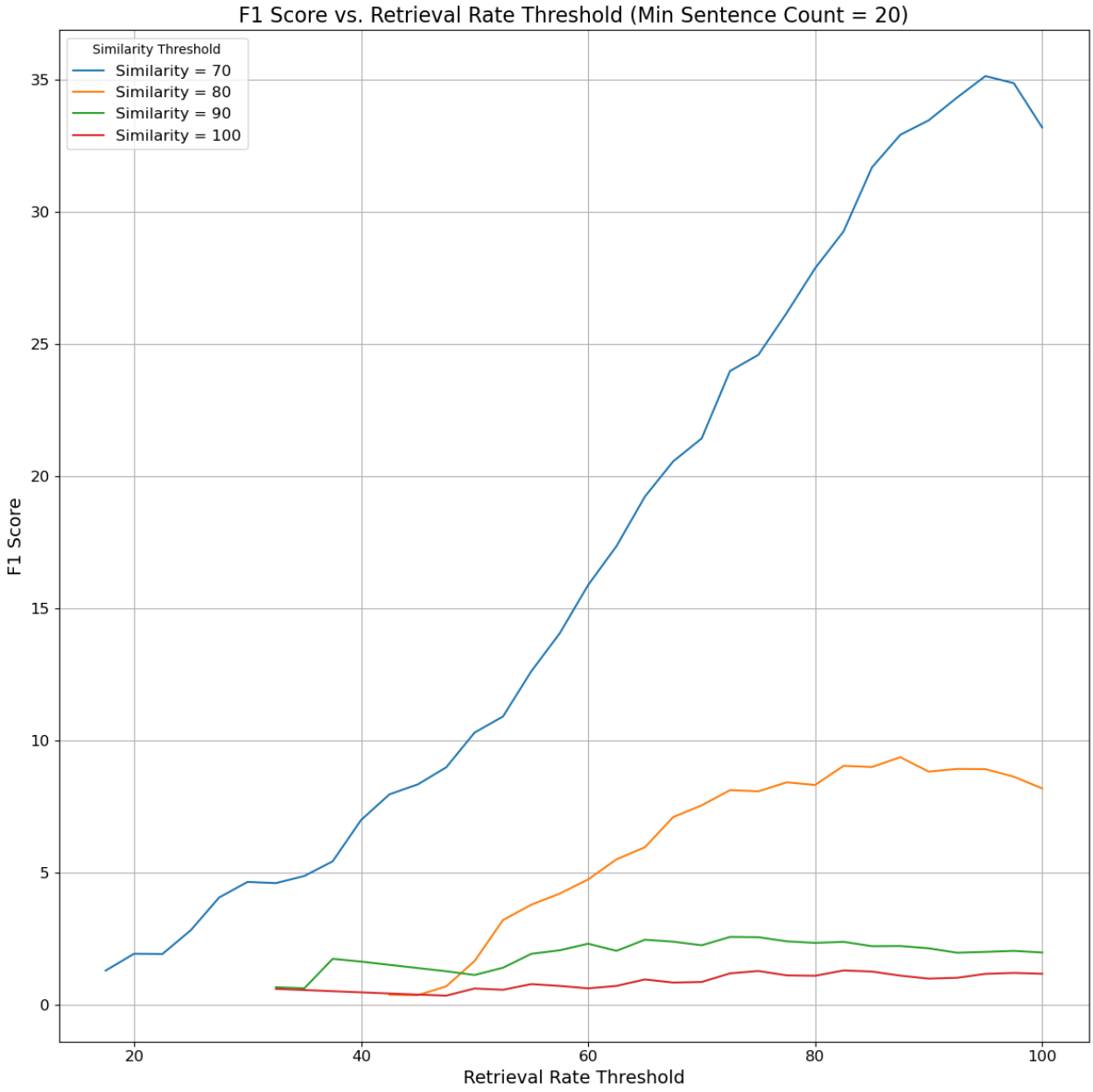

Figure 66: Dialect-specific Words List Comparison - Evolution of F1-Score for Min. Sent.: 20

9.8 Regional Analysis: Heatmaps

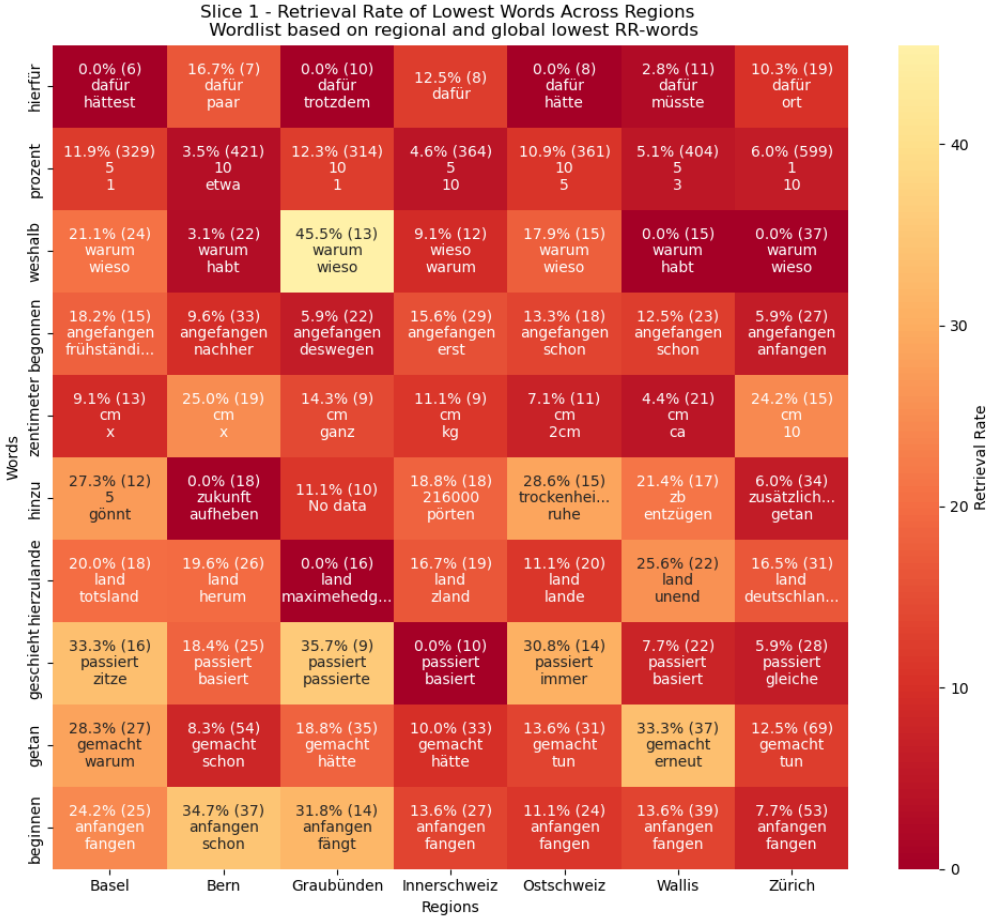

Figure 67: HeatMap: RR across Regions - Slice 1

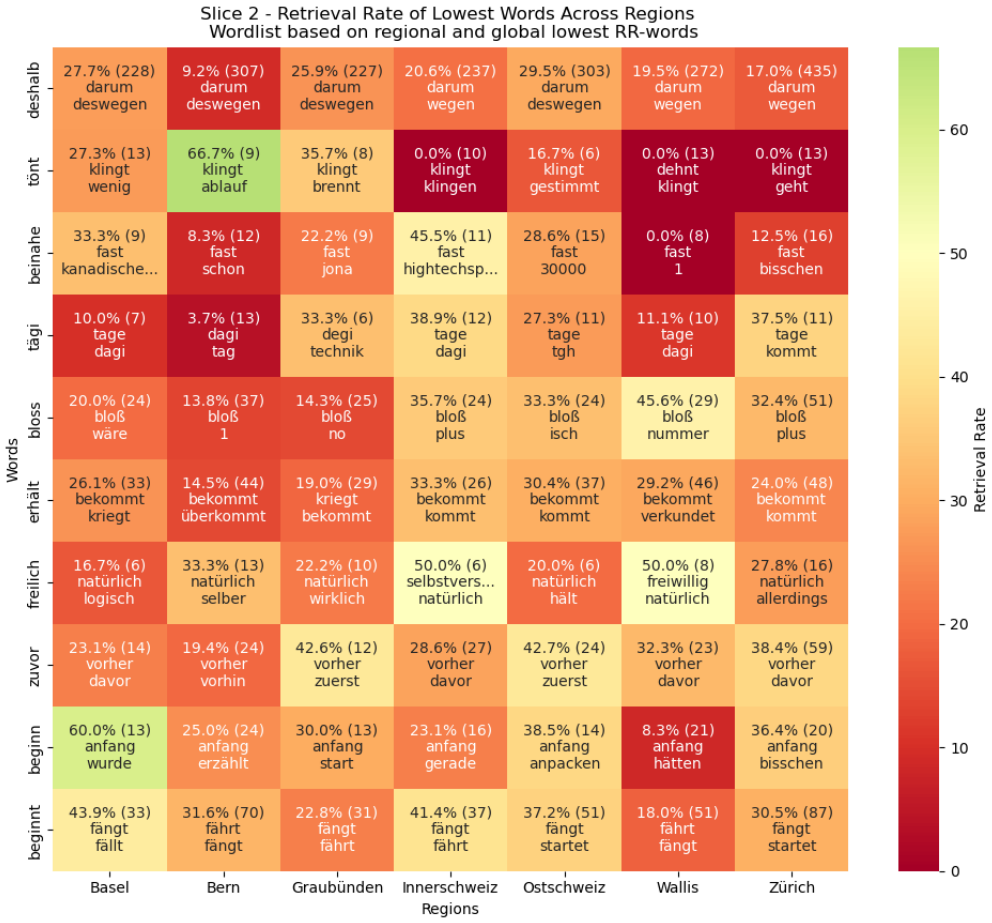

Figure 68: HeatMap: RR across Regions - Slice 2

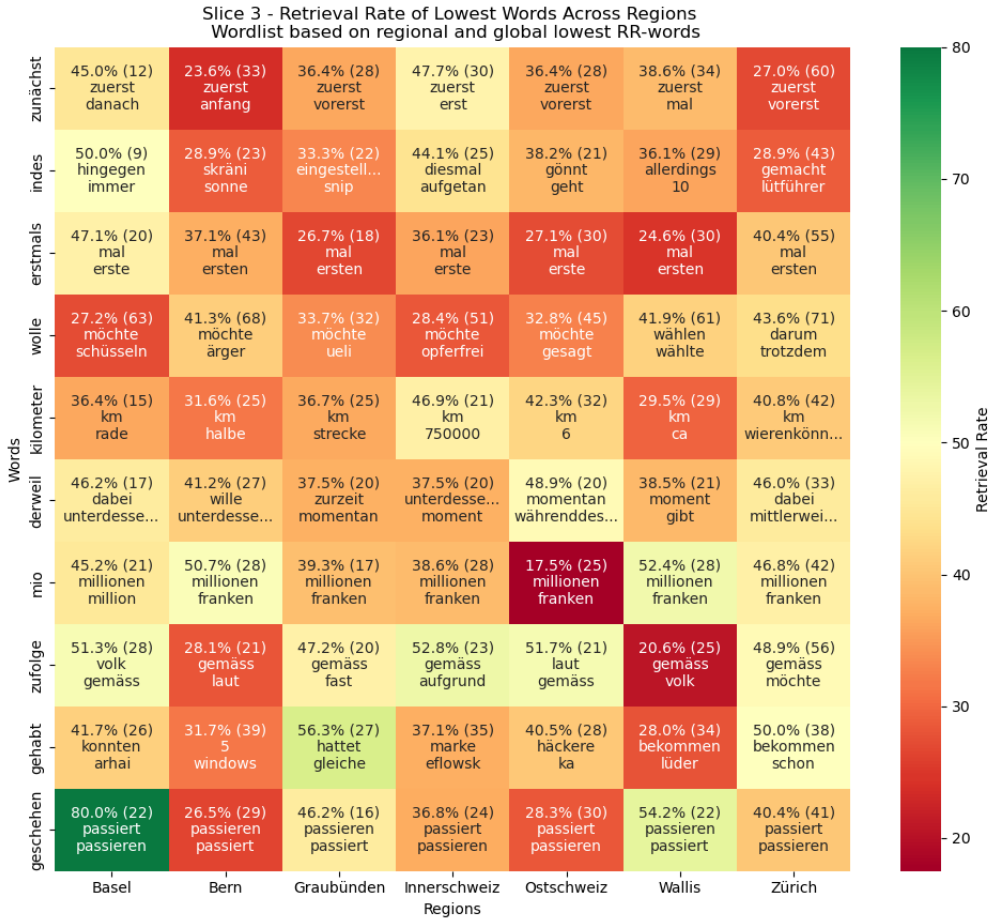

Figure 69: HeatMap: RR across Regions - Slice 3

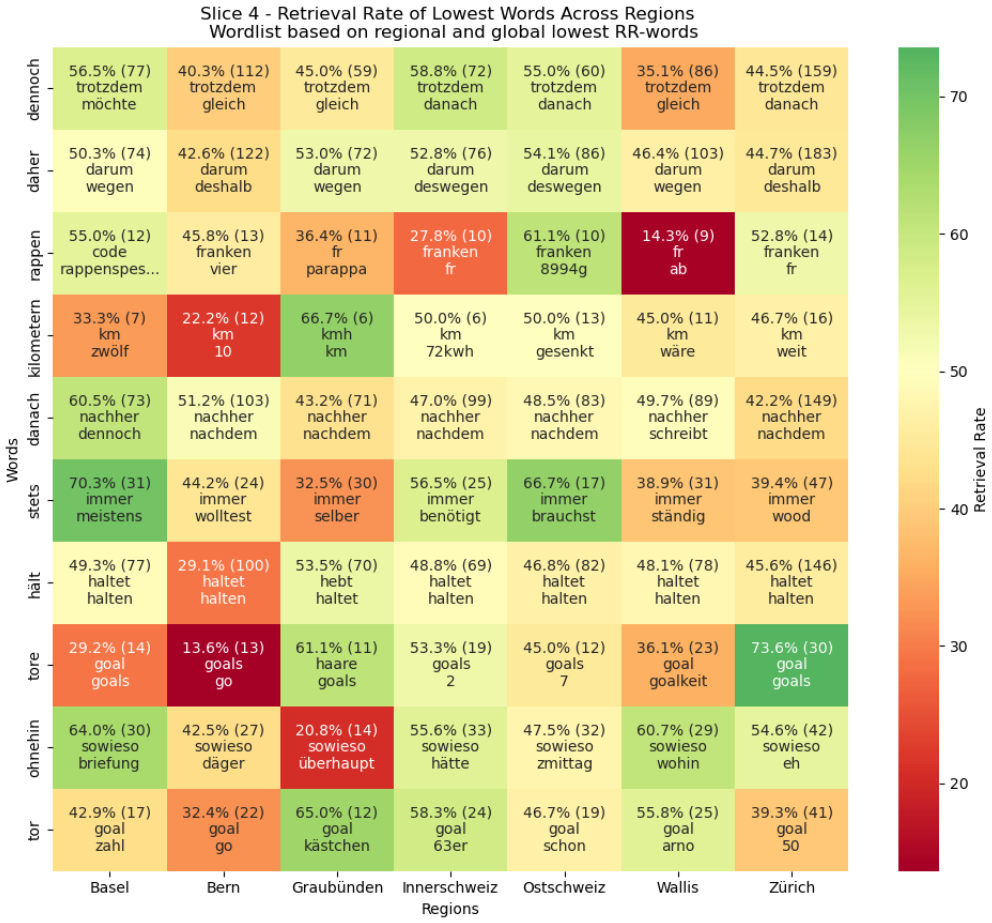

Figure 70: HeatMap: RR across Regions - Slice 4

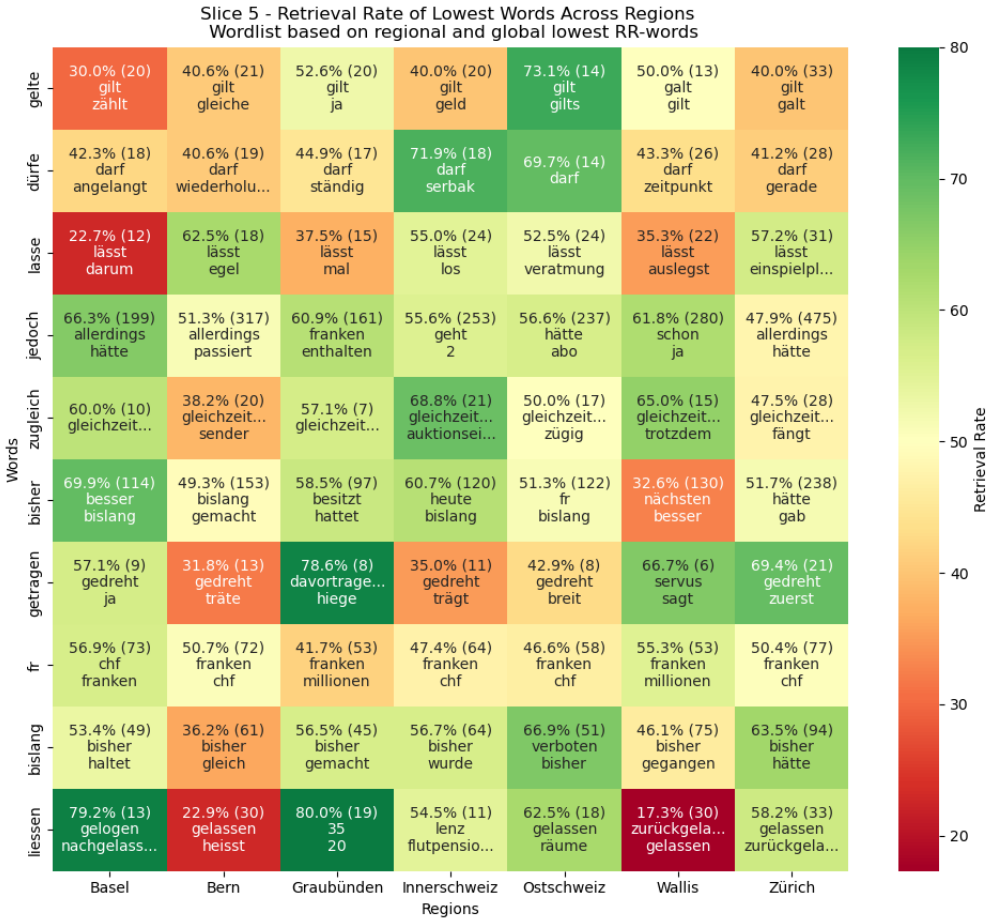

Figure 71: HeatMap: RR across Regions - Slice 5

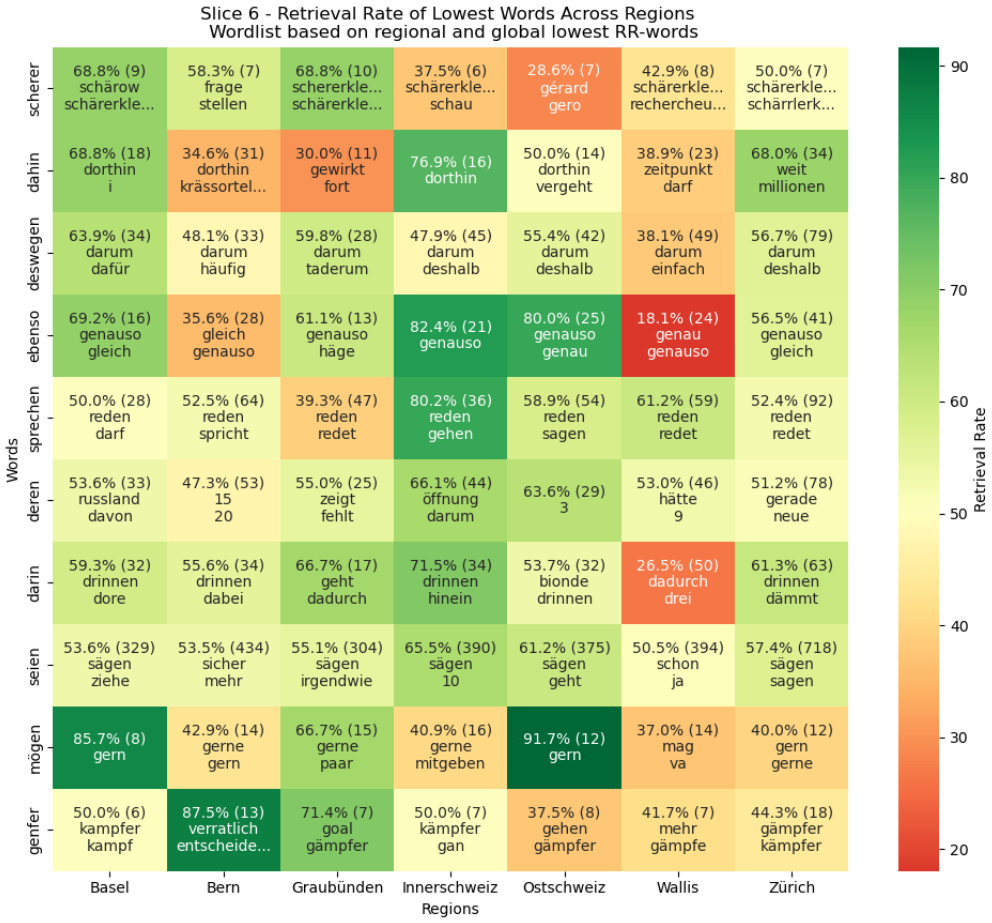

Figure 72: HeatMap: RR across Regions - Slice 6

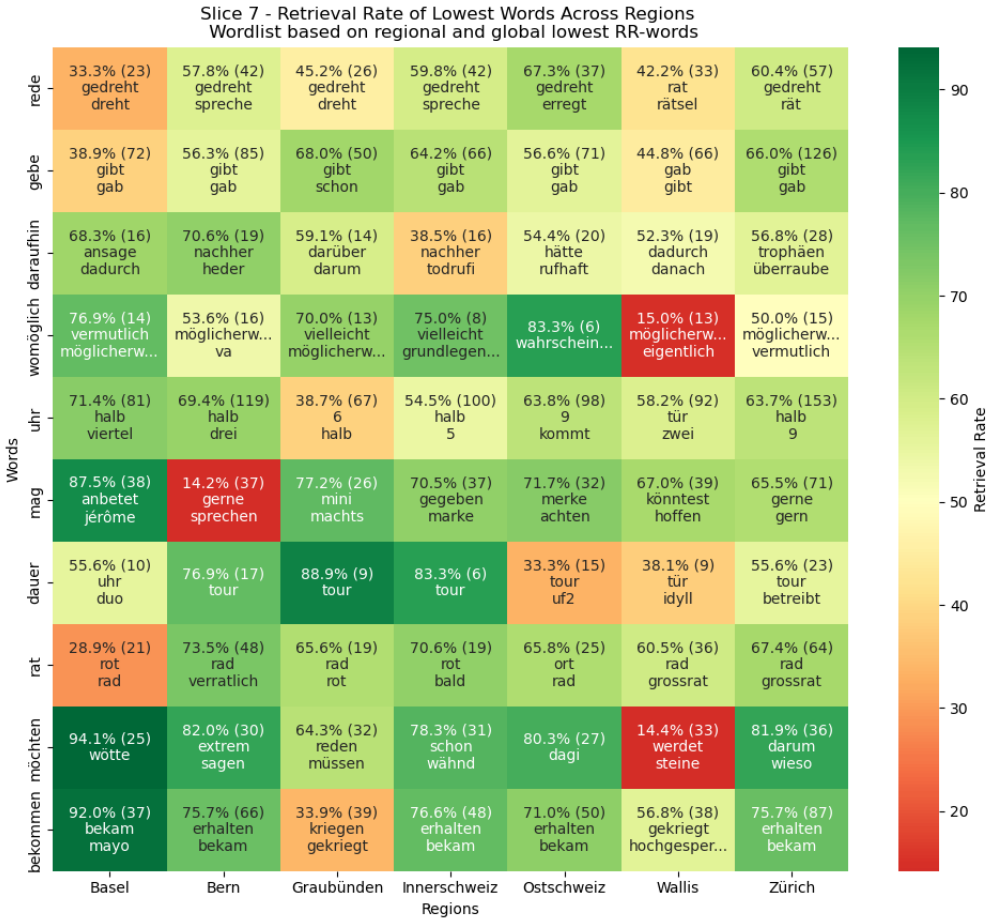

Figure 73: HeatMap: RR across Regions - Slice 7

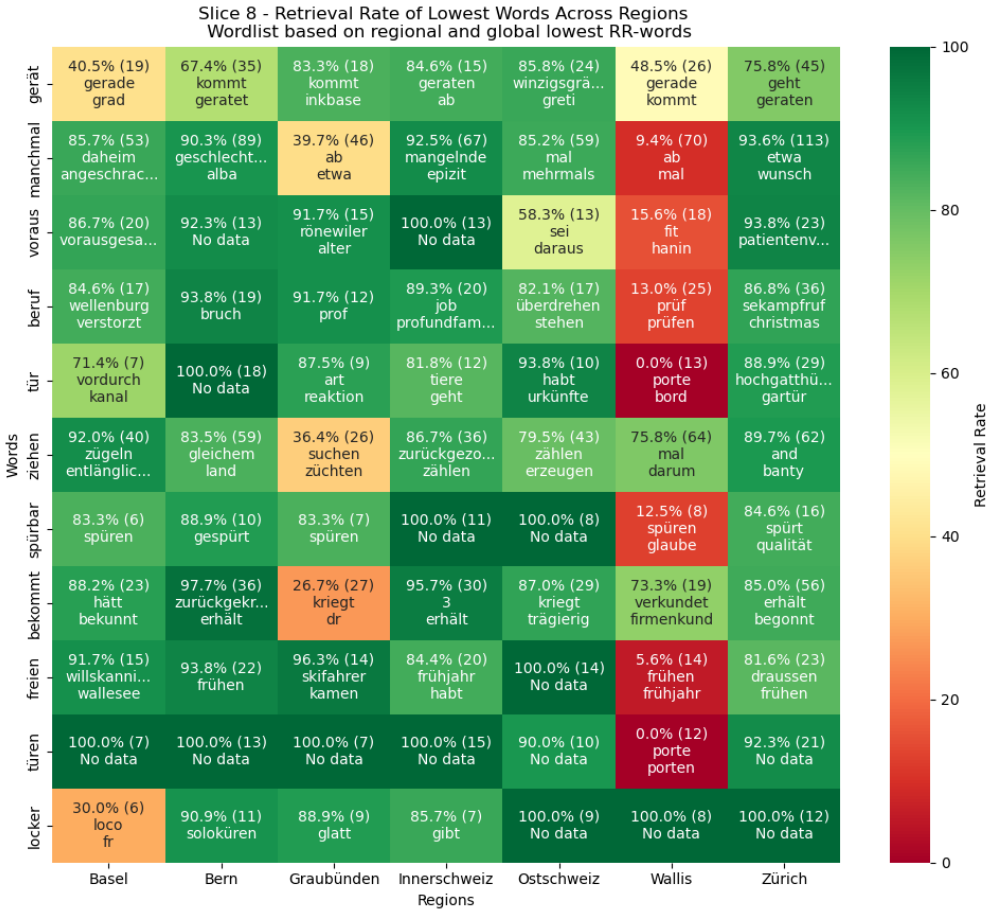

Figure 74: HeatMap: RR across Regions - Slice 8

10 Appendix B: Code Listings and Manuals

10.1 Installation Instructions

This section provides an overview of how to install and use the code repository from this thesis. Before using the scripts, attention should be paid to the following requirements.

It is also important to download the corpora data so that it is accessible for the scripts.

10.1.1 Code Adjustements

There are two config-sample files that have to be adjusted.

General Config-File: This file is placed in 'code/config-sample.py'. It should be renamed to 'config.py', and it is necessary to fill out the local paths to fit the local environment. An OpenAI-Key has to be generated[32] and implemented in the config-file as well as the Huggingface Key[33].

Additional Config-File: This file is placed in 'additional/configadd-sample.py'. It should be renamed to 'configadd.py', and it is necessary to fill out the local paths to fit the local environment.

10.1.2 Required Modules

This subsection lists the required modules. Libraries like 'os', 'shut', 're', 'time', 'math', and 'CSV' should already be part of the standard Python library. The other modules can be installed with 'pip' or 'brew' or another command specified for the operation system.

- fuzzywuzzy
- librosa
- matplotlib
- nemo_toolkit[all]
- nltk
- numpy
- openai
- pandas
- pickle
- pyannote.audio[34]
- pyannote.segemntation3.0[35]
- pydub
- seaborn
- spacy
- text-to-num
- torch
- torchaudio
- tqdm

10.1.3 Execution

Python 3.0 on the command line or Jupyter Notebook via Anaconda Navigator was used to run the scripts.

[32] Can be done with the instructions on the Developer Quickstart
[33] Can be done with the instructions on Quicktour
[34] Installation-Help can be found on GitHub `https://github.com/pyannote/pyannote-audio`.
[35] This is a prerequisite for pyannote.audio, found on `https://huggingface.co/pyannote/segmentation-3.0`.

10.2 Code Repository

For a detailed view of the implementation and algorithms used in this study, the reader is referred to the attached file, 'speech-pipeline.zip'.

10.3 Technical outputs

10.3.1 Dialektwörter.ch word list

The complete single-word list of Dialektwörter.ch used in the chapters 6.2 and 6.3.2 of this thesis can be found in the file code/output/dialect_specific_word_count.tsv of the repository.