Zürcher Hochschule
für Angewandte Wissenschaften

School of Engineering

CAI Centre for Artificial Intelligence

Project thesis (Engineering & Management)

Corpus Evaluation for Automatic Speech Recognition 2.0 (“CEASR 2.0”)

Author	Aurora Alitjaha
	Mario Küng
Main supervisor	Mark Cieliebak
Sub supervisor	Katsiaryna Mlynchyk
Date	23.12.2022

Zürcher Hochschule
für Angewandte Wissenschaften

School of Engineering

DECLARATION OF ORIGINALITY
Project Thesis at the School of Engineering

By submitting this project 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.

City, Date:	Name students:
Winterthur, 23.12.2022	Aurora Alitjaha
Winterthur, 23.12.2022	Mario Küng

Abstract

Automatic speech recognition (ASR) has come a long way and have considerably improved speech-to-text quality. These development leads to increasing adoption of real-life applications of ASR such as live captioning, virtual assistants and chatbots.

A high-level comparative system evaluation in terms of word error rate (WER) and latency for French and Italian is performed with four popular state-of-the-art ASR engines. The data originates from widely known public speech corpora such as Common Voice and Multilingual LibriSpeech.

A comparison of three evaluations from 2019 to 2022 all based on English speech data accessed from CEASR corpus was conducted. Yielding no meaningful results in terms of accuracy development as the engines were configurated differently in each of the examined evaluations. Highlighting the importance of configuration and parameter settings.

The evaluation framework is based on the work of Cieliebak, Benites, Germann, Hürlimann and Ulasik published as “CEASR: A Corpus for Evaluating Automatic Speech Recognition”.

Our experiments show significant differences in accuracy between French and Italian. The results cannot be relied upon in terms of absolute accuracy as there were issues with the configuration of the ASR engines and shall be viewed as indicative of relative measures between the ASR engines.

We hope that this project thesis will serve as an outline for students on performing an evaluation study in the fields of speech-to-text and its pitfalls.

Zusammenfassung

Die automatische Spracherkennung (Automatic Speech Recognition, ASR) hat einen langen Weg zurückgelegt und die Qualität von Speech-to-Text erheblich verbessert. Diese Entwicklung führt zu einer zunehmenden Verbreitung von ASR-Anwendungen im realen Leben, wie z. B. Live-Untertitel, virtuelle Sprachassistenten und Chatbots.

Eine vergleichende Systemevaluation in Bezug auf die word error rate (WER) und die Latenzzeit für Französisch und Italienisch wird mit vier populären state-of-the-art ASR Engines durchgeführt. Die Daten stammen aus bekannten öffentlichen Sprachkorpora wie Common Voice und Multilingual LibriSpeech.

Es wurde ein Vergleich von drei Auswertungen aus den Jahren 2019 bis 2022 durchgeführt, die alle auf englischen Sprachdaten aus dem CEASR-Korpus basieren. Es wurden keine aussagekräftigen Ergebnisse in Bezug auf die Entwicklung der Transkriptionsgenauigkeit erzielt, da die Engines in jeder der untersuchten Evaluierungsdurchführungen anders konfiguriert waren. Dies unterstreicht die Wichtigkeit der Konfiguration und der Parametereinstellungen.

Der Evaluierungsrahmen basiert auf der Arbeit von Cieliebak, Benites, Germann, Hürlimann und Ulasik, die unter dem Titel "CEASR: A Corpus for Evaluating Automatic Speech Recognition" veröffentlicht wurde.

Unsere Experimente zeigen signifikante Unterschiede in der Genauigkeit zwischen Französisch und Italienisch. Die Ergebnisse können nicht als absolute Kennzahl für die Genauigkeit gewertet werden, da es Probleme mit der Konfiguration der ASR-Engines gab, und sollen als Indikator für relative Vergleiche zwischen den ASR-Engines dienen.

Wir hoffen, dass diese Projektarbeit Studenten als Leitfaden für die Durchführung einer Evaluierungsstudie im Bereich Speech-to-Text und deren Fallstricke dienen kann.

Contents

Abstract 3

Zusammenfassung 3

1 Introduction 6

1.1 Framework of CESAR 2.0 6

1.2 Objectives 6

2 Foundations 7

2.1 Automatic Speech Recognition 7

2.2 Speech Corpora 7

2.2.1 Common Voice 11.0 (CV) 8

2.2.2 Multilingual Librispeech (MLS) 9

2.2.3 Fleur 10

2.3 ASR Systems Evaluation 11

2.3.1 Evaluation Techniques and Metrics 11

2.3.2 Word Error Rate 11

2.3.3 Latency 11

2.3.4 Real-time-factor 12

2.3.5 BeSTT 12

3 Experimental Setup 13

3.1 Corpora 13

3.1.1 Corpora selection criteria 13

3.1.2 Corpora research and evaluation 14

3.1.3 Overview of selected corpora 14

3.2 ASR Engines selection 15

3.3 Engine Implementation 15

3.4 Selected ASR Engines and configurations 16

3.5 Data structure and transformation 16

3.5.1 Pre-processing data checks 17

3.5.2 Post-processing 17

3.5.3 Decoding and encoding 17

3.5.4 Contractions 17

3.5.5 Number to words 18

3.5.6 General string manipulations 18

3.5.7 Exclusion of bad samples 18

3.5.8 WER recalculation 18

3.6 Benchmarking 19

4 Results 20

4.1 French and Italian Data Sets 20

4.1.1 Number of Samples 20

4.1.2 WER per Corpus and Engine 21

4.1.3 Audio Duration and Transcription Length 23
4.1.4 Speaker and Gender 27
4.1.5 Age 30
4.1.6 Latency 32
4.2 English Data Sets 35
4.2.1 Overall performance 35
4.2.2 Speaking Style 36
4.2.3 Language Skills 36
4.2.4 Latency 38
4.2.5 Comparison of systems over time 39
5 Conclusion 41
6 Discussion and Outlook 42
7 References 43
7.1 Bibliography 43
7.2 Glossary 45
7.3 List of Figures 46
Appendix 48
A. Corpora Documentation 48
B. Overview Benchmark Execution 57
C. Python code used for post-processing and WER calculation 58
Project management 59
A. Description 59
B. Timetable 59
C. Protocol 60

1 Introduction

Automatic speech recognition (ASR) or alternatively Speech-To-Text (STT) technology has come a long way in recent years, and many systems are now able to achieve high levels of accuracy, even in noisy environments. With the improved accuracy, decreasing costs in terms of fees and computational power and reduced latency. ASR systems are becoming increasingly important in everyday applications such as live captioning and transcription, virtual assistants and chatbots.

Against this background the evaluation of these systems is also becoming increasingly relevant outside of academic research. The evaluation is often performed on speech corpora (large collection of transcribed audio recordings) the results are used for comparing different systems and to further improve or develop such systems.

The project at hand focuses on the continuation of "Corpus Evaluation for Automatic Speech Recognition" short CEASR. Supplementary to English and German, French and Italian locale will be added to the corpora on which the evaluation of the ASR engines is performed. The goal is to represent all four official languages in Switzerland (excluding Romansh as it is not supported by any of the ASR engines).

The results of CEASR 1.0 will be compared with the results of this project work. Furthermore, the data analysis will be extended and intensified covering further fields such as gender and transcription length.

1.1 Framework of CESAR 2.0

This project thesis is a contribution to the continuation of the framework laid out by Corpus Evaluation for Automatic Speech Recognition" short CEASR 1.0 [1]. CEASR 1.0 is based on English and German public speech corpora which were standardized and include the normalized transcript texts and metadata. This paper has focused on the performance of open-source vs. proprietary systems and found that proprietary systems notably outperformed.

Supplementary to English and German, French and Italian data sets (in CESAR format) shall be added to make this more relevant to evaluations regarding Switzerland. As it contains all national languages (excluding Romansh) and English which is often used to bridge the divides in relation to language regions. Also new engines and models shall be selected to be onboarded to Benchmark Evaluation Speech to Text platform ("BeSTT"), an evaluation tool built for and on the basis of CESAR. This primarily includes the configuration of these ASR systems and implement them within the BeSTT framework. The configuration and implementation are not in scope for this project.

1.2 Objectives

The aim of this project thesis is to collect a wide range of data of various domains and speech types from public available speech corpora for the three major official languages of Switzerland and English. This was narrowed down to English, French and Italian during the course of the work.

Also, ASR engines should be identified, which are widely used in practice or have attracted attention in recent publications by good performance for the languages in scope (i.e. English, French and Italian). In addition, these providers should - if possible - give us permission to mention them by name in a potential later publication and to grant a discount for their use for academic purposes. As cost is a restrictive factor in the academic evaluation of proprietary systems on large amounts of audio samples.

Once the data is collected and the engines are selected to do the evaluation, both need to be configured and implemented on BeSTT. The data must be put into a specific format (i.e., unified format) and ultimately uploaded. The engines must be configured for the use on the platform. The configuration and implementation of the selected ASR engines was abandoned during the course of this work because as it was beyond our understanding of IT infrastructure and programming capabilities.

Nevertheless, the engines already implemented on BeSTT from previous projects (see 3.2 ASR Engines selection) and the newly collected data sets for Italian and French will be evaluated. Also, for English,

the evaluation will be repeated with the same data from 2019 and mid-2022 to see if the performance has improved over time.

We hope that our work can contribute and help setting some guidelines to the colleagues who will continue this project as a bachelor thesis after us.

2 Foundations

2.1 Automatic Speech Recognition

Automatic speech recognition (ASR) is a technology that allows computers to recognize and transcribe spoken words into written format [2]. It is used in a variety of applications, including voice-enabled virtual assistants, transcription services, and call centers [3].

ASR systems typically use machine learning algorithms to analyse speech patterns and identify the words being spoken. They may also use natural language processing (NLP) techniques to understand the meaning of the words and sentences in the context of the overall conversation.

ASR systems can be trained on large amounts of data (see 2.2 Speech Corpora) to improve their accuracy and ability to understand a wide range of accents and dialects [4]. In general, the more data an ASR system is trained on, the better it will perform [4].

ASR technology has come a long way in recent years, and many systems are now able to achieve high levels of accuracy, even in noisy environments. However, there are still challenges to be overcome, such as accurately recognizing words with similar sounds and dealing with rapid, spontaneous speech [5].

Overall, ASR technology has the potential to greatly improve our ability to communicate with computers and other devices and is likely to play an increasingly important role in our daily lives in the coming years [6].

2.2 Speech Corpora

A speech corpus is a collection of recorded speech data that is used to train and evaluate speech recognition systems. These corpora typically consist of hundreds or thousands of hours of recorded speech, transcribed to provide a ground truth text for the words being spoken. This ground truth text constitutes the basis for the evaluation of such systems.

Speech corpora are an essential part of developing and improving automatic speech recognition (ASR) systems. They provide the training data that ASR systems use to learn the patterns of human speech and the sounds of different words and phrases. [4]

The size and composition of a speech corpus can have a significant impact on the performance of an ASR system. In general, larger corpora tend to produce better results, as they provide the system with more examples to learn from. It is also important for the corpus to be diverse, representing a wide range of accents, dialects, and speaking styles. [7]

In addition to being used for training ASR systems, speech corpora are also often used to evaluate the performance of different systems (see 2.3 ASR Systems Evaluation). This can involve comparing the output of an ASR system to the ground truth transcriptions in the corpus, to see how accurately the system is able to transcribe the words being spoken.

Overall, speech corpora are a key tool in the development of speech recognition technology and continue to play a crucial role in advancing the field.

2.2.1 Common Voice 11.0 (CV)

In the following Figure 1 all information about the Common Voice 11.0 (CV) corpus can be found for both languages, French and Italian. More details such as number of samples per gender and age can be found in chapter 4.1.

	GENERAL INFORMATION
Summary	Largest open-source multi-language voice dataset based on voices of volunteer contributors. A sentence or word is displayed on the screen and the volunteer read this text aloud. No specific domain of sentences defined.
URL	https://commonvoice.mozilla.org/en/datasets
Owner / Authors	The Mozilla Foundation
License	Creative Commons Attribution Share-Alike 3.0 Unported license; https://www.mozilla.org/en-US/foundation/licensing/website-content/
	CORPUS PROPERTIES
Speaking Style	Scripted, monologue read-aloud speech
Accented Speech	Unknown
Dialectical Variation	No
Overlapping Speech	No
Speaker Noise	No
Acoustic Environment	Unknown
Recording Device	Unknown
	TRANSCRIPTION INPUTS
Reference	Segmented on speaker utterance level (json per utterance)
Audio	Segmented on speaker utterance level (mp3 per utterance)
Metadata	Segmented on speaker utterance level (json per utterance)
	TESTSET
Test Set Definition	Subset of default test set (see chapter 3.1.3)
Test Set Duration	8.4 hours (it), 8.3 hours (fr)
Number Utterances	4889 (it), 5172 (fr)
Number Speakers	1779 (it), 2338 (fr)
Average Utterance	6.2 sec (it), 5.8 sec (fr)

Figure 1 Detailed overview of Common Voice 11.0 (CV) Corpus (French and Italian)

2.2.2 Multilingual Librispeech (MLS)

In the following Figure 2 all information about the Multilingual Librispeech (MLS) corpus can be found for both languages, French and Italian. More details such as number of samples per gender can be found in chapter 4.1.

	GENERAL INFORMATION
Summary	Multilingual LibriSpeech (MLS) dataset is a large multilingual corpus suitable for speech research. The dataset is derived from read audio-books from LibriVox and consists of 8 languages
URL	https://www.openslr.org/94
Owner / Authors	Vineel Pratap, Qiantong Xu, Anuroop Sriram, Gabriel Synnaeve and Ronan Collobert
License	Creative Commons Attribution 4.0 International, https://creativecommons.org/licenses/by/4.0/
	CORPUS PROPERTIES
Speaking Style	Scripted, monologue read-aloud speech
Accented Speech	Unknown
Dialectical Variation	No
Overlapping Speech	No
Speaker Noise	Yes
Acoustic Environment	Unknown
Recording Device	Unknown
	TRANSCRIPTION INPUTS
Reference	Segmented on speaker utterance level (json per utterance)
Audio	Segmented on speaker utterance level (.flac per utterance)
Metadata	Segmented on speaker utterance level (json per utterance)
	TESTSET
Test Set Definition	Default test set
Test Set Duration	5.3 hours (it), 10.1 hours (fr)
Number Utterances	1262 (it), 2426 (fr)
Number Speakers	10 (it), 18 (fr)
Average Utterance	15 sec (it), 14.9 sec (fr)

Figure 2 Detailed Overview of Multilingual Librispeech Corpus (French and Italian)

2.2.3 Fleur

In the following Figure 3 all information about the Fleur corpus can be found for both languages, French and Italian. More details such as number of samples per gender can be found in chapter 4.1.

GENERAL INFORMATION	
Summary	Fleurs is a new multilingual speech understanding evaluation dataset in 102 languages.
URL	https://arxiv.org/abs/2205.12446
Owner / Authors	Alexis Conneau, Min Ma, Simran Khanuja, Yu Zhang, Vera Axelrod, Siddharth Dalmia, Jason Riesa, Clara Rivera and Ankur Bapna
License	Creative Commons Attribution 4.0 International, https://creativecommons.org/licenses/by/4.0/
CORPUS PROPERTIES	
Speaking Style	Scripted, monologue read-aloud speech
Accented Speech	Unknown
Dialectical Variation	No
Overlapping Speech	No
Speaker Noise	Unknown
Acoustic Environment	Unknown
Recording Device	Unknown
TRANSCRIPTION INPUTS	
Reference	Segmented on speaker utterance level (json per utterance)
Audio	Segmented on speaker utterance level (.wav per utterance)
Metadata	Segmented on speaker utterance level (json per utterance)
TESTSET	
Test Set Definition	Subset of default test set
Test Set Duration	3.5 hours (it), 2 hours (fr)
Number Utterances	865 (it), 676 (fr)
Number Speakers	346 (it), 332 (fr)
Average Utterance	14.7 sec (it), 10.4 sec (fr)

Figure 3 Detailed Overview of Fleur Corpus (French and Italian)

2.3 ASR Systems Evaluation

Evaluation of automatic speech recognition (ASR) systems is an important part of the development and improvement of these technologies. It involves assessing the performance of ASR systems and comparing different systems to one another, in order to determine their accuracy, efficiency, and overall effectiveness.

2.3.1 Evaluation Techniques and Metrics

Evaluation techniques and metrics are used to assess the performance of automatic speech recognition (ASR) systems. These techniques and metrics allow researchers and developers to compare different ASR systems and evaluate their accuracy and efficiency.

One common evaluation technique is to compare the output of an ASR system to a ground truth transcription of the same speech data. This can be done by calculating the error rate, which is the percentage of words in the transcription that are not accurately recognized by the ASR system.

Another common evaluation metric is the word error rate (WER), which is a more sophisticated measure of the accuracy of an ASR system. The WER takes not only the words that are recognized incorrectly into account, but also the words that are inserted, deleted, or substituted by the ASR system. (see 2.3.2 Word Error Rate)

In addition to these accuracy metrics, ASR systems may also be evaluated based on other factors, such as their speed and efficiency (see 2.3.3 Latency). This can involve measuring the time it takes for an ASR system to transcribe a given amount of speech data, as well as the amount of computational resources (such as memory and processing power) that the system requires.

Overall, evaluation techniques and metrics are important tools for assessing the performance of ASR systems and for comparing different systems to one another. Additional work is required to identify the nature and main sources of translation errors [8].

2.3.2 Word Error Rate

The WER is a more sophisticated measure of ASR accuracy than the error rate, which only considers the words that are recognized incorrectly. By accounting for additional errors, such as inserted or deleted words, the WER provides a more comprehensive picture of an ASR system's performance.

The WER is computed as follows:

$$WER = \frac{I + D + S}{\mathrm{N}} \times 100$$

where **I** is the number of insertions, **D** is the number of deletions, **S** is the number of substitutions and **N** is the total number of words in the reference transcription [9].

In general, a low WER is desirable, as it indicates that an ASR system is accurately transcribing the majority of the words in a given piece of speech data. However, it is important to note that the WER can vary depending on factors such as the quality of the speech data and the complexity of the words and phrases being spoken.

2.3.3 Latency

Latency is a measure of the delay between the time that an automatic speech recognition (ASR) system receives an input (i.e., a spoken utterance) and the time that it produces an output (i.e. a transcription of the utterance) [2].

ASR systems are typically designed to operate in real-time (streaming), meaning that the latency should be minimal and unnoticeable to the user. However, there are many factors that can affect the latency of an ASR system, such as the complexity of the speech data and the computational resources available to the system.

High latency can be a problem for use cases of ASR systems, as it can make the system feel slow and unresponsive to the user. This can lead to frustration and a poor user experience. It can also cause

problems in applications where real-time transcription is critical, such as in live transcription services (e.g. live subtitles) or voice-enabled virtual assistants [10].

To reduce latency, ASR systems may use techniques such as caching and parallel processing to speed up the transcription process. In addition, ASR systems may be designed to use less computational resources, such as memory and processing power, to improve their speed and efficiency [11].

Overall, latency is an important consideration in the design and evaluation of ASR systems, and efforts are ongoing to reduce latency and improve the speed and responsiveness of these systems.

2.3.4 Real-time-factor

The real time factor (RTF) is a metric commonly used to evaluate the speed of an ASR engine at run-time ("decoding phase"). As it is normalized by the audio duration it makes it easier to interpret and compare across different ASR engines [12].

RTF is defined as processing time divided by audio duration.

Usually, a state-of-the-art speech-to-text cloud-based service supplied by Google, Azure, AWS, etc. has values between 0.2 and 0.6 [12].

2.3.5 BeSTT

BeSTT is proprietary framework for the evaluation of STT engines jointly build and realized by CONTEXITY AG, SpinningBytes AG and Zurich Applied University of Science (ZHAW). This application is used to perform the evaluation of the ASR engines.

In Figure 4 you can see the entire workflow from obtaining data and transforming them into the unified format (see 3.5 Data structure and transformation) to ultimately getting the results of the evaluation in a unified, standardized format back.

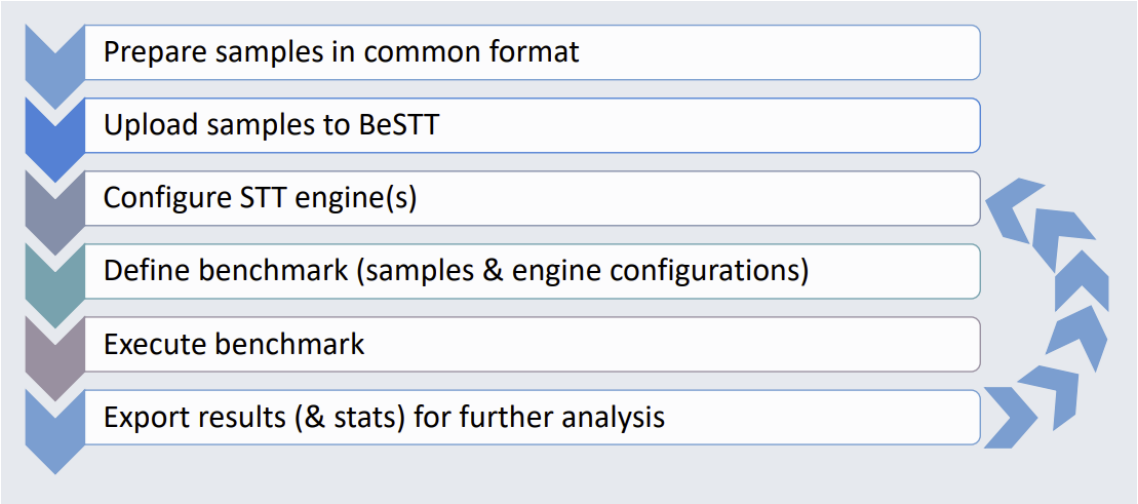

Figure 4 Visualization of the evaluation workflow on BeSTT [13]

Figure 5 shows the above workflow in terms of the system architecture of the plattform. We refrain from describing the exact mode of operation at this point, as this is irrelevant for the further work.

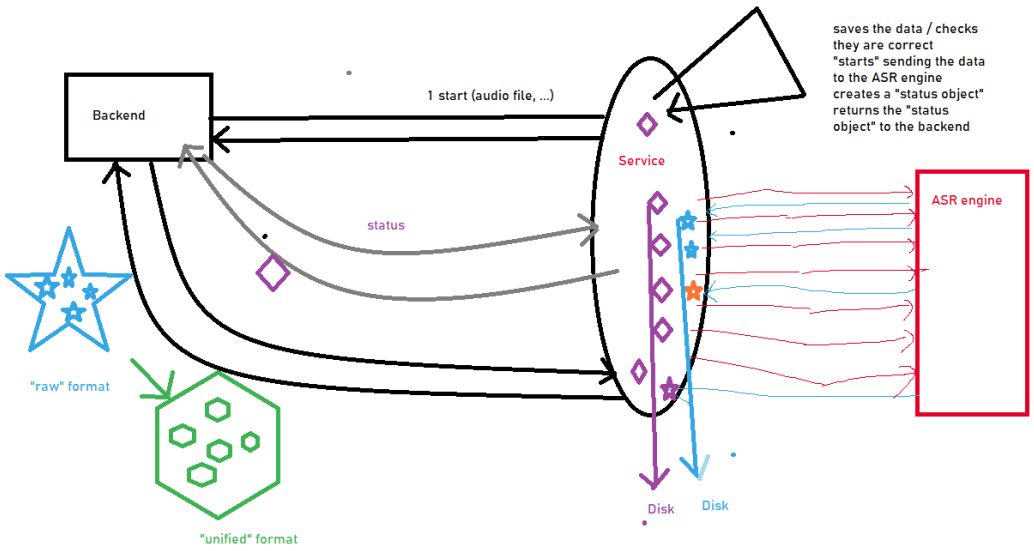

Figure 5 Sketch of the BeSTT system architecture (Source: Alexandros Paramythis)

The evaluation tool supports real-time API as well as batch execution. This can be specified during the configuration phase of ASR engine. For this evaluation the real-time API of the ASR engines was used to transcribe the audio samples (“Online” in terminology of BeSTT).

3 Experimental Setup

This chapter discusses the methods and criteria how the ASR engines and corpora were selected.

In this chapter it will be explained how the project was realized. It shows the criterions for the corpora selection and ASR engines, the reason why those had to be made and how the corpora were acquired.

Furthermore, the following data transformations are shown, how it was done, the format which had to be achieved, the statistics on how many utterances had to be deleted and how many utterances had to be cut out after transcription. The reasons behind the various benchmarks and why they had to be created will be explained as well.

3.1 Corpora

This chapter will focus on the statistics of all the corpora in this project (English, French and Italian). The newly found corpus in French and Italian will be presented as well as the method and procedure of acquiring the data.

3.1.1 Corpora selection criteria

The Corpora selection criterions had to be specified in advance. Learnings from CEASR 1.0 were considered and therefore certain corpora not used.

Great importance was put into the ground truth text and if it was manually corrected. Many open-source corpora transcribed the ground truth text via ASR engines and did not manually correct them. This means that even if the transcription in this project had been right, it would have been counted as a mistake (incorrect transcription) since the ground truth text had a faulty transcription. This potentially would have led to higher and faulty WER’s, therefore such corpora were not considered.

Since the English subset had to be kept identical to the CEASR 1.0 Version, so an appropriate comparison could be made, there were no new selection criterions differing from CEASR 1.0.

3.1.2 Corpora research and evaluation

Based on the selection criteria defined in chapter 3.1.1, various corpora were examined and evaluated (see Figure 6). This proved to be more challenging than anticipated. For French and Italian, in contrast to English, the availability of open-source quality data is considerably worse. In addition, many corpora (especially with spontaneous dialogues) have not validated the ground truth text and are thus not suitable for this project. Likewise, our efforts to connect with local associations in the field of Natural Language Processing and Speech Recognition have been unsuccessful.

KIP Parla corpus, CLIPS and Parole Publique did not contribute to CEASR 2.0 as no satisfactory answer from the corpus owner has been received by the time the evaluation has been performed.

Corpus	French	Italian	breach of guidelines	reason for exclusion	source
CLAPI	Yes	No	passive	annotation of groundtruth not consistent across utterances	http://clapi.ish-lyon.cnrs.fr/V3_Corpus_Liste.php?interface_langue=EN
CLIPS	No	Yes	none	not able to register and download	http://www.clips.unina.it/en/corpus.jsp
Parole Publique	Yes	No	none	no download link for audio files, non-responsive to multiple inquireies	https://www.info.univ-tours.fr/~antoine/parole_publique/corpus.
KIP parla corpus	No	Yes	none	not downloadable with the provided link; non-responsive to multiple	http://130.136.148.2/bonito/run.cgi/first_form
OGI Multilanguage Corpus	Yes	No	active	machine to human conversation	https://doi.org/10.35111/9bkm-qa61
CSLU: Multilanguage Telephone Speech Version 1.2	Yes	No	passive	similarity of utterances (e.g. weekdays), not possible to extract those with reasonable effort	https://doi.org/10.35111/j0p6-f049
2003 NIST Language Recognition Evaluation	Yes	No	active	fee of $500	https://doi.org/10.35111/38cj-3k75
ECI Multilingual Text	No	Yes	active	fee of $75	https://doi.org/10.35111/h2vd-p896
2008 NIST Speaker Recognition Evaluation	No	Yes	active	fee of $600	https://doi.org/10.35111/fyxw-v682
The SIWIS French Speech Synthesis Database	Yes	No	passive	consits of only 1 female speaker for 10 hr	http://datashare.is.ed.ac.uk/download/DS_10283_2353.zip
Augmented LibriSpeech	Yes	No	active	not validated groundtruth (i.e., translated)	https://persyval-platform.univ-grenoble-alpes.fr/datasets/DS91
MuST-C v1.0	Yes	Yes	active	not validated groundtruth (i.e.,	https://ict.fbk.eu/must-c-release-v1-0/
African Accented French	Yes	No	active	not validated groundtruth	http://www.openslr.org/57/
M-AILABS French-v0.9 Corpus	Yes	No	passive	subset of MLS	https://www.caito.de/2019/01/03/the-m-ailabs-speech-dataset/
Snips SLU Corpus	Yes	No	active	subset of MLS	https://paperswithcode.com/paper/snips-voice-platform-an-embedded-spoken/
TED-LIUM Release 3	Yes	Yes	active	not validated groundtruth (i.e.,	https://lium.univ-lemans.fr/en/ted-lium3/
M-AILABS Italian Corpus	No	Yes	active	subset of MLS	https://www.caito.de/2019/01/03/the-m-ailabs-speech-dataset/

Figure 6 Overview of examined but excluded corpora for Italian and French

3.1.3 Overview of selected corpora

Due to the lack of diversification across the corpora in Italian and French and Common Voice contributing 65% of audio samples, it was decided to randomly subset (reduced audio duration by 60%) the default test set of Common Voice to distribute the contribution of the single corpora more equally. This decision was supported also against the background of overall cost for the evaluation. For the remainder of the French and Italian corpora the full test dataset provided was applied. The English data sets were taken over from CESAR 1.0.

	default test set	after subsetting		samples applied to evaluation		attribution to	attribution to
French	[n size]	[n size]	[hr]	[n size]	[hr]	**total sample size**	**total audio**
Common Voice 11.0	12'930	5'172	8.32	3'285	5.44	*53.5%*	*32.6%*
MLS	2'426	2'426	10.07	2'361	9.80	*38.5%*	*58.7%*
Fleur	676	676	1.95	494	1.46	*8.0%*	*8.8%*
Total	**16'032**	**8'274**	**20.34**	**6'140**	**16.71**	***56.2%***	***55.8%***
Italian							
Common Voice 11.0	12'224	4'889	8.45	2'827	5.15	*59.2%*	*39.0%*
MLS	1'262	1'262	5.27	1'085	4.55	*22.7%*	*34.4%*
Fleur	865	865	3.52	864	3.52	*18.1%*	*26.6%*
Total	**14'351**	**7'016**	**17.24**	**4'776**	**13.22**	***43.8%***	***44.2%***

Figure 7 Default test set, subset and evaluation size for selected corpora in French and Italian

In Figure 7 above an overview of basic distribution from the data obtained by the corpus provider ("default test set), after subsetting Common Voice and remaining samples after the applied post-processing (see chapter 3.5.7) is provided.

An overview of all corpora can be found in Figure 8. A more detailed overview of the corpora for French and Italian can be found in Figure 1, Figure 2 and Figure 3 and for English in Appendix A.

Corpus Name	language	test Set	Test Set Duration	Speaking Style	Number of Speaker	accented speech	dialectal variation	overlapping speech
Fleurs	fr	Default test set	2h	Scripted, monologue	332	Unknown	Unknown	No
MLS	fr	Default test set	10.1h	Scripted, monologue	18	Unknown	Unknown	No
Common Voice 11.0	fr	Default test set	8.5h	Scripted, monologue	2338	Unknown	Unknown	No
Fleurs	it	Default test set	3.5h	Scripted, monologue	346	Unknown	Unknown	No
MLS	it	Default test set	5.3h	Scripted, monologue	10	Unknown	Unknown	No
Common Voice 11.0	it	Default test set	8.3h	Scripted, monologue	1779	Unknown	Unknown	No
AMI	en	Random selection	5h	Dialogue spontaneous speech	38	Yes	Yes	Yes
Common Voice	en	Default test set	5h	Monologue read aloud speech	Unknown	Unknown	Yes	No
LibriSpeech Clean	en	Default test set	5.4h	Monologue read aloud speech	49	Unknown	No	No
RT	en	Random selection	3.6h	Dialogue spontaneous speech	30	Yes	Unknown	Yes
ST	en	Random selection	4.7h	Monologue read aloud speech	5	Unknown	No	No
Switchboard	en	Random selection	0h	Dialogue spontaneous speech				
Tedlium	en	Default test set	2.6h	Monologue semi spontaneous speech	11	Unknown	No	No
Timit	en	Default test set	1.4h	Monologue read aloud speech	168	No	No	No
Voxforge	en	Default test set	3.9h	Monologue	171	Unknown	yes	No

Figure 8 Overview of selected corpora across all languages after subsetting Common Voice 11.0

3.2 ASR Engines selection

ASR engines were considered for this project which are popular in practice or have performed very well in previous evaluations. Primarily online research was conducted and input from Prof. Dr. Mark Cieliebak was considered. Although the comparison between open-source and proprietary systems is very interesting, it was decided to exclude them from the selection process, because the implementation of those is very complex and unrealistic for us to perform successfully.

Another criterion is the cost of transcription, as the project is subject to budget constraints. For this purpose, we have written to the eligible providers and clarified the costs and whether a discount can be spoken for academic purposes or to waive the fees altogether. In addition to the costs, the providers were asked whether they could be mentioned by name in the paper if the results were to be published. However, this was not an exclusion criterion. An overview of the different providers and details is not included in this paper but will be submitted separately with regards to confidentiality concerns.

3.3 Engine Implementation

At the beginning it was planned that this work should contribute to the implementation (under implementation we understand to integrate the engines' API in BeSTT and configure its model and parameters) of these investigated and selected engines. This initial approach was also pursued over several

weeks, but ultimately had to be aborted, because we simply lacked the necessary basic knowledge of computer science. For the scope of this work, we decided to use the already implemented engines from previous projects and not to add any new ones.

3.4 Selected ASR Engines and configurations

All engines selected are proprietary, state-of-the art and widely acknowledged as leading in the field of speech recognition. There are no further details available on the configuration and parameter settings of each engine, we were assured by the platform owner of BeSTT that if there is bias, this applies to all engines equally.

Due to confidentiality and legal concerns the engines used for the evaluation will not be mentioned by name. The engines hereinafter will be referred to as Engine 1, Engine 2, Engine 3 and Engine 4.

3.5 Data structure and transformation

The different corpora use different formats, provide various (different) metadata, and use different annotations. For comparability and to evaluate these samples on BeSTT, they need to be put into a unified format with uniform annotations as expected from BeSTT. The diagram below shows the data structure and attributes for the unified format (see Figure 9).

BeSTT expects this format and can only perform the evaluation if this format is followed. In addition, the metadata and the reference to the audio sample must be saved in a json file with the associated audio file in a zip file.

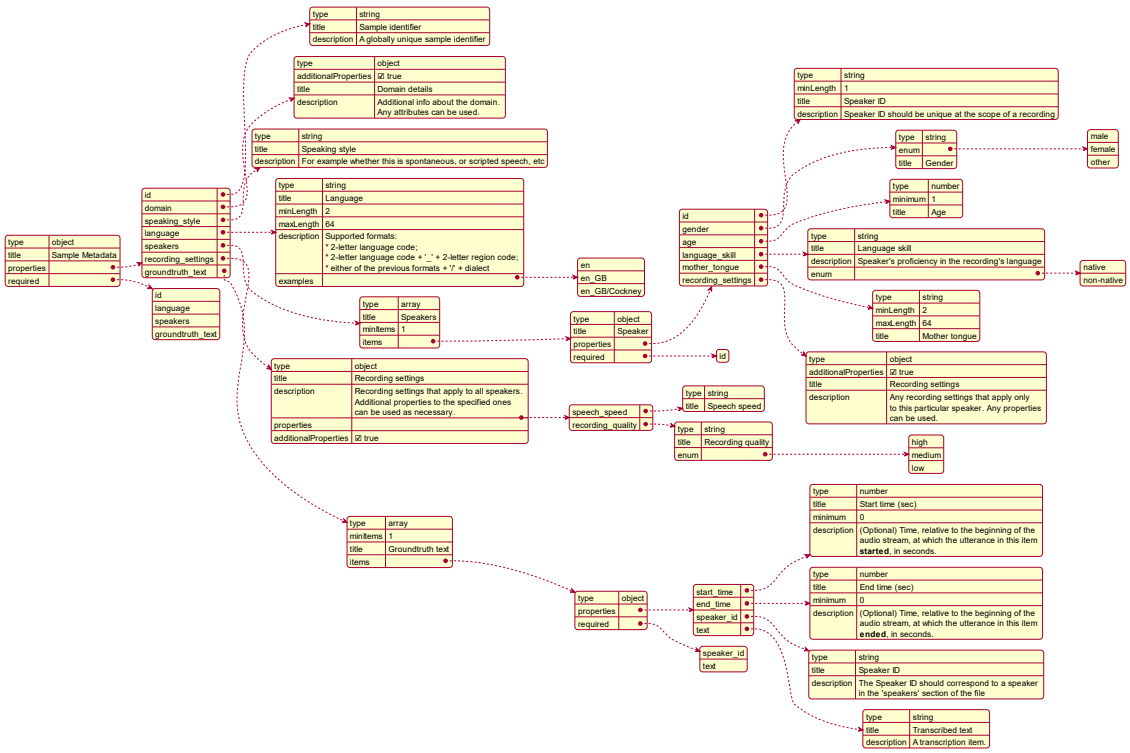

Figure 9 Visualization of the schemata of the unified format of CESAR

This was done for the French and Italian corpora. The English corpora were already available in the required format on BeSTT and have been adopted (as outlined in chapter 3.1.3). In contrast to the English data, the French and Italian corpora were not segmented because the length of the audio files did not make this necessary. For the English data this was done as BeSTT did not handle long duration audio well.

3.5.1 Pre-processing data checks

Prior to the evaluation, sanity checks were performed on the data. Here, using the French data from Common Voice 11 as an example, we first excluded data that had been rated as insufficient by other users (corpus provides columns with up & down votes). The corpus is an open-source project, and the ground truth text can be validated by other users, where this validation was not available or negative, the samples were excluded.

We also checked the correlation between audio duration and the length of the ground truth text. For this purpose, we considered the length of the string of the ground truth text with the duration of the audio file in seconds. Then, applying the z-score [14], we identified samples that differed by more than 0.5 (for MLS and Fleur 1) standard deviation from the mean. These were mainly short samples with an above average audio duration. In the two plots below (see Figure 10) one can see the impact of excluding these samples on the distribution.

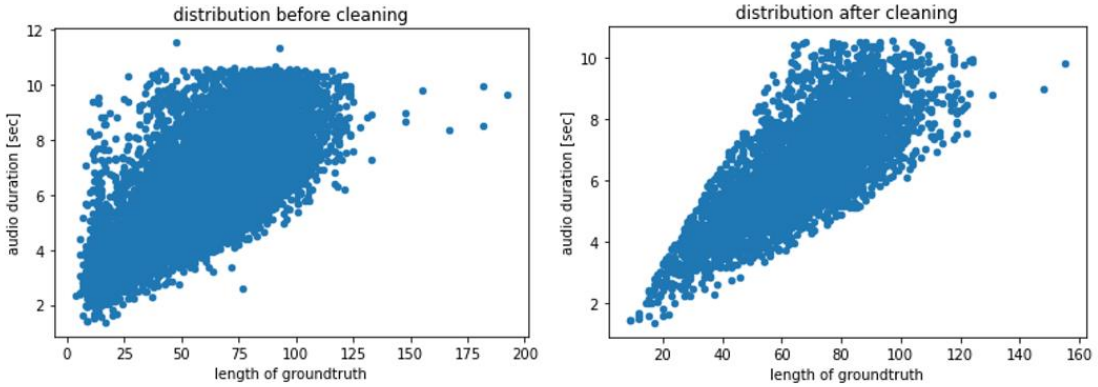

Figure 10 Distribution of French data for audio duration against length of ground truth text from Common Voice 11 pre and post cleaning applying z-score

Spot checks were performed on the excluded samples, and one could not detect any deviation between ground truth and audio for most of them, sometimes the audio was not cut correctly, which led to pauses at the end or beginning of the recording. Therefore, we decided not to exclude these samples and if necessary to exclude them in post-processing (see chapter 3.5.7) after we have findings about the evaluation results on this data.

3.5.2 Post-processing

The following chapters outlines the transformation / cleaning process. These processes were performed separately for French and Italian on both reference and hypothesis.

3.5.3 Decoding and encoding

The reference and hypotheses were transformed applying the python package unidecode to ensure that different encoding of the reference and hypotheses does not inflate the WER. Various approaches were evaluated to account for the French and Italian accent marks (e.g. special encoding / decoding such as “NKDF” and “latin-1”). Ultimately, it was decided to decode with the default package, this led to small loss of information but resolved the issue with different decoded characters in the reference and hypotheses text (see python code in Appendix C).

3.5.4 Contractions

Spaces occurring from contractions (i.e., c’ est to c’est) were removed with Regex. This yielded significant improvement for Engine 3’s performance, for the other engines this has had a positive impact but not significant (see python code in Appendix C).

3.5.5 Number to words

The hypothesis contained numbers such as 10 and the ground truth word-like numbers such as “ten”. With the python package num2words for French and Italian these occurrences were standardized to word-like numbers (see python code in Appendix C).

3.5.6 General string manipulations

The following string manipulations using Regex were performed (see python code in Appendix C):

- Removal of multiple spaces
- Removal of special characters ("$%&+")
- Removal of defined special characters ('!"#()*,./:;<=>?@[\]^_{|}~-')
- Abbreviations (e.g. “mme” for madame in French) were not removed.

3.5.7 Exclusion of bad samples

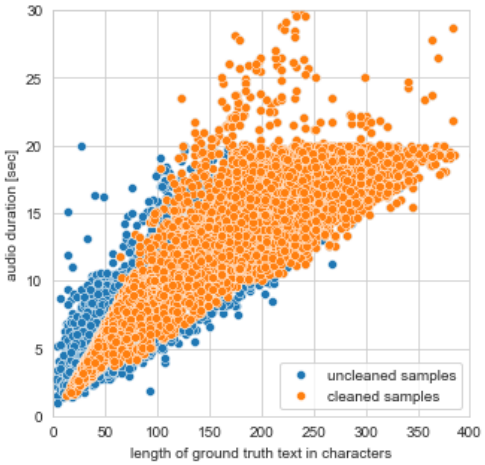

Figure 11 Distribution of audio duration and length of ground truth

The approach that was disregarded in chapter 3.5.1 to exclude certain samples on the basis of the z-score of the string length of the reference text and the duration of the corresponding audio file was applied in post-processing but with marginal (positive) impact on the results.

In Figure 11 the number of characters in the ground truth text is plotted against the audio duration in seconds. The orange datapoints are the samples after cleaning applying z-score. On these samples the evaluation will be performed. The blue datapoints represent the excluded samples. Primarily short samples were excluded from Common Voice, this cannot be traced back to specific root causes. It can be assumed that this was likely due to the volunteers not turning on / off the microphone when recording the sentence display to read-aloud. After the exclusion spot-checks were performed on the remaining samples with high WER (WER >0) These samples often have a non-native speaker (no metadata available on language skills in the corpora) or where the sentence was spoken very rapidly. For the very rapidly spoken texts, we assume that we were able to exclude most of them with applying the z-score. Also, the reference text and the audio were true and correct, but the engine had problems transcribing these due to the fast speech.

Fleur was located between Common Voice and MLS. Over spot checks the samples generally seemed to be cut off right, however the speakers talked rapidly, and some samples were given a lower volume level and therefore lacked quality.

Given the fact, that the data of MLS is derived from audio books and hence are very long recordings, they are by default segmented in 10 to 20 seconds segments and aligned with the ground truth. This led to a couple of audios which were longer than purposely intended and therefore, not fully matching transcripts. When spot checks were carried out even an audio was found where the speaker was singing, even though the speech type for the whole corpora was stated as scripted. This, however, was a single case and was cut out of the evaluation set as the ratio between the ground truth length and audio duration was outside the acceptable interval.

3.5.8 WER recalculation

After the post-processing the WER was recalculated on the transformed (i.e., postprocessed) reference and hypothesis text with the python package jiwer (see python code in Appendix C). This package and the methodology of retrieving the WER is consistent with the methodology of BeSTT (i.e., same settings and same package).

3.6 Benchmarking

BeSTT summarises the evaluation of data and engines up into benchmarks. There are several reasons for this; on the one hand, a lot of data is run at once, which can put a lot of strain on the infrastructure and slow it down. On the other hand, splitting the data into different benchmarks prevents the entire data from having to be run again in the event of errors in the communication between BeSTT and the engines' API.

An overview of the composition of the various data sets and engines can be found in the Appendix B. For French and Italian the benchmark consists of the entire data set and one engine. For English the data set was split into two parts, due to the size (reasoning as mentioned above). In total 16 benchmarks were created, 4 for French, 4 for Italian and 8 for English.

The benchmarks were executed in the beginning of December of 2022. Parallel execution is possible but limited to 1 engine at a time. This means it is only possible to run 4 benchmarks in parallel for our setup. Between one execution the system needed around 20 minutes to execute the next benchmark.

For Italian and French the benchmark execution was performed efficiently, and the results were available within a week. Due to incompatibility of the English data and BeSTT all results were only available after 3 weeks. Eventually the 8th English benchmark which belonged to Engine 3 could not be run successfully, the root cause is at the time of this project not known and could not be investigated in this short timeframe. Because of that Engine 3 was excluded from the English evaluation.

4 Results

The evaluation was performed with two different objectives. The evaluation of the French and Italian dataset is focused on the comparison between the engines across the selected corpora and languages and aims to identify and explain variance and investigating the influence of the available metadata on these measures. The evaluation of the English dataset focuses on comparing the results to prior evaluations performed in 2019 [9] and mid-2022 [13] on the same samples (or similar for 2019) with the same engines (exact model configuration unknown)).

We have no specifics on how the engines were configured and the parameters were set, as this was not part of this project (see chapter 3.4). All executions of benchmarks were using real-time APIs of the engine (i.e., "online" in terms of vocabulary of BeSTT).

4.1 French and Italian Data Sets

Figure 12 provides an overview of the WER for each corpus and each engine as well as the mean WER per corpus and engine. A distinction is made between the means, mean per corpus builds the average of the WER score from each corpus and mean over all utterances builds the average by accounting for all samples. The best result for French was obtained by Engine 4 with the MLS corpus. For Italian the best result was obtained by Engine 2 with the Fleur corpus. In general, it can be said that Italian performed better than French in terms of the WER.

Engine 1	common voice	fleur	mls	mean per corpus	mean over all utterances
French	0.47	0.44	0.37	0.43	0.43
Italian	0.32	0.26	0.33	0.30	0.31
Engine 4					
French	0.15	0.13	0.10	0.13	0.13
Italian	0.09	0.08	0.17	0.11	0.11
Engine 2					
French	0.28	0.26	0.24	0.26	0.27
Italian	0.11	0.07	0.30	0.16	0.15
Engine 3					
French	0.47	0.39	0.36	0.41	0.42
Italian	0.28	0.15	0.38	0.27	0.29
French Ø	0.34	0.31	0.27		
Italian Ø	0.20	0.14	0.29		

Figure 12 Table comparing WER for French and Italian of all corpora and engines

4.1.1 Number of Samples

The initial sample size is visualized in Figure 14. Fleur accounted for the smallest number of samples. For Italian 865 samples were transcribed whereas for French it was 676 samples. Common Voice dominated the number of samples. Since the aim was for all the corpora to be weighted equally the Common Voice corpus was subsetted. Originally Common Voice consisted of 12'224 samples for Italian and 12'930 for French, however, a random subset of 4'889 samples for Italian and 5'172 samples for French were transcribed. The corpus still has the highest number of samples, but this way its influence

and dominance could be significantly reduced. The initial (default test set) and cleaned/transformed data (samples applied to evaluation) size for all corpora can be seen in Figure 7.

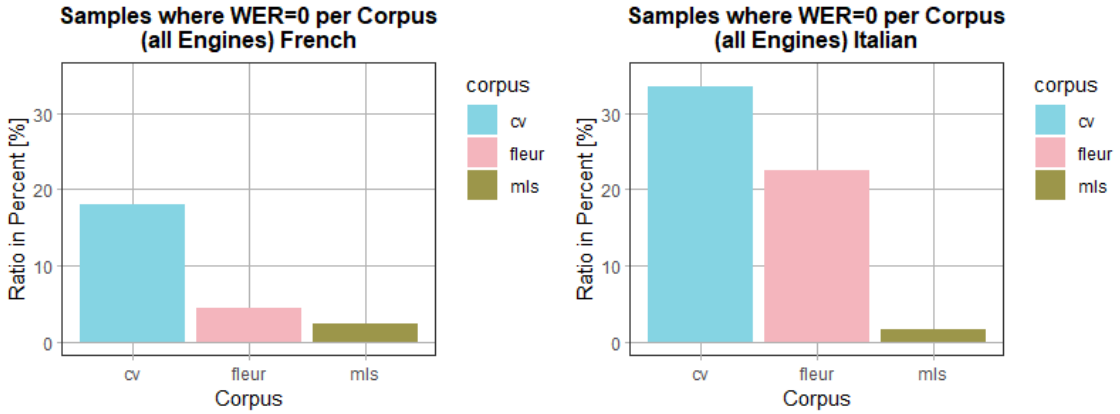

Figure 13 Samples where WER=0 per Corpus (all Engines) for French and Italian

In Figure 13 all the sample per corpus which obtained a WER equal to 0 were calculated and divided by the total number of samples. Illustrating the absolute numbers did not make sense in this context since Common Voice dominated with sample size and therefore, Fleur and MLS would have been too small to detect visually. In comparison from French to Italian it can be easily said that the general outcome was poorer for French. Only MLS' ratio for a WER equal to 0 was higher in French than in Italian. This aligns with Figure 16 where Common Voice and Fleur both obtain better results for Italian and MLS obtains better results for French. This might be traced back to the number of samples per corpus in Figure 14 where we could see that MLS provided less data in Italian.

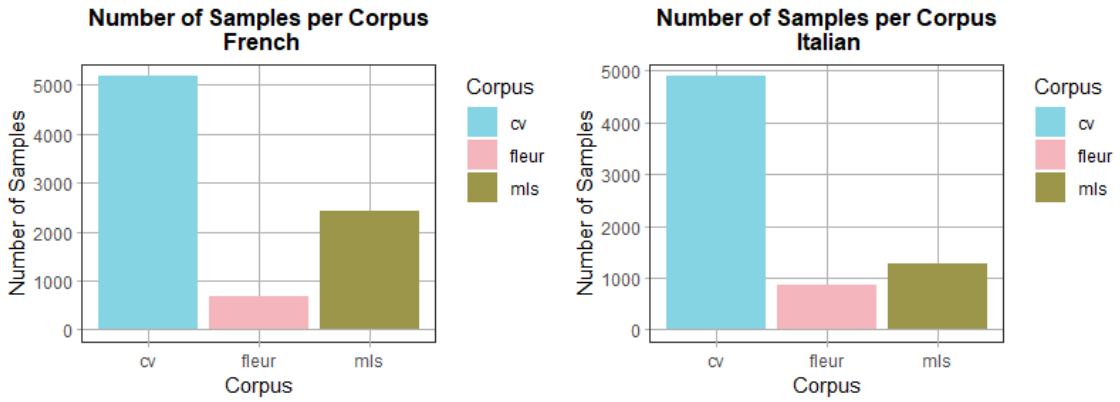

Figure 14 Number of Samples per Corpus for French and Italian which were transcribed

4.1.2 WER per Corpus and Engine

In Figure 15 the WER mean per Corpora and Engine can be seen. For French the WER per Corpora decreases from Common Voice to MLS almost linearly for every engine. Engines 1 and 3 performed similarly, Engine 2 starts at a lower level, and decreases from Common Voice to MLS with a lower slope. Engine 4 performs the best and shows a similar linear fall from Common Voice to MLS. A different shape is detected for the Italian data. Whereas for French Common Voice performed the poorest, now for Italian MLS shows the highest WER for all Engines. After MLS follows Common Voice and the best results are obtained by the Fleur corpus.

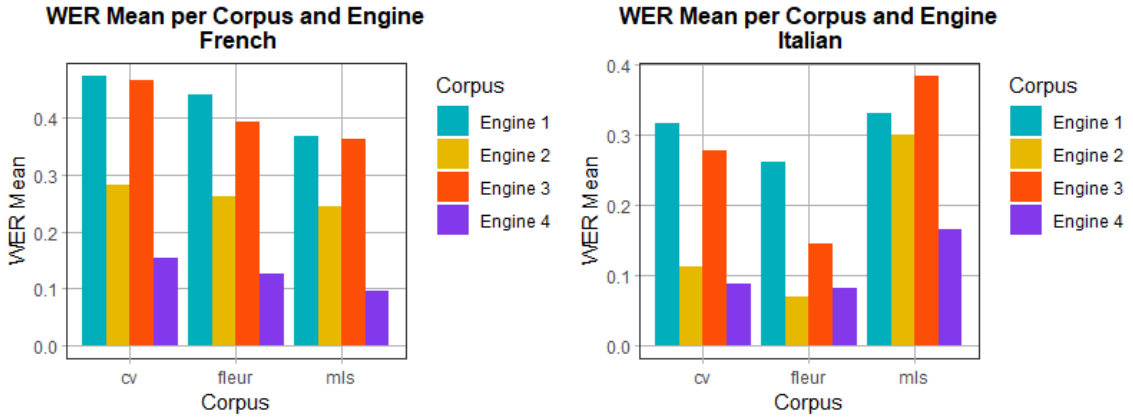

Figure 15 WER Mean per Corpus and Engine for French and Italian

The Figure 16 compares the WER for each engine for each language. The best overall results were received by Engine 4 followed by Engine 2. The poorest results were obtained by Engine 1 which was slightly worse than Engine 3. The performance was except for the level the same for both engines, however, Engine 4 shows many outliers for French. The cause of this might be led back to the observation which will be explained further in Chapter 4.1.3 (Audio Duration and Transcription Length) that French speakers talked faster and therefore were transcribed poorer.

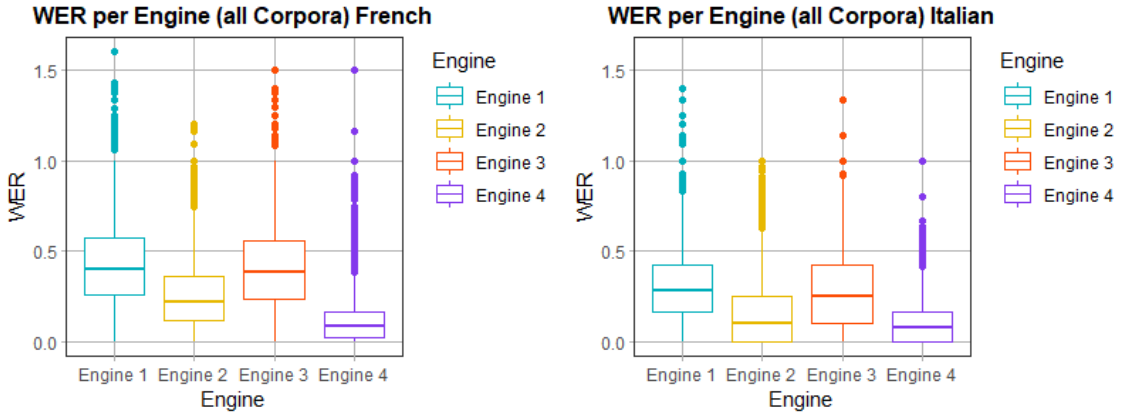

Figure 16 WER per Engine (all Corpora) for French and Italian

Figure 17 illustrates the same as Figure 16 however, the WER is now partitioned by corpus. For French the Engines show similar results for each corpus with a slight decline from Common Voice over Fleur and eventually MLS. Different for Italian. The decline now starts with MLS and goes from Common Voice to Fleur. This picture can bis seen for each engine except for Engine 2, where Common Voice achieved the best results. Still, it does not go unmentioned that in comparison to French MLS half as much data was available for Italian MLS, therefore smaller data size might have had a negative impact for Italian MLS.

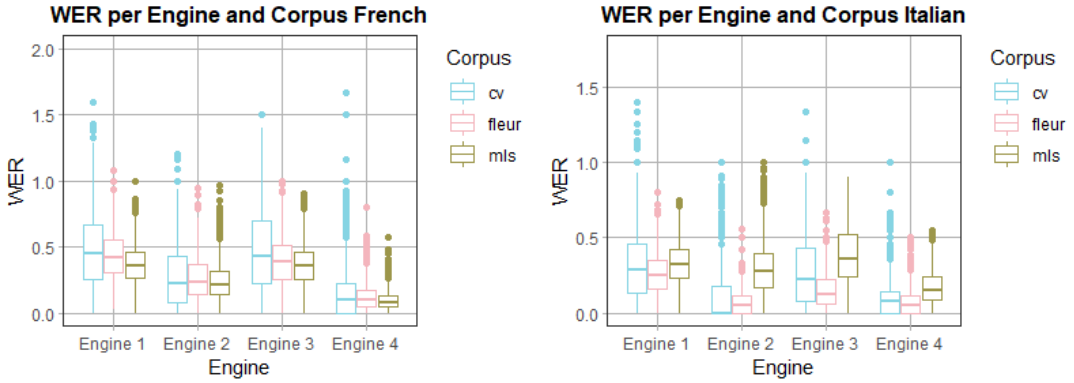

Figure 17 WER per Engine and Corpus for French and Italian

4.1.3 Audio Duration and Transcription Length

In Figure 18 and Figure 19 two similar shapes are illustrated. On one hand the corpora themselves form three clusters for every single graph. On the other the general shape indicated that with increasing sample duration and transcription length the WER decreases. In Figure 18 it is illustrated distinctly where the audios of Common Voice and MLS were cut. Most of the audios of Common Voice were all around 5 to 8 seconds long which can be led back to the data collection where people were asked to read a sentence which was shown on a screen. Unlike Common Voice, MLS cut their audios after every 20 seconds. This also can be led back to the data collection where the speakers were asked to read longer texts aloud and were cut into segments post processing. Consulting Figure 20 and Figure 22 it can be seen that the audio durations of Common Voice and Fleur are more arbitrarily than for MLS. Since Common Voice had the shorter audio durations with less words being spoken in those, a higher WER can be detected. If on has two words in the ground truth, the probability of those two being transcribed wrong and receiving a WER of >1 is much higher than with a sample which consists of more words. Of the three plots of Audio Duration vs. WER in Figure 20 and Figure 22 Common Voice and Fleur show a downward trend with increasing audio duration. MLS, however, does not indicate such a shape, but rather a cluster within its audio duration interval of 10 to 20 seconds.

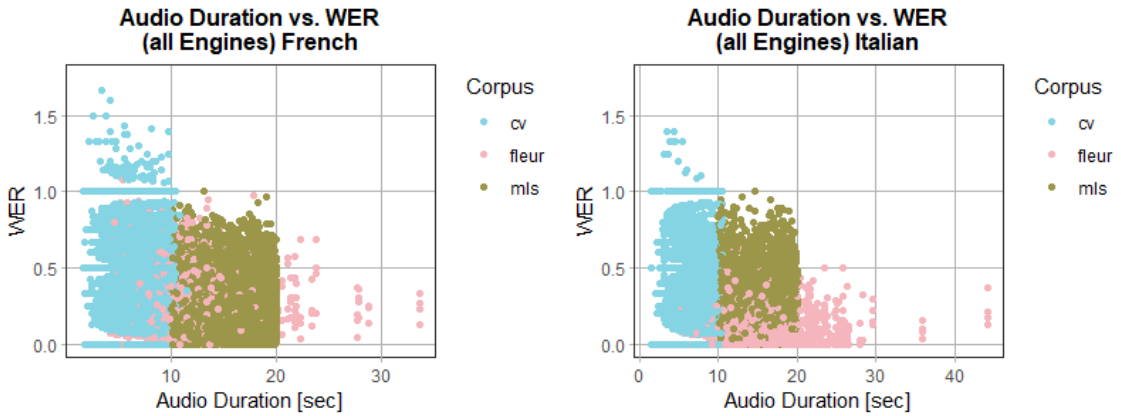

Figure 18 Transcription Duration vs. WER for French and Italian

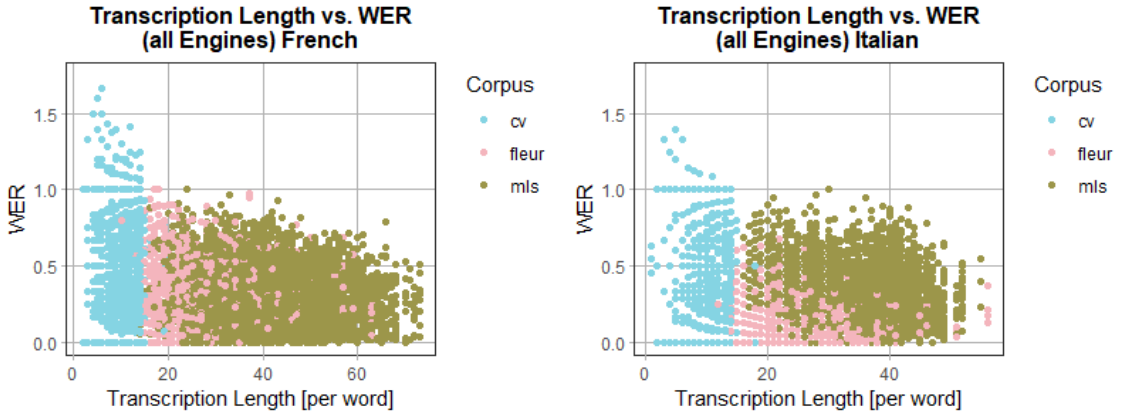

Figure 19 Transcription Length vs. WER for French and Italian

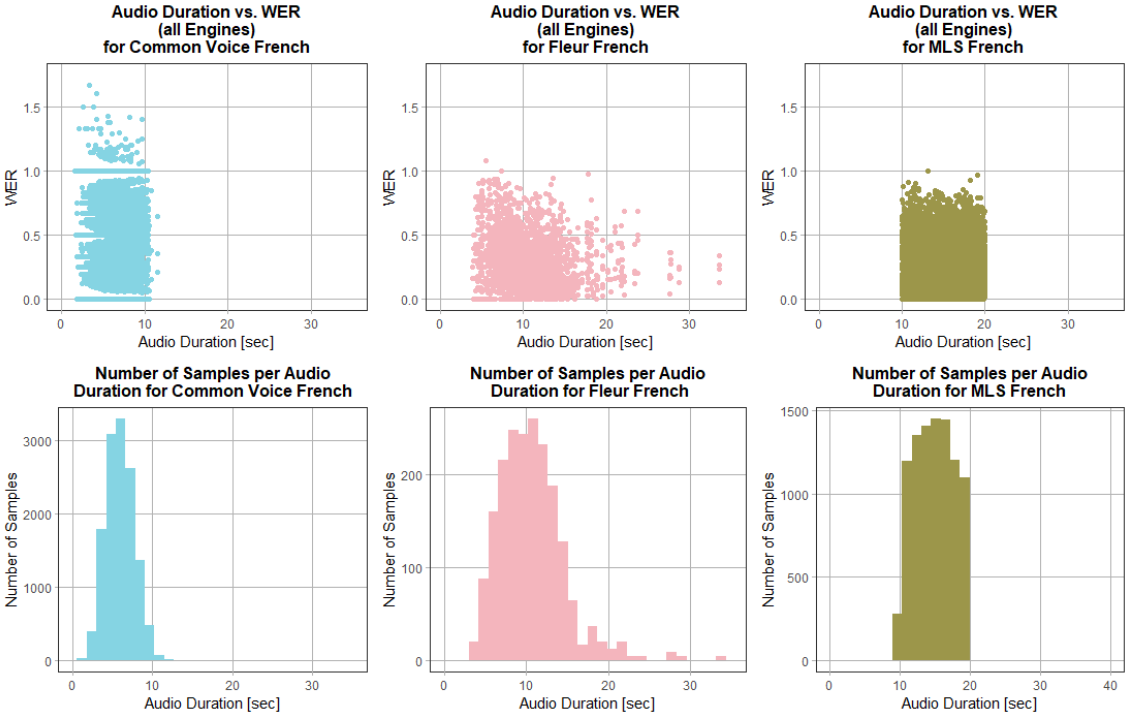

Figure 20 Audio Duration vs. WER separately for each corpus and Number of Samples per Audio Duration separately for each corpus in French

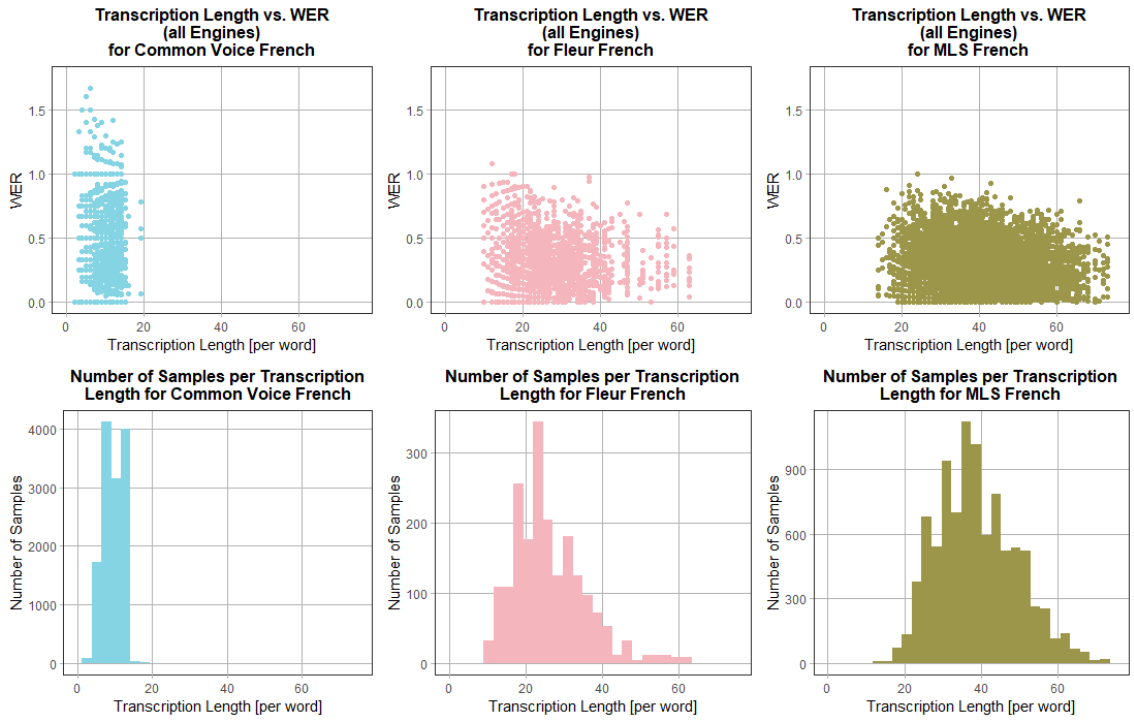

Figure 21 Transcription Length vs. WER separately for each corpus and Number of Samples per Audio Duration separately for each corpus in French

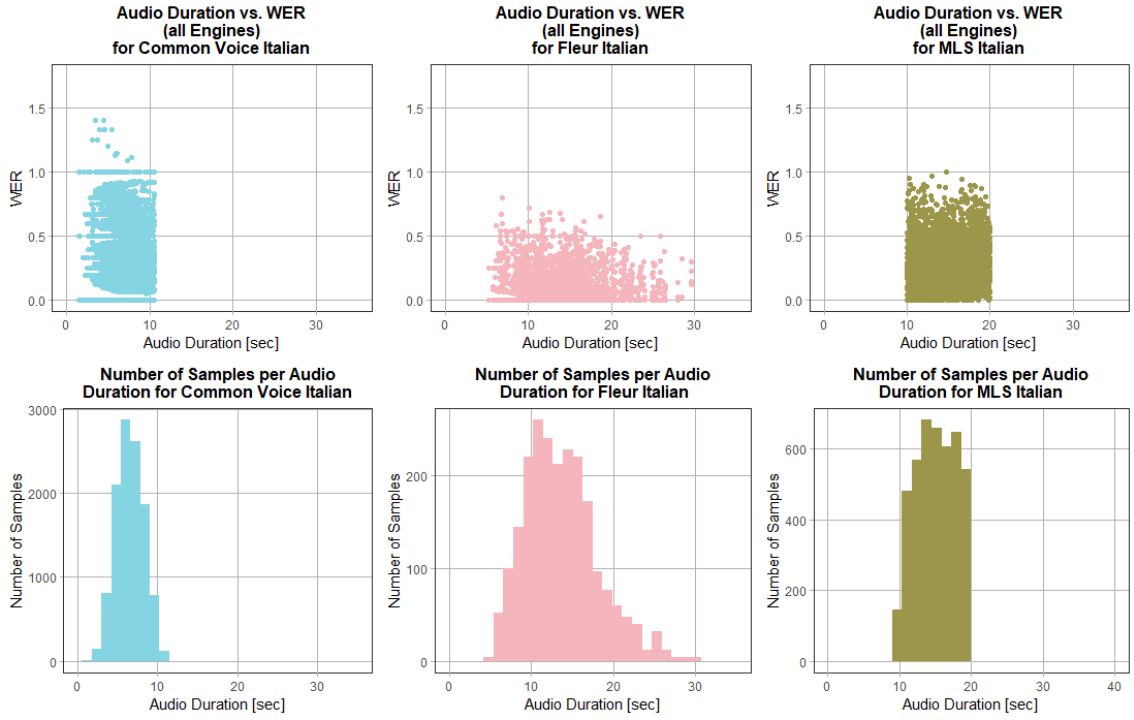

Figure 22 Audio Duration vs. WER separately for each corpus and Number of Samples per Audio Duration separately for each corpus in Italian

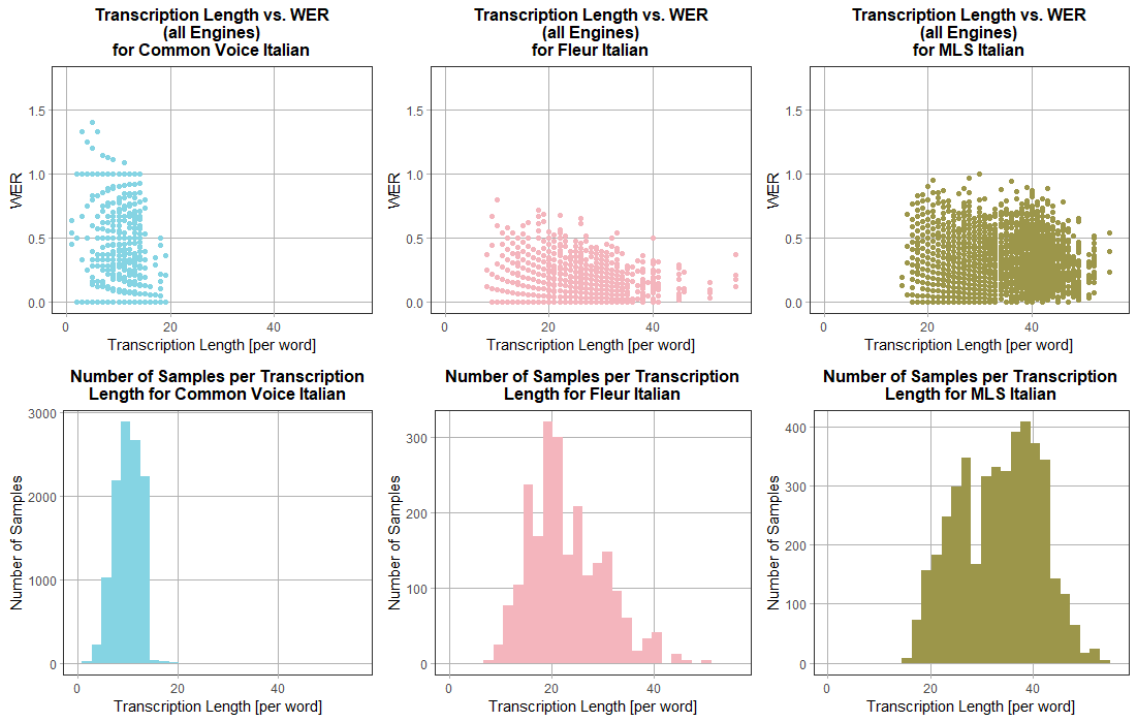

Figure 23 Transcription Length vs. WER separately for each corpus and Number of Samples per Audio Duration separately for each corpus in Italian

Overall in Figure 20, Figure 21, Figure 22, Figure 23 similar shapes can be detected for audio duration and transcription length spread. Common Voice provided the shortest samples, as its longest samples had a max. audio duration of 11.5 seconds in French and 10.5 seconds for Italian. Its longest transcription length consisted of 19 words for both French and Italian. Fleurs longest sample was 33.55 seconds long for French and 44.2 for Italian. The max. transcription length consisted of 63 words in French and 56 in Italian. MLS' samples were segmented post processing therefore its max. audio duration was 20 seconds for both languages, French and Italian. Again, the transcription length for French is longer with 73 words against 55 in Italian.

Considering the max audio duration for French and Italian for the corpora Fleur and MLS, it can be analysed that the ratio between max. audio duration and max. transcription length is lower for French. This indicates that the talking speed of the French speakers was higher, and they might have not pronounced their sentences clear enough which led to the audio being transcribed with higher WER than for Italian.

In the following Figure 24 the transcription length per word was grouped into the detected clusters from Figure 21 and Figure 23. The boxplot for French indicated a higher WER for a lower transcription length per word which stagnates after 3 – 8 words. Different for Italian, the boxplot shows a slight banana shape. The WER starts at a lower level than French and reaches its low at 3-8 words and then increases slightly again as the transcription length per word increases as well but forms an unexpected raise at the group +31. Looking at Figure 23 it can be seen that for increasing audio duration and transcription length per word, less data is available. In the group of 9 – 30 words the three corpora are still represented, however, for the group +31 only Fleur and MLS provided samples. With having less data for that part of audio duration and transcription length per word, follows that no concluding statement can be given. It is only to be speculated how the level would continue to be shaped, but it can be assumed that the WER level would stagnate at the level of the group 3 – 8 / 9 – 30.

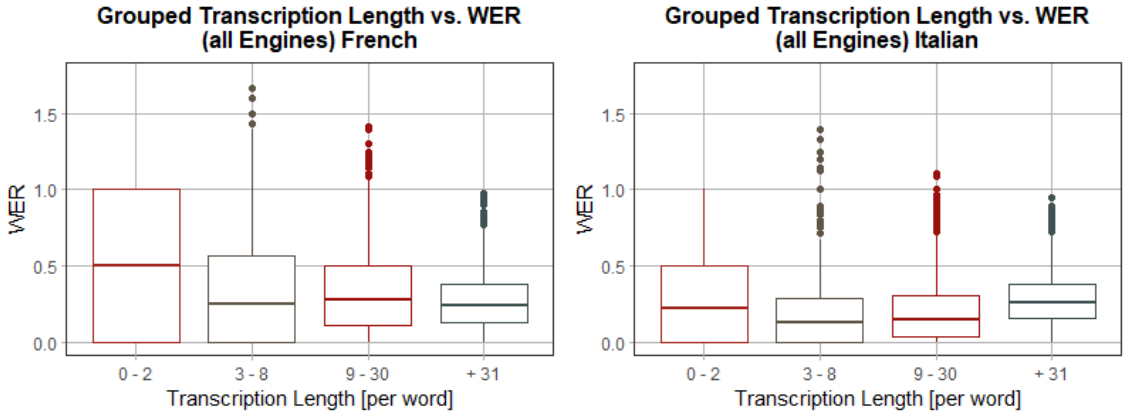

Figure 24 Grouped Transcription Length vs. WER (alle Engines) for French and Italian

4.1.4 Speaker and Gender

As already mentioned before, in comparison to Common Voice and Fleur, MLS initially had fewer transcription audios which were cut into short sequences of 20 seconds. MLS therefore had less speakers than Common Voice and Fleur. For French 18 speakers were transcribed and for Italian 10. In Figure Figure 25 the WER separately per engine for each speaker is shown. The plots return a different picture of the overall performance for each engine. Engines 1 and 3 again lay on a similar WER level, whereas Engine 1 shows the best performance and Engine 2 is situated between those two groups. Engine 2 received a significant outlier with speaker 4482. By listening to random samples of the speakers' audios it was noticeable that the audio lacked quality and the speaker partly mumbled words. He, however, was the only noticeable outlier. The performance for the speaker 2114 audios was only poor for Engine 1. As with speaker 4481 random samples were taken and listened into which showed that the quality was better, however, the speaker pronounced words unclear. Other than that, the results fluctuate with engine 4 having the smallest fluctuation and engine 3 and 4 the greatest. By colouring the speakers genders it also can be seen that the poorest performances for all engines, except for Engine 4, was obtained by male speakers and the best, except for Engine 4, as well by male speakers.

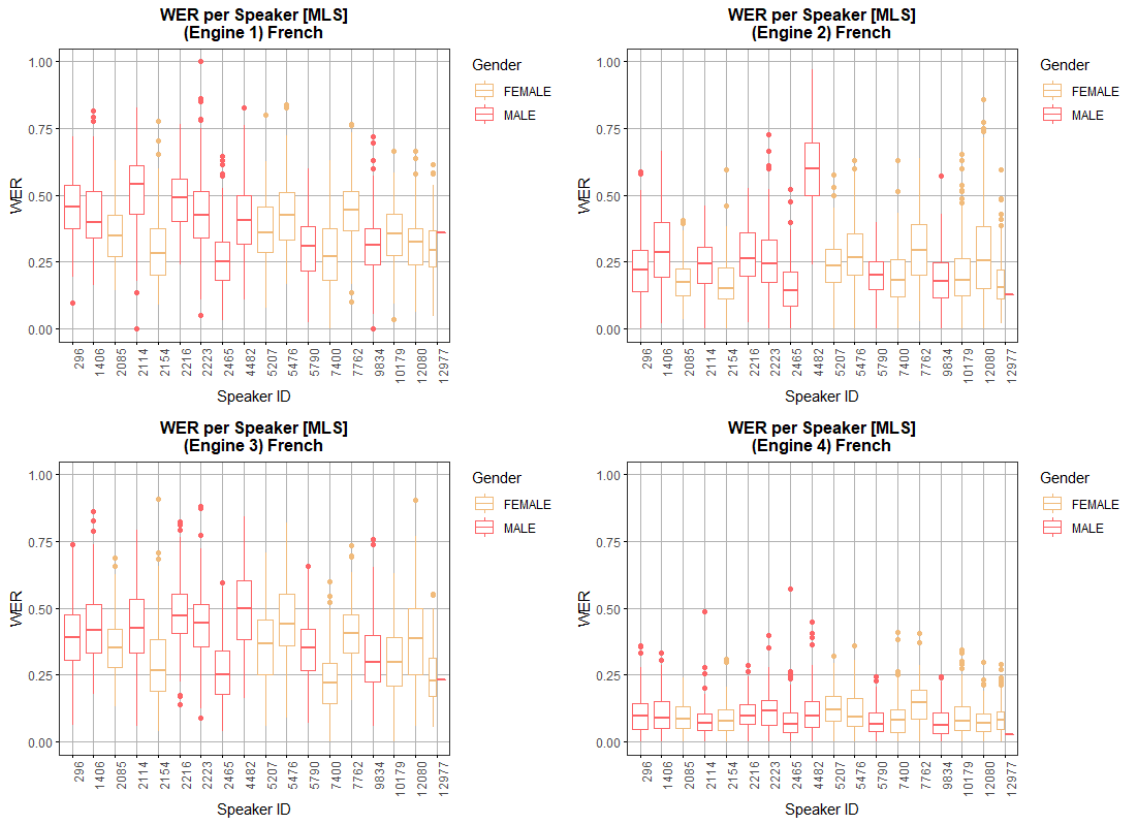

Figure 25 WER per Speaker [MLS] separately for each engine colored by gender French

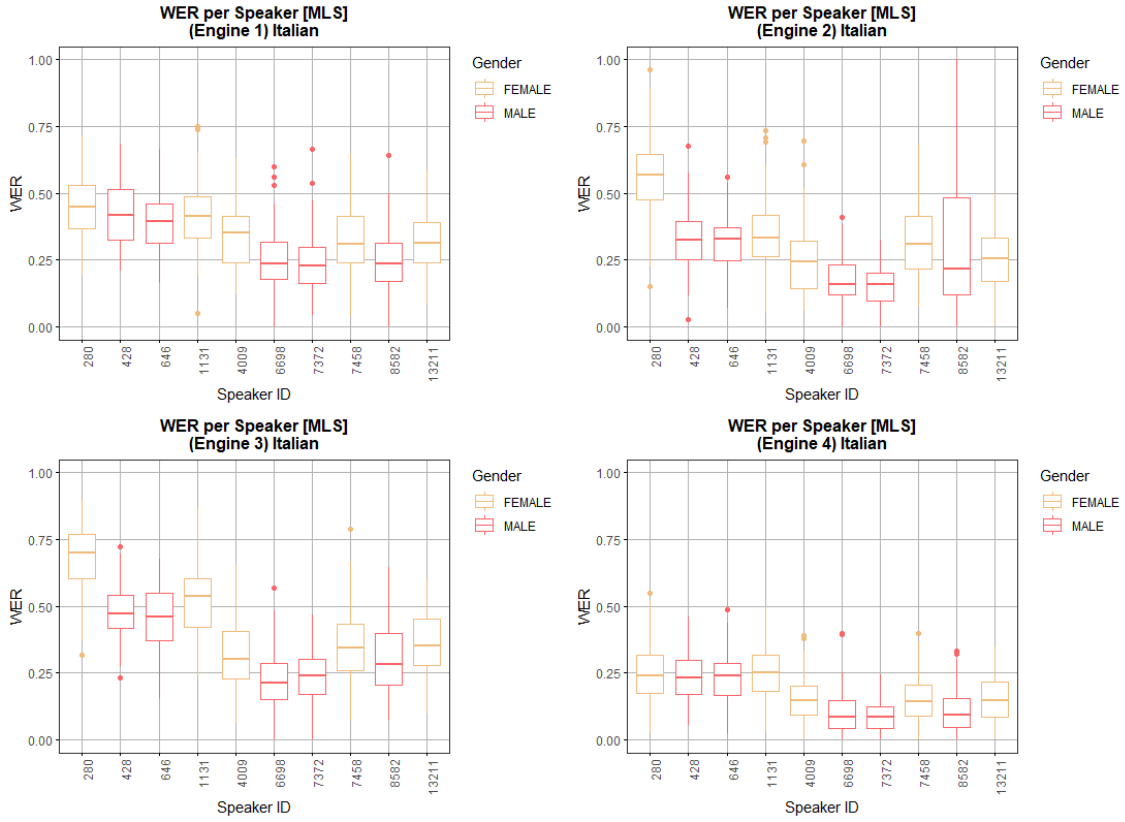

Figure 26 WER per Speaker [MLS] separately for each engine colored by gender Italian

Figure 26 illustrates the same picture as Figure 25, however now with the Italian Data of MLS. The most obvious outlier is speaker 280. She performs the worst for every engine. By listening to random

samples of the speaker it is noticeable that just as with the outlier of the French dataset, the audio lacks quality and the speaker mumbles even more. In contrast to Figure 25 now the poorest results are continuously obtained by female speakers and the best by male speakers.

The following Figure 27 illustrates the overall WER across each Gender. No notably deviation to each gender can be detected. Unknown Genders were excluded from this analysis.

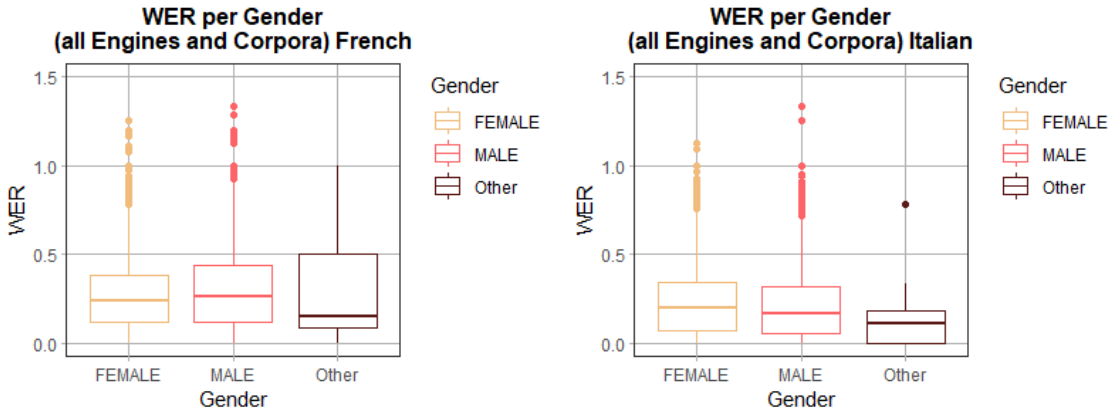

Figure 27 WER per Gender for French and Italian

In the following Figure 28 the results for each engine per gender are shown. For French it can be detected that male speakers obtained a higher WER for each engine followed by females and other genders. This shape can be seen for each engine except for Engine 4 where other genders performed the poorest. For Italian a slightly different picture can be seen. Females obtained the highest WER, followed by males and eventually other gender which aligned with the picture of Figure 26 where female speakers obtained the poorest performance. This shape can be identified for all engines except for Engine 3 where other genders performed the poorest.

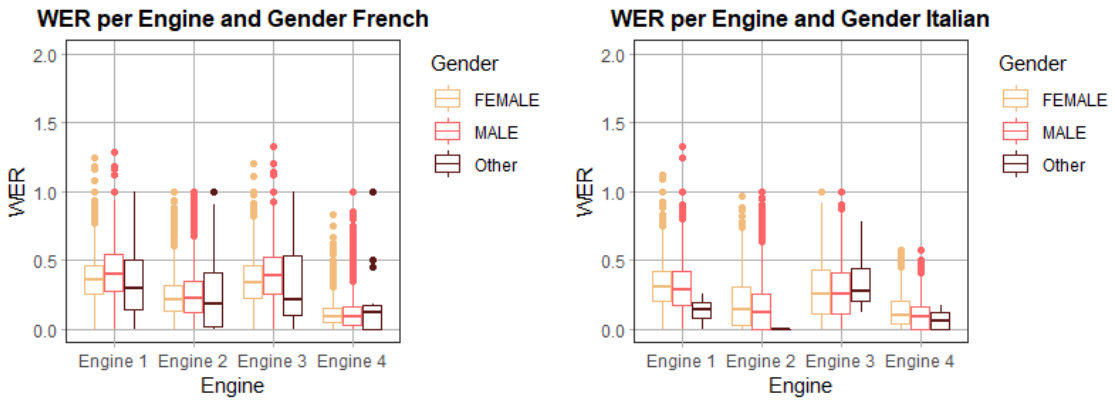

Figure 28 WER per Engine and Gender for French and Italian

Picking up the results of Figure 27 and Figure 28 it is of importance to show the provided number of samples for each Gender per Corpus which can be seen in Figure 29. Other genders are strongly underrepresented and the performance of those can therefore be assumed as not meaningful. Furthermore, a great part of the samples was labelled without a gender. In Addition, Fleur provided a low number of male samples for Italian and Common Voice dominated with the male samples.

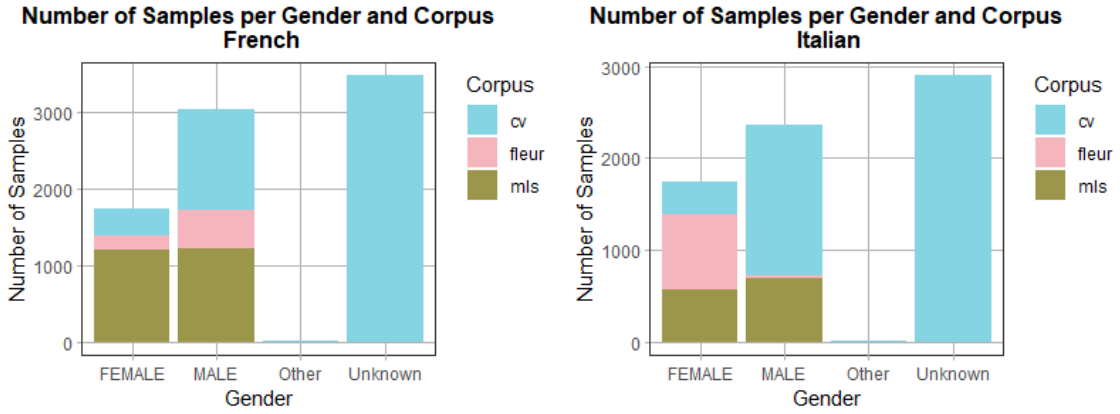

Figure 29 Number of Samples per Gender and Corpus for French and Italian

Considering the findings from Figure 29 it is plausible to consider the WER per Gender and Corpus as in Figure Figure 30. For each language a different shape can be detected. Common Voice shows in both languages a slightly poorer performance for female speakers. A Similar picture occurs for Fleur in French. Considering that the Common Voice corpus contains less data for female speakers than MLS which has overall a smaller sample size, it is difficult to determine whether the WER for female speakers for Common Voice would either in- or decrease with increasing female speaker sample size. The contrary is the case for Fleur Italian as male speakers obtain a higher WER. However, in Figure Figure 29 it was analysed that the sample number of male speakers in the Fleur Corpus is notably lower than for females, therefore no concluding statement can be made for Fleur in Italian. MLS shows a similar performance for both genders in French but returns a higher WER for female speakers in Italian. Since the number of samples for other genders is exceedingly low no conclusive analysis can be done.

No concluding statement can be done concerning the performance of the WER under consideration of gender, however overall corpora, and engines the results showed that there is no obvious and significant difference.

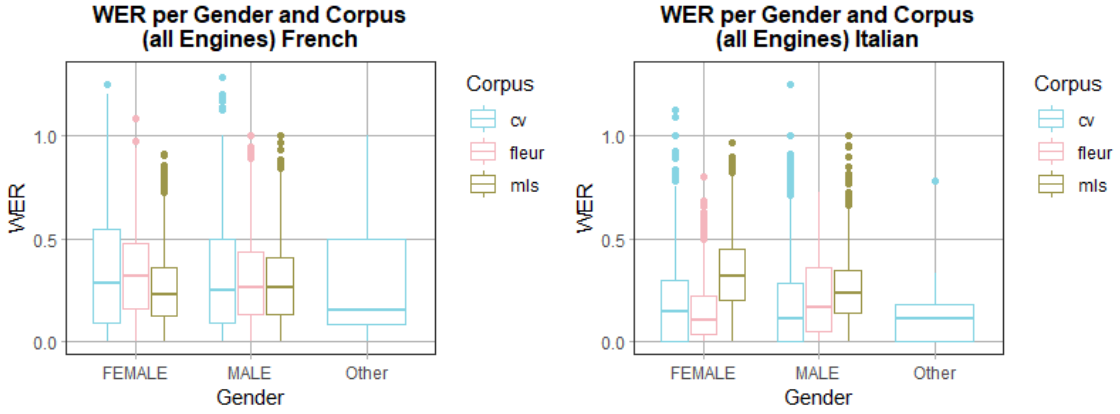

Figure 30 WER per Gender and Corpus for French and Italian

4.1.5 Age

Common Voice was the only corpus to provide the age decade the speakers were at the time of recording. In Figure 31 for French there a small-scale fluctuation to be noticed. For Italian there is no cluster or such as to be detected, therefore it can be assumed that age has no influence on the WER. The level of the WER replicates the general outcome where Italian obtained a better performance than French.

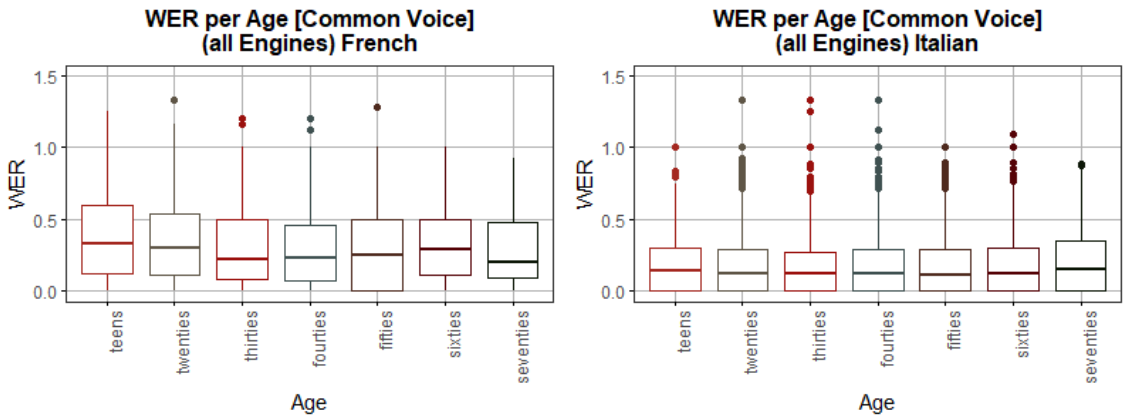

Figure 31 WER per Age [Common Voice] (all Engines) for French and Italian

In Figure 31 it could be observed that the overall level of the French WER is higher than for Italian. In Addition to that it can be seen that most of the Italian samples touch the target of WER=0. In Figure 32 the metrics are observed closer for both languages. Except for the level the shape for the Number of Samples per Age and the Number of Samples per Age where the WER is equal to 0 does not differ significantly for a particular age group. Putting these two number into a ratio and dividing the number of samples per age where the WER is equal to 0 by the total number of samples per age a new shape is detected for French. It can be seen that the ratio increases linearly from teens to fifties and eventually decreases to the same level as teens, however this partly might be led back to the number of samples as teens, sixties and seventies which performed the poorest are also the groups which have the lowest number of samples. Still for French this might indicate an influence by age. For Italian such an observation cannot be made as the ratio shows no positive or negative impact with age.

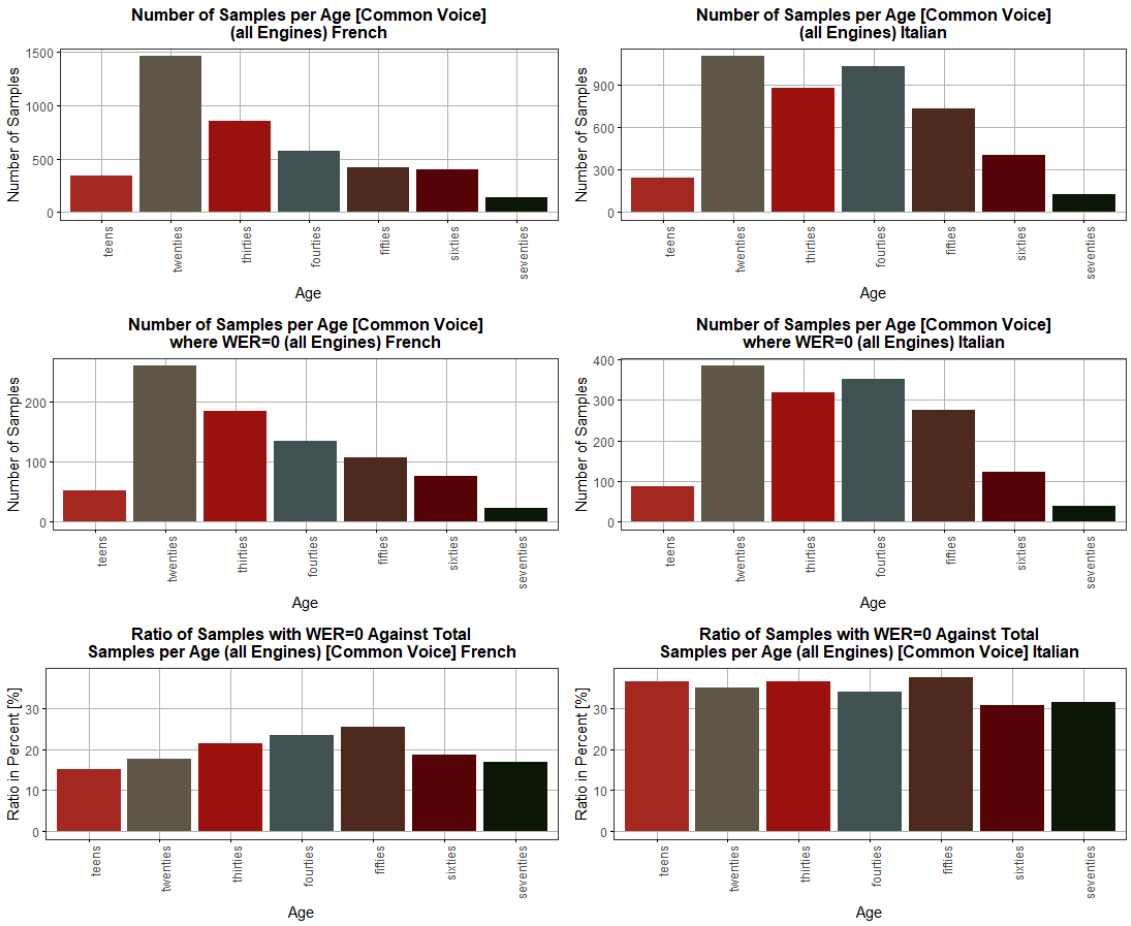

Figure 32 Number of Samples per Age [Common Voice] (all Engines) for French and Italian, Number of Samples per Age [Common Voice] where WER=0 (all Engines) for French and Italian and Ratio of Samples with WER=0 Against total Samples per Age (all Engines) [Common Voice] for French and Italian

4.1.6 Latency

The latency values applied to the computation of the RTF (see chapter 2.3.4) are the mean latency figures calculated by BeSTT from min and max latency measured during the benchmark execution. These mean latency values are not guaranteed to be exact but should be “close enough to reality” to apply comparative metrics such as RTF that are based on absolute latency numbers. Moreover, any bias they have is similar between the engines. Considering the above we will go ahead with using the RTF measure to compare the engines’ latency. In the following results latency refers always to mean latency of the evaluation results from BeSTT. Also please note that since the following figures were plotted in Python and the above in R they differ in appearance and color scheme.

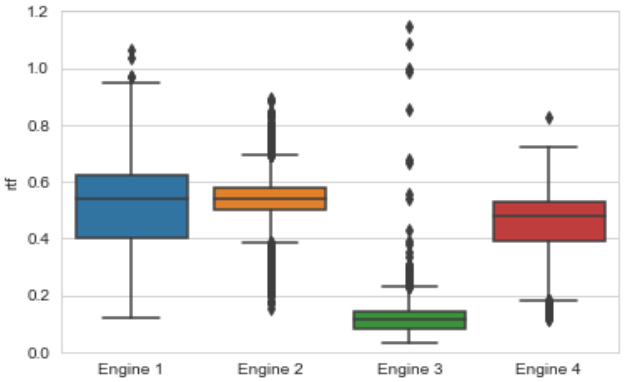

Figure 33 RTF grouped by engine for French data

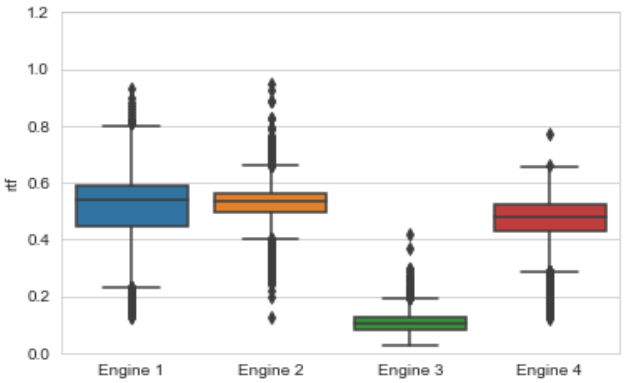

Figure 34 RTF grouped by engine for Italian data

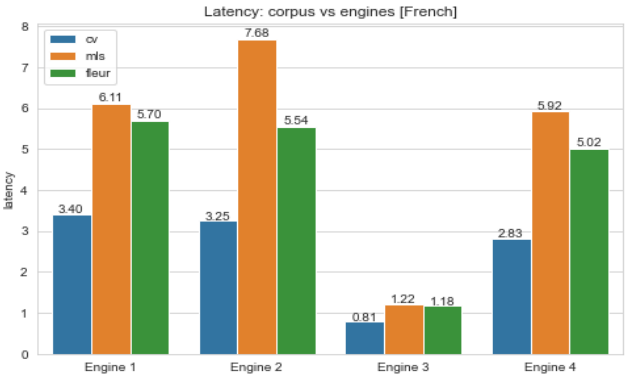

Figure 35 Barplot of engine's latency per French corpus

In Figure 33 it can be observed that Engines 1, 2 and 4 perform similarly in terms of median RTF around 0.5 and 0.6. Engine 1 has the highest variance, Engine 4 has slightly lower variance and Engine 2 has the lowest variance out of these three engines. Engine 3 has a lower median by factor 4 compared to the other engines and has also low variance. This is interesting as Engine 3 did not perform well in terms of WER.

In Figure 34 it can be observed that in terms of low RTF, variance the engines follow the same ranking as for the French data. RTF is very similar to the French data, although the variance of RTF is lower for Italian data compared to French. Since RTF is normalized by audio duration the different average audio duration for French and Italian does not affect the results and they can be compared to each other.

In Figure 35 it can be observed that that the sequence from low latency to high latency is the same for the 4 engines across three corpora, although on different levels and different increments between the corpora. To compare the latency fairly between the engines one need to compare them on the same corpus, as latency is influenced by the audio duration and the corpora have different average audio duration (see Figure 20 and chapter 2.2). The latency for MLS differed by a factor 6 between the best and worst performing engine.

Figure 36 cannot be directly compared to the Figure 35 as the data is not normalized, and audio duration is different for the samples of each language. Given that the data seems very similar to French. The order in the ranking of best to worst has changed. Engine 4 seems to be handling Italian worse than Engine 1 as they have switched places compared to French Data. Best performing engine remains Engine 3 by a large factor.

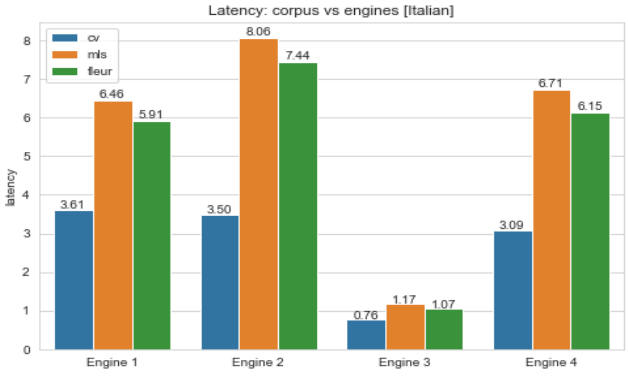

Figure 36 Barplot of engine's latency per Italian corpus

Figure 37 further visualizes the above findings. This plot is not ideal as some

data points are overlapping, nonetheless the key information we want to show is retained. One can nicely see that latency positively correlates with audio duration. This is true for all engines though at different slopes of correlation.

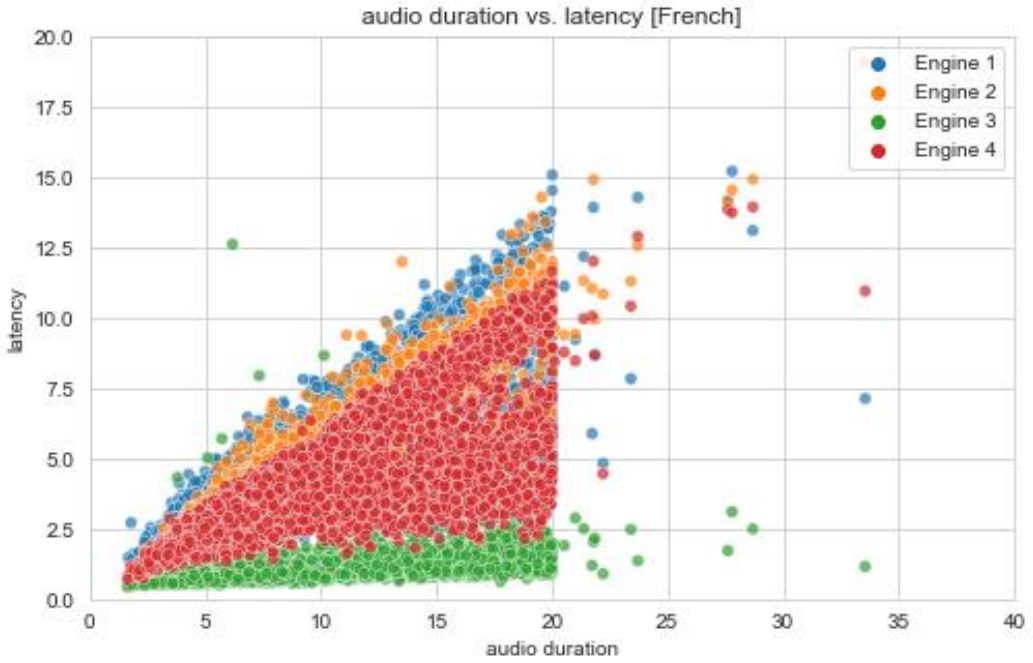

Figure 37 Scatterplot latency vs. audio duration on French data

In Figure 38 below the overall behavior is similar to Figure 37, but Italian has more data points with longer audio duration (see histogram Figure 20 and Figure 22). As shown in the histogram only fleur has provided Italian samples longer than 20 seconds, that is the reason that the levels of correlation shift. This is visible for Engine 1 and Engine 4, for Engine 2 this shift of correlation behavior cannot be observed, and the latency increases at a very similar slope as for the audio files with duration less than 20 seconds duration. This difference in latency mostly for long audio files of fleur is interesting between the Engine 1,2 and 4 and leaves room for further investigation on another project.

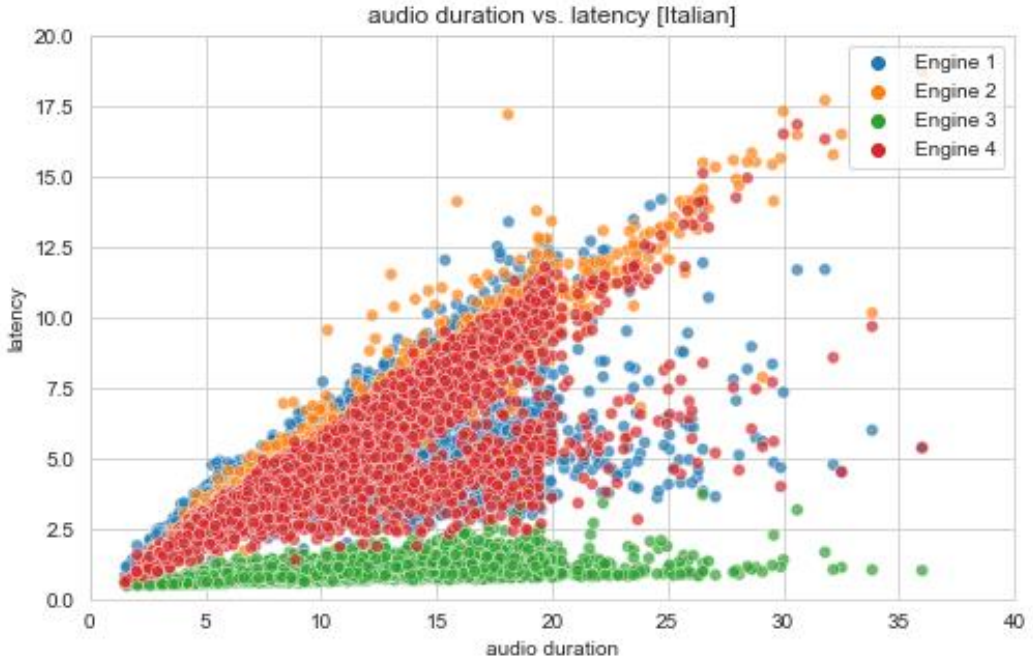

Figure 38 Scatterplot latency vs. audio duration on Italian data

4.2 English Data Sets

The evaluation of the English dat set (details to the individual datsets can be found in the Appendix A) focuses on the camparison between prior evaluation results of the same test sets or very similar with the same engines. The first evaluation was done in 2019 as part of a bachelor thesis (referenced in the following tables as “WER 2019”) [1]. Comparing the WER of this evaluation (referenced as “WER” or “WER 2022”) to the WER 2019 offers limited insight on the development of the respective ASR engines, as the evaluation was performed on different engine configurations and parameter settings. Additionally these configurations and parameter settings were optimized for the respective corpus, and the evaluation performed here uses the same confuguration and parameter settings for all corpora. The second evaluation was performed in mid-2022 (referenced as “WER 2022(1)”) for a publication at a conference with the same evaluation tool, parameter setting and engine configuration [13]. Definitons for the below mentioned metadata (i.e., language skill etc.) can be found in Figure 9.

Engine 3 is excluded from the evaluation of the English data sets, this is due to problems in communication between BeSTT and the engines’ API and no results of the evaluation could be extracted in due course to be considered in this project thesis. Also corpus switchboard was excluded from the evaluation due to very high WER of 0.87, this would need further investigation if this is related to the data or the configuration and this is not the scope of this work.

4.2.1 Overall performance

Figure 39 provides an overview of the WER for each corpus and each engine as well as the mean WER per corpus and engine. A distinction is made between “mean per corpus” and “mean over all utterances”. “mean per corpus” is the average of the WER scores from each corpus and “mean over all utterances” builds the average by accounting for all utterances across all corpora. The difference in these means can be explained that the worst performing corpus (i.e., ami and rt) are also the largest corpus in terms of utterances in the evaluation contributing together 40% of samples. Engine 4 performed the best across the corpora with the exception for corpora with spontaneous speech, where Engine 1 yielded the best results.

	ami	commonvoice	librispeech	rt	st	tedlium	timit	voxforge	mean per corpus	mean over all utterances
Engine 1 en-US	0.40	0.15	0.12	0.31	0.08	0.11	0.14	0.13	0.18	0.21
Engine 4 en-US	0.46	0.06	0.06	0.41	0.06	0.07	0.07	0.07	0.16	0.21
Engine 2 en-US	0.60	0.21	0.12	0.52	0.05	0.17	0.10	0.14	0.24	0.29
Ø	0.49	0.14	0.10	0.41	0.06	0.12	0.10	0.11	0.19	0.24

Figure 39 Overview of WER per engine per corpora

4.2.2 Speaking Style

In Figure 40 it can be observed that speaking style has a significant impact on the WER. Engine 1 handles spontaneous speech the best with a median WER of 0.25 and smaller variance than Engine 4. Engine 2 performed worse in recognizing spontaneous speech by a factor 2 in terms of median WER.

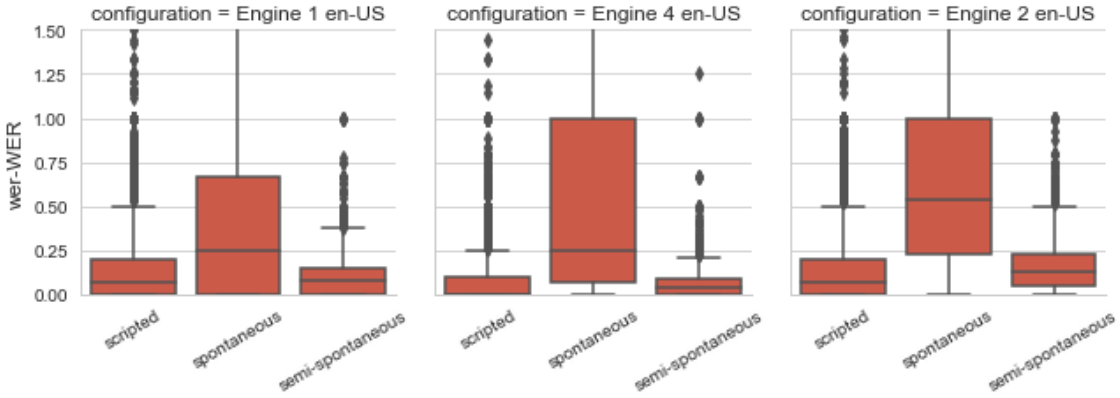

Figure 40 Boxplots of speaking style against WER per engine

This explains the poor performance of ami and rt in Figure 39, as these corpora are mostly spontaneous speech.

In terms of scripted and semi-spontaneous speech, Engine 4 outperforms Engine 1 and 2 in terms of median WER and variance (i.e., lowest variance). Engine 1 and 2 performed very similarly in terms of median WER and variance of WER.

4.2.3 Language Skills

In Figure 41 language skill is plotted against all samples and suggests that native is recognized better by the ASR than non-native.

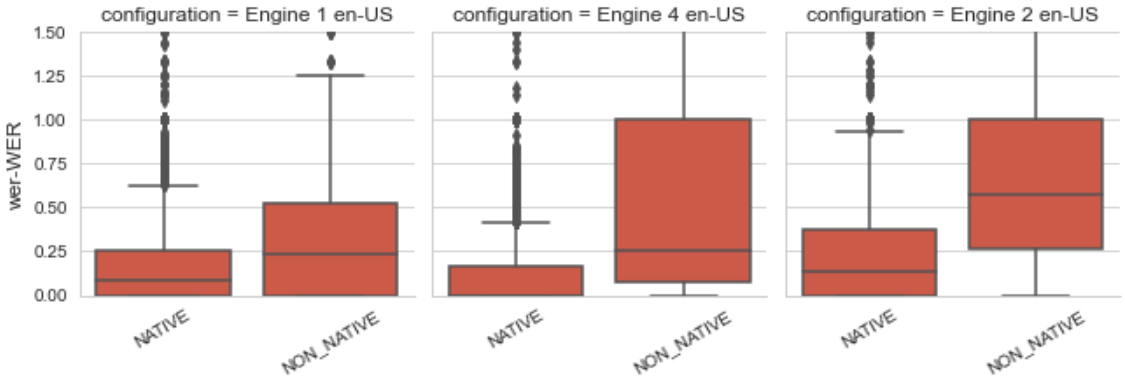

Figure 41 Boxplots of language skill of speaker against WER per engine across all corpora

As the plot has similar characteristics than Figure 40 we were interested which speaking style is represented in native and non-native. Speaking style and language skill was grouped and it could be observed that only spontaneous speech contains samples with the attribute non-native, while scripted speech only contains native samples, given the drastic difference of the speaking styles in terms of WER this is no fair comparison. In Figure 42 one can see scripted speech has only native language skills and semi-spontaneous speech does not have samples with this attribute.

speaking style	language skill	samples
scripted	native	24693
spontaneous	native	11880
	non-native	13827

Figure 42 Overview of number of occurrences of language skill in attribute speaking style

For a fair comparison of language skill, the data was filtered for spontaneous speech and the plot in Figure 41 was repeated. In Figure 43 it can be observed that non-native performs worse than native, but not as significant as first suggested from the plot in Figure 41. The overall increase in WER is expected as this data contains only samples with spontaneous speech and spontaneous speech has a high WER. To compare the performance in terms of WER and language skill on the filtered data one needs to be careful as the WER is affected by the difference of the ASR sensitivity to spontaneous speech.

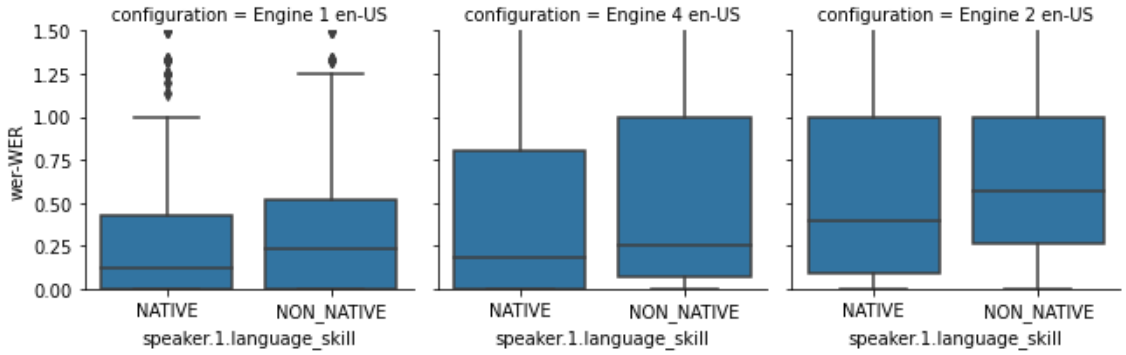

Figure 43 Subset of English data filtered for spontaneous speech and plotted language skill against WER

In Figure 44 language skills are broken down more granularly in mother tongue. A direct relationship between non-native and non-native and mother tongue cannot be ascertained. This is due to unclear allocation of mother tongue to native or non-native (explained in Figure 45).

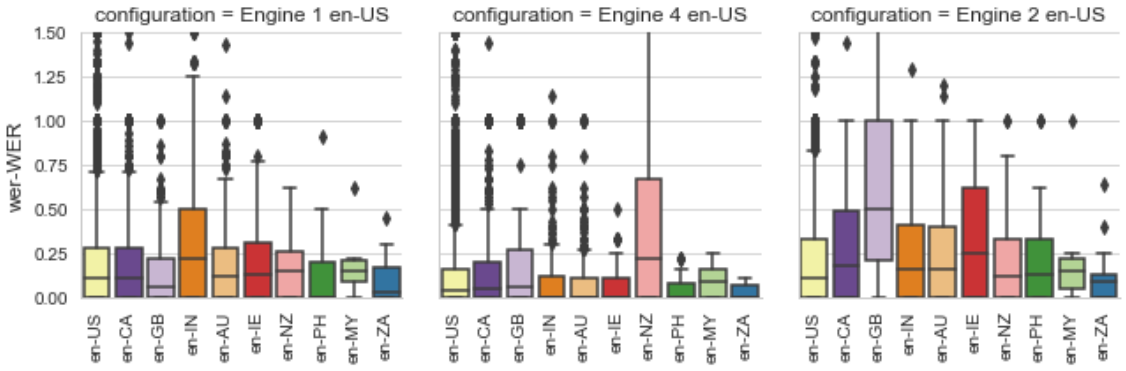

Figure 44 Boxplots of mother tongue of native and non-native speakers against WER per engine

Engine 1 performed balanced across various mother tongues with en-IN (“Indian”) performing the worst. Engine 4 performed balanced across the various mother tongues with en-NZ (“New Zealanders) performing the worst in terms of variance and median WER. Between the evaluated engines, Engine 4 performed the best.

To establish the relationship between native and non native and mother tongues. In Figure 45 all mother tongues assigned to non-native in the metadata are plotted against each other.

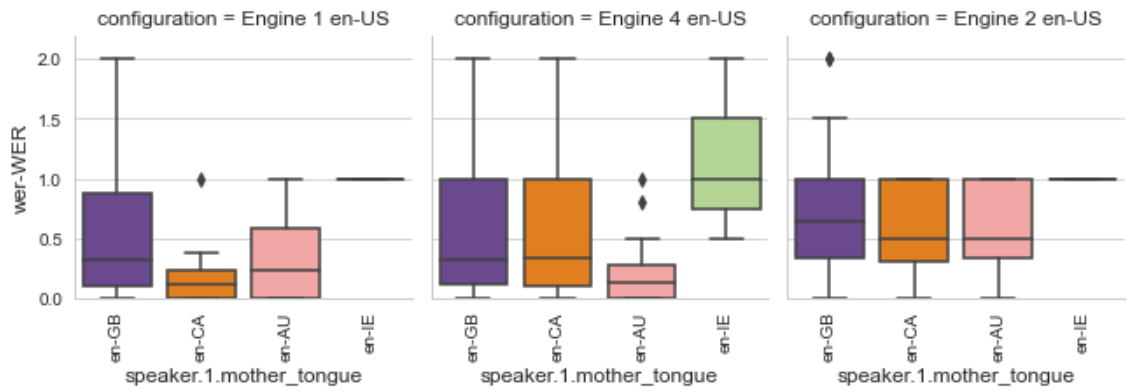

Figure 45 Boxplots of mother tongues assigned to non-native in the metadata against WER per engine

No fair evaluation can be performed on filtered samples for non-native and looking on the mother tongue assigned to these samples in the metadata (as shown above in Figure 45). As British, Canadian and Australian English mother tongues cannot be considered non-native language skilled.

4.2.4 Latency

In Figure 46 it can be observed that the engines performed very similar in terms of mean latency for each corpus. Engine 4 performed best across all corpora, followed by Engine 2 and Engine 1. Engine 2 performed significantly worse than Engine 1 and 4 for the data set from ami. All engines show high latency means for tedlium (speaking style semi-spontaneous) and librispeech (speaking style scripted). For corpora rt latency mean was low but WER was high (see Figure 39).

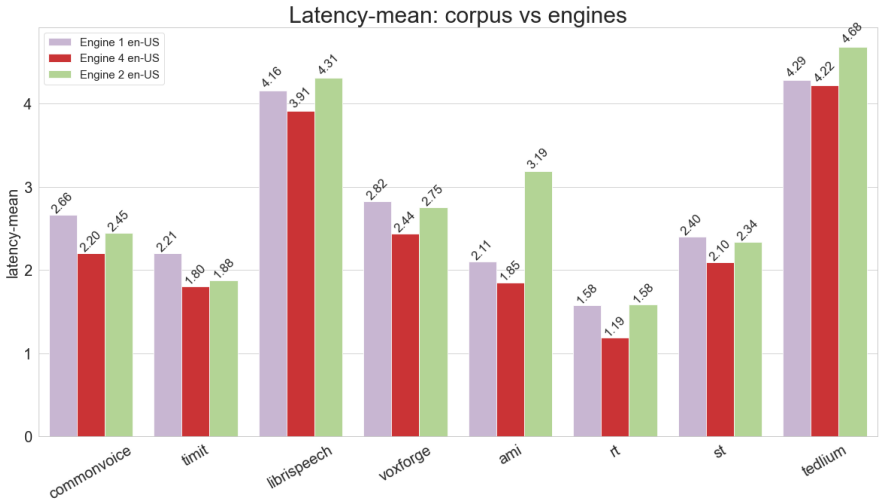

Figure 46 mean latency per engine per corpora

In Figure 47 a right skewed distribution for all engines with a negative correlation between WER and latency can be observed. This means the lower the latency the higher the WER, this is an expected behavior. The difference in distribution of the datapoints is mainly related to plotting against the WER on the x-axis, as the engines performed similarly in terms of latency but performed more mixed in terms of WER.

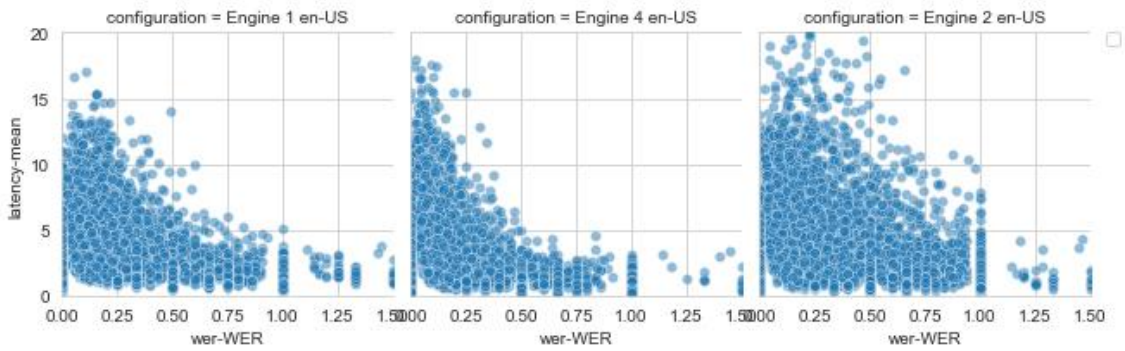

Figure 47 WER vs latency across all corpora per engine

4.2.5 Comparison of systems over time

Figure 48 shows it can be observed that the overall performance of the engines in terms of WER has declined between the evaluation in 2019 [1] and 2022 (this project thesis). On the contrary, an improvement was expected over the three-year period. As mentioned in chapter 4.2 the evaluation in 2019 was performed with a different evaluation framework and applying optimization to the parameter settings of the engines' configuration for at least some corpora. No fair comparison can therefore be performed between the two evaluations in terms of WER. As the parameter settings used in 2019 are unknown to us, also no comparison can be performed in terms of comparing the parameter settings and their impact on performance of overall WER and WER per corpus. Nonetheless, it was decided to show the WER for both evaluations in Figure 48 below to emphasize the influence of the engine's configuration on its performance in recognizing speech.

The distinction between mean per corpus and mean over all utterances introduced in chapter 4.1 and 4.2 is discontinued hereafter, as only mean per corpus is available for the previous evaluations [9] [13].

Engine 1 en-US	ami	commonvoice	librispeech	rt	st	tedlium	timit	voxforge	mean per corpus
WER 2022	0.40	0.15	0.12	0.31	0.08	0.11	0.14	0.13	0.18
WER 2019	0.39	0.13	0.08	0.32	0.06	0.08	0.09	0.10	0.16
Engine 4 en-US									
WER 2022	0.46	0.06	0.06	0.41	0.06	0.07	0.07	0.07	0.16
WER 2019	0.38	0.09	0.09	0.25	0.04	0.10	0.08	0.09	0.15
Engine 2 en-US									
WER 2022	0.60	0.21	0.12	0.52	0.05	0.17	0.10	0.14	0.24
WER 2019	0.40	0.10	0.07	0.30	0.04	0.06	0.64	0.08	0.15
mean per corpora across all engines									
Ø WER 2022	0.49	0.14	0.10	0.41	0.06	0.12	0.10	0.11	0.19
Ø WER 2019	0.39	0.11	0.08	0.29	0.05	0.08	0.27	0.09	0.17

Figure 48 Table comparing WER of 2019 [9] with 2022 across all engines and corpora

The dissimilarities in evaluation framework and parameter settings outlined above for the comparison between 2019 and 2022 don't apply to the evaluation results in Figure 49, as these were performed on identical data sets, identical evaluation framework (i.e., BeSTT) and identical postprocessing. For Engine 2 and 4 the results do not differ, as was expected. Less is true for Engine 1 the performance in terms of mean WER per corpus increased by 50% with corpora contributing equally to the increase except for st for which the WER doubled.

Engine 1 en-US	commonvoice	librispeech	st	tedlium	timit	voxforge	mean per corpus
WER 2022	0.15	0.12	0.08	0.11	0.14	0.13	0.12
WER 2022(1)	0.11	0.09	0.04	0.07	0.09	0.10	0.08
Engine 4 en-US							
WER 2022	0.06	0.06	0.06	0.07	0.07	0.07	0.07
WER 2022(1)	0.07	0.07	0.05	0.07	0.07	0.08	0.07
Engine 2 en-US							
WER 2022	0.21	0.12	0.05	0.17	0.10	0.14	0.13
WER 2022(1)	0.21	0.13	0.05	0.17	0.10	0.14	0.13
mean per corpora across all engines							
WER 2022	0.14	0.10	0.06	0.12	0.10	0.11	0.11
WER 2022(1)	0.13	0.10	0.05	0.10	0.09	0.11	0.10

Figure 49 Table comparing WER of 2022(1) [13] with 2022 across all engines and corpora

The poor performance of Engine 1 and the identical performance of Engine 2 and 4 compared to the previous evaluation in mid-2022 (WER 2022(1)) [13] suggests that either Engine 1 was misconfigured on BeSTT, or during the uploading or creating the benchmarks a mistake was made. Thus, Engine 1 cannot be fairly compared to the previous evaluation. The difference in mean WER compared to Figure 39 is explained with the exclusion of ami and rt (corpora containing spontaneous speech) in the evaluation in mid-2022, which both have high WER.

5 Conclusion

French and Italian analysis

In conclusion, it can be said that there are various factors which influence the performance of ASR engines. The analysis of French and Italian showed that the sample duration and transcription length has a significant impact on the WER. Common Voice returned the poorest results for its corpus in French and MLS in Italian. The probability for a full sample to be transcribed wrong is higher if the sample is short, as can be seen for Common Voice. As a corpus provider, cutting the audios into too short segments can therefore backfire and influence the WER negatively. It is only to be assumed how the performance with long samples could have looked like, but in terms of practicability most corpus providers segmented their samples. Fleur had a wider spread in connection to audio duration and transcription length obtained an overall good performance.

When the performance for the speakers of the MLS corpus were plotted and coloured by gender one could detect that for Italian females obtained the poorest and males the best results. Except for that observation for French and Italian there was no significant gender bias detected concerning the WER, however this might be traced back by the unevenness of samples per corpus and gender as it is known that the performance of ASR engines is poorer for females [15].

Concerning the influence of age on the performance of the ASR engines for Italian no such relation could be detected. By plotting the ratio of the samples which obtained a WER equal to 0 divided by the total samples per age a shape could be detected where speakers in their thirties, forties and fifties obtained a better performance. In terms of RTF derived from the mean latency Engine 2 and 4 performed similar and on the acceptable upper threshold of 0.6 for state-of-the-art engines. Engine 3 has a lower RTF score by a factor 4 and shows the lowest correlation between latency and audio duration.

In terms of performance of the engines themselves there were clear differences to be detected. Engine 4 showed the overall best performance. The poorest performance was obtained by Engine 1 however, it was followed closely by Engine 2. This is true for French and Italian, while the engines transcribed Italian with higher accuracy than French. It is to be assumed that during of the implementation of the engines into the BeSTT-Platform the engines were not configurated properly and the capability of engine 1 would have been greater.

English analysis

The comparison between 2019 and 2022 is not meaningful in terms of assessing changes in performance of ASR, as the engines are configurated differently. Further development and improvement in the fields of ASR and declined performance emphasizes the importance of configuring the engines correctly and adoption for their use case. The comparison of the evaluations both performed in 2022 has revealed the changed configuration of Engine 1, which lead to an increase of the overall WER by 50 %, while the WER declined for spontaneous speech, again emphasizing the importance of the engines' configuration and its effect on accuracy.

All engines are sensitive to spontaneous speech resulting in an increased WER compared to scripted speech. For Engine 2 the mean WER increased by a factor 6.7x and the mean WER for Engine 4 increased by 4.1x. In terms of overall WER (over all utterances and corpora) Engine 4 has a 50 % higher WER with 0.30 WER than Engine 2 with 0.20. In interpreting these accuracy measures, one needs to be wary of the fact that the configuration may differ and its effects on accuracy.

The assignment of language skill in non-native and native English speaker and corresponding metadata mother tongue for the same sample could not be comprehended. As for example British mother tongue was assigned to non-native English speaker. Assuming the assignment of language skill is

correct, non-native was only assigned to samples containing spontaneous speech, which has a high WER of around 0.5 for Engine 2 and 4, making it difficult to draw conclusions about sensitivity to language skills and its effects on accuracy.

In terms of RTF derived from the mean latency all engines performed similar and on the acceptable upper threshold of 0.6 for state-of-the-art engines.

6 Discussion and Outlook

The evaluation of ASR engines is a complex undertaking, which we did not know how to assess at the beginning and therefore underestimated it. The configuration of the ASR engines and the implementation of those into an existing framework was not only difficult due to the lack of knowledge, but also because the engines themselves are very diverse and use complex frameworks.

The same is true for the corpora, which are very diverse. Each corpus has its own structure, with different metadata and in different annotations and formats. Likewise, the documentation is very different from precise to very sparse or the documentation has to be gathered from different sources. Furthermore, the availability of the data in good quality and with validated ground truth text was especially for French and Italian more challenging than expected.

The achieved results do not meet the expectations at the beginning of the project, as we evaluated fewer engines and fewer corpora than planned. In addition, some difficulties arose during the evaluation with the engines and corpora. Different encoding/decoding of the data led to distorted WER at the beginning and interruptions and aborts during the execution of the benchmarks on BeSTT delayed the availability of the transcripts results. Given the circumstances and the short timeframe we have had to do the evaluation, we are satisfied with the results and believe that they are relevant apart from Engine 1, which is obviously misconfigured. If systemic errors were made during the process, they equally affected the results of the engines and still allow comparability. The accuracy of the results cannot be judged conclusively, but we think that with the given setup they are to be expected and are in-line with previous evaluations with comparable setup.

It is interesting that these problems probably provided the initial motivation to design CEASR and BeSTT, a unified corpus composed of sub-corpora in a unified format and easily accessible and an "easy" to use evaluation tool.

However, what surprised us very positively were the provider of the ASR engines. They were very interested in this project and responsive to it. We were able to establish contacts with a large number of well-known providers and had interesting discussions. We were able to realize our goals to get their agreement to be mentioned by name and a possible discount for the fees that accrue for transcription of large amounts of data

In retrospect, we have to admit that initial outline of this project was unrealistic from the beginning and that this led to, our resources and those of third parties not being used in the best possible way and looking back substantially misused. We should have communicated more proactively that the initial outline is not feasible for us and discussing an alternative that was more in line with our capabilities such as providing us with data from previous runs to do more in-depth analysis on those. Furthermore, we believe that further analysis of language skill, accents, dialects and mother tongue on a broader set of data containing various speaking style could yield more relevant results to assess the sensitivity of ASR engines on these attributes. A more in-depth analysis on the French and Italian data could provide insight to why French was predicted less accurately than Italian.

7 References

7.1 Bibliography

[1] Malgorzata Anna Ulasik et al., "CEASR: A Corpus for Evaluating Automatic Speech Recognition," *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020),* p. 64776485, 2020.

[2] S. K. a. J. Bahadur, "https://developer.nvidia.com/," Nvidia, 31 08 2022. [Online]. Available: https://developer.nvidia.com/blog/solving-automatic-speech-recognition-deployment-challenges/. [Accessed 09 12 2022].

[3] S. Rella, "developer.nvidia.com," Nvidia, 08 08 2022. [Online]. Available: https://developer.nvidia.com/blog/essential-guide-to-automatic-speech-recognition-technology/. [Accessed 09 12 2022].

[4] Georgescu et al., "Progress on automatic annotation of speech corpora using complementary ASR systems," *2019 42nd International Conference on Telecommunications and Signal Processing (TSP),* Vols. 571-574, 2019.

[5] Wellekens et al., "Automatic speech recognition and speech variability: A review," *Speech Communication,* vol. 49, no. 10, pp. 763-786, 2007.

[6] M. Chen, "www.notta.ai," notta ai, 29 06 2022. [Online]. Available: https://www.notta.ai/en/blog/automatic-speech-recognition. [Accessed 09 12 2022].

[7] H. a. O. S. Polat, "Building a Speech and Text Corpus of Turkish: Large Corpus Collection with Initial Speech Recognition Results," *Symmetry,* vol. 12, no. 2, p. 290, 2022.

[8] D. K. a. J. Peters, "Testing the correlation of word error rate and perplexity," *Speech Communication,* vol. 38, no. 1, pp. 19-28, 2002.

[9] F. G. a. M. A. Ulasik, "Evaluation of Automatic Speech," Bachelor thesis, ZHAW, 2019.

[10] Singh et al., "2019 IEEE Fifth International Conference on Multimedia Big Data (BigMM)," *Automatic Speech Recognition for Real Time Systems,* pp. 189-198, 2019.

[11] Mouchtaris et al., "Caching networks: Capitalizing on common speech for ASR," in *ICASSP 2022*, Singapore, 2022.

[12] O.-T. Wiki, 12 12 2021. [Online]. Available: https://openvoice-tech.net/index.php/Real-time-factor. [Accessed 20 12 2022].

[13] E. Altman, "Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy," *Journal of Finance,* 1968.

[14] Adamantidou et al., "Bestt – a framework for evaluation of STT benchmarks," in *SwissText*, 2022.

[15] A. V. V. D. Thayabaran Kathiresan, "Gender bias in voice recognition: An i-vector-based gender-specific," Zurich University, Zurich, 2021.

[16] K. Moskvitch, “www.bbc.com,” 15 02 2017. [Online]. Available: https://www.bbc.com/future/article/20170214-the-machines-that-learned-to-listen.

[17] Statista Research Department, "www.statista.com," 05 08 2022. [Online]. Available: https://www.statista.com/statistics/266808/the-most-spoken-languages-worldwide/. [Accessed 11 11 2022].

[18] S. Bondolfi, "www.swissinfo.ch," 23 05 2017. [Online]. Available: https://www.swissinfo.ch/ger/politik/raetoromanisch_stirbt-sprache-der-bergler-aus/43203472. [Accessed 11 11 2022].

[19] L. Tate, "www.kardome.com," www.kardome.com, [Online]. Available: https://www.kardome.com/blog-posts/speech-recognition-technology-workplace#:~:text=The%20primary%20benefit%20of%20speech,any%20information%20into%20a%20machine.. [Accessed 07 12 2022].

7.2 Glossary

This glossary is mostly extracted from the Bachelor Thesis “Evaluation of Automatic Speech Recognition Systems” written by Fabian Germann and Malgorzata Anna Ulasik [1].

Word Error Rate (WER)	the edit distance between a reference and its automatic transcription, normalised by the length of the reference.
Automatic Speech Recognition (ASR)	a technology enabling machines to process speech input and translate it into text. It is also known as Speech Recognition and Speech-to-text (STT).
Benchmark Evaluation Speech to Text platform (BeSTT)	a proprietary framework for the evaluation of STT engines
Speech Corpus	a collection of digital recordings of speech together with their annotations, meta data, and documentation.
Evaluation Metric	a measure used for evaluation of an ASR system. It should be objective and clearly interpretable. Most well-known ASR evaluation metric is Word Error Rate.
Utterance	a unit of speech which is provided as input for speech recognition. It can be a single word, a phrase, a complete or incomplete sentence, or multiple sentences.
CEASR	a Corpus for Evaluation of Automatic Speech Recognition
Hypothesis	refers to the prediction of an ASR engine on the reference of a given audio sample and is used interchangeably with transcription in this project thesis.
Reference	refers to the true transcription of a given audio samples. Reference and ground truth are interchangeably used as the bear the same meaning.

7.3 List of Figures

Figure 1 Detailed overview of Common Voice 11.0 (CV) Corpus (French and Italian) 8
Figure 2 Detailed Overview of Multilingual Librispeech Corpus (French and Italian) 9
Figure 3 Detailed Overview of Fleur Corpus (French and Italian) 10
Figure 4 Visualization of the evaluation workflow on BeSTT [13] 12
Figure 5 Sketch of the BeSTT system architecture (Source: Alexandros Paramythis) 13
Figure 6 Overview of examined but excluded corpora for Italian and French 14
Figure 7 Default test set, subset and evaluation size for selected corpora in French and Italian 14
Figure 8 Overview of selected corpora across all languages after subsetting Common Voice 11.0 15
Figure 9 Visualization of the schemata of the unified format of CESAR 16
Figure 10 Distribution of French data for audio duration against length of ground truth text from Common Voice 11 pre and post cleaning applying z-score 17
Figure 11 Distribution of audio duration and length of ground truth 18
Figure 12 Table comparing WER for French and Italian of all corpora and engines 20
Figure 13 Samples where WER=0 per Corpus (all Engines) for French and Italian 21
Figure 14 Number of Samples per Corpus for French and Italian which were transcribed 21
Figure 15 WER Mean per Corpus and Engine for French and Italian 22
Figure 16 WER per Engine (all Corpora) for French and Italian 22
Figure 17 WER per Engine and Corpus for French and Italian 23
Figure 18 Transcription Duration vs. WER for French and Italian 23
Figure 19 Transcription Length vs. WER for French and Italian 24
Figure 20 Audio Duration vs. WER separately for each corpus and Number of Samples per Audio Duration separately for each corpus in French 24
Figure 21 Transcription Length vs. WER separately for each corpus and Number of Samples per Audio Duration separately for each corpus in French 25
Figure 22 Audio Duration vs. WER separately for each corpus and Number of Samples per Audio Duration separately for each corpus in Italian 25
Figure 23 Transcription Length vs. WER separately for each corpus and Number of Samples per Audio Duration separately for each corpus in Italian 26
Figure 24 Grouped Transcription Length vs. WER (alle Engines) for French and Italian 27
Figure 25 WER per Speaker [MLS] separately for each engine colored by gender French 28
Figure 26 WER per Speaker [MLS] separately for each engine colored by gender Italian 28
Figure 27 WER per Gender for French and Italian 29
Figure 28 WER per Engine and Gender for French and Italian 29
Figure 29 Number of Samples per Gender and Corpus for French and Italian 30
Figure 30 WER per Gender and Corpus for French and Italian 30
Figure 31 WER per Age [Common Voice] (all Engines) for French and Italian 31
Figure 32 Number of Samples per Age [Common Voice] (all Engines) for French and Italian, Number of Samples per Age [Common Voice] where WER=0 (all Engines) for French and Italian and Ratio of Samples with WER=0 Against total Samples per Age (all Engines) [Common Voice] for French and Italian 32
Figure 33 RTF grouped by engine for French data 33
Figure 34 RTF grouped by engine for Italian data 33
Figure 35 Barplot of engine's latency per French corpus 33
Figure 36 Barplot of engine's latency per Italian corpus 33
Figure 37 Scatterplot latency vs. audio duration on French data 34
Figure 38 Scatterplot latency vs. audio duration on Italian data 34
Figure 39 Overview of WER per engine per corpora 35
Figure 40 Boxplots of speaking style against WER per engine 36
Figure 41 Boxplots of language skill of speaker against WER per engine across all corpora 36
Figure 42 Overview of number of occurrences of language skill in attribute speaking style 37

Figure 43 Subset of English data filtered for spontaneous speech and plotted language skill against WER _____ *37*
Figure 44 Boxplots of mother tongue of native and non-native speakers against WER per engine _____ *37*
Figure 45 Boxplots of mother tongues assigned to non-native in the metadata against WER per engine *38*
Figure 46 mean latency per engine per corpora _____ *38*
Figure 47 WER vs latency across all corpora per engine _____ *39*
Figure 48 Table comparing WER of 2019 [9] with 2022 across all engines and corpora _____ *39*
Figure 49 Table comparing WER of 2022(1) [13] with 2022 across all engines and corpora _____ *40*
Figure 50 Python code belonging to chapters 3.5.2, 3.5.3, 3.5.4 and 3.5.6 _____ *58*
Figure 51 Python code belonging to chapter 3.5.5 _____ *58*
Figure 52 Python code belongs to chapter 3.5.8 _____ *58*

Appendix

A. Corpora Documentation

The following documentation concerning the English corpora is extracted from the Bachelor Thesis "Evaluation of Automatic Speech Recognition Systems" written by Fabian Germann and Malgorzata Anna Ulasik [1].

A.1 AMI

	GENERAL INFORMATION
Summary	The AMI Meeting Corpus is a multi-modal data set consisting of 100 hours of meeting recordings. Around two-thirds of the data has been elicited using a scenario in which the participants play different roles in a design team, taking a design project from kick-off to completion over the course of a day. The rest consists of naturally occurring meetings in a range of domains. Although the AMI Meeting Corpus was created for the uses of a consortium that is developing meeting browsing technology, it is designed to be useful for a wide range of research areas.
URL	http://groups.inf.ed.ac.uk/ami/corpus/
Owner / Authors	University of Edinburgh
License	Creative Commons Attribution 4.0 International Public License http://creativecommons.org/licenses/by/4.0/legalcode
	COPRUS PROPERTIES
Speaking Style	Dialog spontaneous speech
Accented Speech	Yes
Dialectal Variation	Yes
Overlapping Speech	Yes
Filled Pauses	Yes
Speaker Noise	Yes
Acoustic Environment	Meeting room
Recording Device	RealMedia audio mix, Headset mix, Lapel mix, Individual lapels, Individual headsets, Microphone array
	TRANSCRIPTION INPUTS
Reference	Segmented on word level (.xml file per speaker per meeting)
Audio	Unsegmented (one .wav file per meeting)
Applied Segmentation (ref and audio)	Segments on speaker utterance level
	TESTSET
Test Set Definition	Random selection
Test Set Duration	5 hours
Number Utterances	4563
Number Speakers	38
Average Utterance Duration	4 seconds
Average Speaking Rate	141 wpm

A.2 Common Voice (English)

	GENERAL INFORMATION
Summary	An open source, multi-language dataset of voices that anyone can use to train speech-enabled applications.
URL	https://voice.mozilla.org/en/datasets
Owner / Authors	Mozilla
License	CC BY-SA 3.0 Zusammenfassung: https://creativecommons.org/licenses/by-sa/3.0/deed.locale Details: https://creativecommons.org/licenses/by-sa/3.0/legalcode
	COPRUS PROPERTIES
Speaking Style	Monologue read-aloud speech
Accented Speech	Unknown
Dialectal Variation	Yes
Overlapping Speech	Yes
Filled Pauses	Yes
Speaker Noise	No
Acoustic Environment	Unknown
Recording Device	
	TRANSCRIPTION INPUTS
Reference	Segmented on speaker utterance level (one csv with all utterances)
Audio	Segmented on speaker utterance level (mp3 per utterance)
Applied Segmentation (ref and audio)	Segments on speaker utterance level
	TESTSET
Test Set Definition	Default test set
Test Set Duration	5 hours
Number Utterances	3995
Number Speakers	Unknown
Average Utterance Duration	4.5 seconds
Average Speaking Rate	132 wpm

A.3 LibriSpeech

GENERAL INFORMATION	
Summary	A corpus of read English speech, suitable for training and evaluating speech recognition systems. The LibriSpeech corpus is derived from audiobooks that are part of the LibriVox project, and contains 1000 hours of speech. We have made the corpus freely available for download, along with separately prepared language-model training data and pre-built language models.
URL	http://www.openslr.org/12/
Owner / Authors	Vassil Panayotov, Daniel Povey
License	CC BY 4.0 https://creativecommons.org/licenses/by/4.0/legalcode
COPRUS PROPERTIES	
Speaking Style	Monologue read-aloud speech
Accented Speech	Unknown
Dialectal Variation	No
Overlapping Speech	No
Filled Pauses	No
Speaker Noise	No
Acoustic Environment	Unknown
Recording Device	Unknown
TRANSCRIPTION INPUTS	
Reference	Segmented on speaker utterance level (one txt with all utterances)
Audio	Segmented on speaker utterance level (flac file per utterance)
Applied Segmentation (ref and audio)	Segments on speaker utterance level
TESTSET	
Test Set Definition	Default test set
Test Set Duration	5.4 (clean), 5.3 (other)
Number Utterances	2620(clean), 2939 (other)
Number Speakers	40 (clean), 33 (other)
Average Utterance Duration	7.52 seconds (clean), 6.54 seconds (other)
Average Speaking Rate	163 wpm (clean), 161 (other)

A.4 RT

GENERAL INFORMATION	
Summary	The evaluation data consists of an approximately 180-minute multi-site test set containing 7 meeting excerpts from 7 meetings. The test data was collected at EDI, IDI, and NIST. Each meeting excerpt contains a head-mic recording for each subject and one or more distant microphone recordings (whatever the data collection sites provided to NIST).
URL	https://www.nist.gov/itl/iad/mig/rich-transcription-evaluation
Owner / Authors	Prepared by the National Institute of Standards and Technology (NIST) Multimodal Information Group and distributed by the Linguistic Data Consortium (LDC)
License	Linguistic Data Consortium (LDC) license
COPRUS PROPERTIES	
Speaking Style	Dialog spontaneous speech
Accented Speech	Yes
Dialectal Variation	Unknown
Overlapping Speech	Yes
Filled Pauses	Yes
Speaker Noise	Yes
Acoustic Environment	Meeting room
Recording Device	individual lapels, individual headsets, headset sum, microphone array, distant microphone, summed distant microphone, source localization arrays, KEMAR mannequin
TRANSCRIPTION INPUTS	
Reference	Segmented on speaker utterance level (one .tdf file with all speaker utterances per meeting)
Audio	Segmented on meeting level (one .sph file per meeting)
Applied Segmentation (ref and audio)	Segments on speaker utterance level
TESTSET	
Test Set Definition	Random selection
Test Set Duration	3.6 hours
Number Utterances	6334
Number Speakers	30
Average Utterance Duration	2 seconds
Average Speaking Rate	195 wpm

A.5 Switchboard

GENERAL INFORMATION	
Summary	The Switchboard-1 Telephone Speech Corpus (LDC97S62) consists of approximately 260 hours of speech and was originally collected by Texas Instruments in 1990-1, under DARPA sponsorship. The first release of the corpus was published by NIST and distributed by the LDC in 1992-3. Since that release, a number of corrections have been made to the data files as presented on the original CD-ROM set and all copies of the first pressing have been distributed. Switchboard is a collection of about 2,400 two-sided telephone conversations among 543 speakers (302 male, 241 female) from all areas of the United States. A computer-driven robot operator system handled the calls, giving the caller appropriate recorded prompts, selecting and dialing another person (the callee) to take part in a conversation, introducing a topic for discussion and recording the speech from the two subjects into separate channels until the conversation was finished. About 70 topics were provided, of which about 50 were used frequently. Selection of topics and callees was constrained so that: (1) no two speakers would converse together more than once and (2) no one spoke more than once on a given topic.
URL	https://catalog.ldc.upenn.edu/LDC97S62
Owner / Authors	Collected by Texas Instruments in 1990-1, under DARPA sponsorship Authors: John J. Godfrey, Edward Holliman
License	Linguistic Data Consortium (LDC) license
COPRUS PROPERTIES	
Speaking Style	Dialog spontaneous speech
Accented Speech	Yes
Dialectal Variation	Unknown
Overlapping Speech	Yes
Filled Pauses	Yes
Speaker Noise	Yes
Acoustic Environment	Unknown
Recording Device	Phone
TRANSCRIPTION INPUTS	
Reference	Segemented on speaker utterance level (one .txt file with all utterances) Transcripts created by Mississippi State transcripts
Audio	Segemented on call level (one audio file per call)
Applied Segmentation (ref and audio)	Segements on speaker utterance level
TESTSET	
Test Set Definition	Random selection
Test Set Duration	5 hours
Number Utterances	4105
Number Speakers	483
Average Utterance Duration	4.39 seconds
Average Speaking Rate	128 wpm

A.6 ST

GENERAL INFORMATION	
Summary	A free American English corpus by Surfingtech (www.surfing.ai), containing utterances from 10 speakers, Each speaker has about 350 utterances. The data set is a subset of a much bigger data set (about 1000hours) which was recorded in the same environment as this open source data. This corpus were recorded in silence in-door environment using cellphone. It has 10 speakers. Each speaker has about 350 utterances. All utterances were carefully transcribed and checked by human. Transcription accuracy is guaranteed.
URL	http://www.openslr.org/45/
Owner / Authors	Surfingtech (www.surfing.ai)
License	Creative Common BY-NC-ND 4.0 https://creativecommons.org/licenses/by-nc-nd/4.0/deed.de
COPRUS PROPERTIES	
Speaking Style	Monologue read-aloud speech
Accented Speech	Unknown
Dialectal Variation	No
Overlapping Speech	No
Filled Pauses	No
Speaker Noise	No
Acoustic Environment	Silence in-door environment
Recording Device	Cellphone
TRANSCRIPTION INPUTS	
Reference	Segmented on speaker utterance level (one .txt with all utterances)
Audio	Segmented on speaker utterance level (one .wav file per utterance)
Applied Segmentation (ref and audio)	Segments on speaker utterance level
TESTSET	
Test Set Definition	Random selection
Test Set Duration	4.7 hours
Number Utterances	3842
Number Speakers	5
Average Utterance Duration	4.44 seconds
Average Speaking Rate	109 wpm

A.7 TedLium

GENERAL INFORMATION	
Summary	The TED-LIUM corpus is English-language TED talks, with transcriptions, sampled at 16kHz. It contains about 452 hours of speech.
URL	https://www.openslr.org/51/
Owner / Authors	Created through a collaboration between the Ubiqus company and the LIUM (University of Le Mans, France)
License	Creative Commons BY-NC-ND 3.0
COPRUS PROPERTIES	
Speaking Style	Monologue semi-spontaneous speech
Accented Speech	Unknown
Dialectal Variation	No
Overlapping Speech	No
Filled Pauses	Yes
Speaker Noise	No
Acoustic Environment	Unknown (most probably audience hall)
Recording Device	Unknown
TRANSCRIPTION INPUTS	
Reference	Segmented on utterance level (one .stm file split in short utterances with time stamps)
Audio	Segemented on lecture level (one .sph file with the whole lecture recording)
Applied Segmentation (ref and audio)	Segments on speaker utterance level (segmented) Segments on lecture utterance level (unsegmented)
TESTSET	
Test Set Definition	Default test set
Test Set Duration	2.6 hours (segmented),
Number Utterances	1155 (segmented)
Number Speakers	11
Average Utterance Duration	8.15 seconds (segmented)
Average Speaking Rate	172 wpm

A.8 Timit

GENERAL INFORMATION	
Summary	The TIMIT corpus of read speech is designed to provide speech data for acoustic-phonetic studies and for the development and evaluation of automatic speech recognition systems. TIMIT contains broadband recordings of 630 speakers of eight major dialects of American English, each reading ten phonetically rich sentences. The TIMIT corpus includes time-aligned orthographic, phonetic and word transcriptions as well as a 16-bit, 16kHz speech waveform file for each utterance. Corpus design was a joint effort among the Massachusetts Institute of Technology (MIT), SRI International (SRI) and Texas Instruments, Inc. (TI). The speech was recorded at TI, transcribed at MIT and verified and prepared for CD-ROM production by the National Institute of Standards and Technology (NIST). The TIMIT corpus transcriptions have been hand verified. Test and training subsets, balanced for phonetic and dialectal coverage, are specified. Tabular computer-searchable information is included as well as written documentation.
URL	https://catalog.ldc.upenn.edu/LDC93S1
Owner / Authors	Authors: John S. Garofolo, Lori F. Lamel, William M. Fisher, Jonathan G. Fiscus, David S. Pallett, Nancy L. Dahlgren, Victor Zue
License	Linguistic Data Consortium (LDC) license
COPRUS PROPERTIES	
Speaking Style	Monologue read-aloud speech
Accented Speech	No
Dialectal Variation	No
Overlapping Speech	No
Filled Pauses	Yes
Speaker Noise	No
Acoustic Environment	Unknown
Recording Device	Unknown
TRANSCRIPTION INPUTS	
Reference	Segmented on speaker utterance level (one .txt file per speaker utterance)
Audio	Segmented on speaker utterance level (one .wav file per speaker utterance)
Applied Segmentation (ref and audio)	Segments on speaker utterance level
TESTSET	
Test Set Definition	Default test set
Test Set Duration	1.4 hours
Number Utterances	1680
Number Speakers	168
Average Utterance Duration	3.09 seconds
Average Speaking Rate	172 wpm

A.9 VoxForge

GENERAL INFORMATION	
Summary	The volunteer-supported speech-gathering effort Voxforge3, on which the acoustic models we used for alignment were trained, contains a certain amount of LibriVox audio, but the dataset is much smaller than the one we present here, with around 100 hours of English speech, and suffers from major gender and per- speaker duration imbalances.
URL	http://www.voxforge.org
Owner / Authors	VoxForge
License	GNU General Public License: http://www.gnu.org/copyleft/gpl.html
COPRUS PROPERTIES	
Speaking Style	Monologue read-aloud speech
Accented Speech	Unknown
Dialectal Variation	Yes
Overlapping Speech	No
Filled Pauses	No
Speaker Noise	No
Acoustic Environment	Unknown
Recording Device	Unknown
TRANSCRIPTION INPUTS	
Reference	Segmented on speaker utterance level (one .txt with all utterances)
Audio	Segmented on speaker utterance level (.flac/.wav file per utterance)
Applied Segmentation (ref and audio)	Segments on speaker utterance level
TESTSET	
Test Set Definition	Default test set
Test Set Duration	3.9 hours
Number Utterances	2929
Number Speakers	171
Average Utterance Duration	4.78 seconds
Average Speaking Rate	171 wpm

B. Overview Benchmark Execution

id	engine	local	fleur	cv 11	mls	TIMIT	VoxForge	cv_en	librispeech (clean)	switchboard	ted-lium	ami	st	rt	benchmark	total sample	total time [hr]	BeSTT id
1000	**Engine_1**	fr	674 ; 1.95h	5169; 8.45	2426 ; 10.07h										1000_Engine_1_fr_fleur_cv_mls	8269	20.46	3131118
1001	**Engine_4**	fr	674 ; 1.95h	5169; 8.45	2426 ; 10.07h										1001_Engine_4_fr_fleur_cv_mls	8269	20.46	3131119
1002	**Engine_2**	fr	674 ; 1.95h	5169; 8.45	2426 ; 10.07h										1002_Engine_2_fr_fleur_cv_mls	8269	20.46	3131120
1003	**Engine_3**	fr	674 ; 1.95h	5169; 8.45	2426 ; 10.07h										1003_Engine_3_fr_fleur_cv_mls	8269	20.46	3131121
2001	**Engine_1**	it	864 ; 3.52h	4889 ; 8.32h	1262 ; 5.27h										2001_Engine_1_it_fleur_cv_mls	7015	17.12	3629152
2002	**Engine_4**	it	864 ; 3.52h	4889 ; 8.32h	1262 ; 5.27h										2002_Engine_4_it_fleur_cv_mls	7015	17.12	3629153
2003	**Engine_2**	it	864 ; 3.52h	4889 ; 8.32h	1262 ; 5.27h										2003_Engine_2_it_fleur_cv_mls	7015	17.12	3629154
2004	**Engine_3**	it	864 ; 3.52h	4889 ; 8.32h	1262 ; 5.27h										2004_Engine_3_it_fleur_cv_mls	7015	17.12	3629155
3001	**Engine_1**	en				1680 ;1.4h	2929 ; 3.9h	3995 ; 5h	2620 ; 5.4h	4105 ; 5h					3001_Engine_1_en_TIMIT_VoxForge_cv_librispeech (clean)_switchboard	15329	20.7	3629156
3002	**Engine_4**	en				1680 ;1.4h	2929 ; 3.9h	3995 ; 5h	2620 ; 5.4h	4105 ; 5h					3002_Engine_4_en_TIMIT_VoxForge_cv_librispeech (clean)_switchboard	15329	20.7	3629157
3003	**Engine_2**	en				1680 ;1.4h	2929 ; 3.9h	3995 ; 5h	2620 ; 5.4h	4105 ; 5h					3003_Engine_2_en_TIMIT_VoxForge_cv_librispeech (clean)_switchboard	15329	20.7	3629158
3004	**Engine_3**	en				1680 ;1.4h	2929 ; 3.9h	3995 ; 5h	2620 ; 5.4h	4105 ; 5h					3004_Engine_3_en_TIMIT_VoxForge_cv_librispeech (clean)_switchboard	15329	20.7	3629159
3101	**Engine_1**	en									1155 ; 2.6h	4539 ; 5h	2422 ; 4.7h	5324 ; 3.6h	3101_Engine_1_en__ted-lium_ami_st_rt	13440	15.9	3629160
3102	**Engine_4**	en									1155 ; 2.6h	4539 ; 5h	2422 ; 4.7h	5324 ; 3.6h	3102_Engine_4_en__ted-lium_ami_st_rt	13440	15.9	3629161
3103	**Engine_2**	en									1155 ; 2.6h	4539 ; 5h	2422 ; 4.7h	5324 ; 3.6h	3103_Engine_2_en__ted-lium_ami_st_rt	13440	15.9	3629162
3104	**Engine_3**	en									1155 ; 2.6h	4539 ; 5h	2422 ; 4.7h	5324 ; 3.6h	3104_Engine_3_en__ted-lium_ami_st_rt	13440	15.9	5729851

C. Python code used for post-processing and WER calculation

```python
def preprocess_text(text: str): #in our case this was done for postprocessing
    """ Performs text preprocessing, including:
    1. Lowers all letters
    2. Removes all excessive spaces
    3. Removes punctuation except apostrophes
    :param text: groundtruth or transcript text to preprocess
    :type text: str
    :rtype: str
    """

    #text = re.sub(r'\[[^A-Z]{2,}:\[[^\]]+\]\]', "", text) ##not accounting for french accents letters
    text = re.sub(r'\[[^À-ú]{2,}:\[[^\]]+\]\]', "", text) ##accounting for french accents letters
    text = text.lower()
    text = unidecode.unidecode(text)#replaces french and italian accents with ascii conform letters
    #text = unicodedata.normalize('NFKD', text)
    punctuation = '!"#()*,./:;<=>?@[\]^_{|}~'
    special_chars = "$%&+"
    text = text.translate(str.maketrans('', '', punctuation))
    text = text.translate(str.maketrans({x: f" {x} " for x in special_chars}))
    text = text.replace("-", " ")
    text = re.sub(r"\s+", " ", text)
    text = re.sub(r"'\s+", "'", text)
    text = text.strip()
    return text
```

Figure 50 Python code belonging to chapters 3.5.2, 3.5.3, 3.5.4 and 3.5.6

```python
def strNum2Words_fr(match):
    text = match.group()
    return num2words(int(text), lang= "fr")



def strNum2Words_it(match):
    text = match.group()
    return num2words(int(text), lang= "it")
```

Figure 51 Python code belonging to chapter 3.5.5

```python
transformation = jiwer.Compose([
    jiwer.ToLowerCase(),
    jiwer.RemovePunctuation(),
    jiwer.RemoveWhiteSpace(replace_by_space=True),
    jiwer.RemoveMultipleSpaces(),
    jiwer.ReduceToListOfListOfWords(word_delimiter=" ")
])
error_rate = []
for i in range(len(samples)):
    error = wer(groundtruth[i], transcripts[i], truth_transform=transformation,
    hypothesis_transform=transformation)
    error_rate.append(float(error))
```

Figure 52 Python code belongs to chapter 3.5.8

Project management

A. Description

Automatic Speech Processing (ASR) is used in various applications, such as chatbots, voice control car devices, subtitling etc. several providers exist that offer ready-built solutions, both commercial and open source.

Goal of this project is to create a corpus (data collection) with transcripts of different ASR engines on different input audio files in the languages used in Switzerland (e.g. German, French, Italian, English and Swiss German).

For this, existing public ASR corpora in these languages should be collected and enhanced with some proprietary data. These will then be run through various ASR engines, and the output will be compared with a ground truth transcript

B. Timetable

SW	Date	AIM
2	27.09.2022	Auswahl Anbieter, Erstellung Liste potenzieller Corpora, grober Zeitplan
3	04.10.2022	Anbieter anschreiben, introduction teaching session BESTT
4	11.10.2022	Start Integrierung erster Engine(s), Ablauf 1. Frist für Rückmeldung Anbieter
5	18.10.2022	Fortsetzung Integrierung Engines
6	25.10.2022	Fortsetzung Integrierung Engines, Ablauf 2. Frist für Rückmeldung Anbieter
7	01.11.2022	Fortsetzung Integrierung Engines, Start Niederschrift PA
8	08.11.2022	Fortsetzung Integrierung Engines,
9	15.11.2022	Abschluss von Durchlaufen lassen der Corpora für Start Datenanalyse
10	22.11.2022	Beginn Datenanalyse
11	29.11.2022	Datenanalyse fortsetzten, Graphen / Tabellen mit Kennzahlen erstellen
12	06.12.2022	Datenanalyse fortsetzen, abschliessen,
13	13.12.2022	Rohfassung PA fertig, in Korrektur geben
14	20.12.2022	Korrigieren, letzte Anpassungen & abgeben

C. Protocol

SW	Date	Status	To do
2	27.09.2022	Start of project work	General research of engine and corpus providers, read in into theoretical foundations & existing papers
3	04.10.2022	First list of engine and corpora providers	Do more research on engines and corpus providers (especially French and Italian), write engine providers
4	11.10.2022	Definite list of engine providers is done, Corpora are still not definite, discussion whether we use the same Corpora for English and German, first replies of engine providers were received, Correspondence with engine providers	write Alex für BeSTT introduction, do more research on corpus providers with weight on whether corpus was manually verified, research on how to implement an engine
5	18.10.2022	First Introduction with Alex was done, first sight into BeSTT, first research for implementation was done, Correspondence with engine providers	trying to implement an engine from online to batch (recommendation of Mark: Engine 2)
6	25.10.2022	Correspondence with engine providers, trying to implement an engine from online to batch (failing at this)	provide definite corpus list, trying to implement an engine from online to batch (Engine 2)
7	01.11.2022	Definite corpus list is done, trying to implement an engine (still failing), Rejection in engine implementation	Starting to transform the corpora, writing a few corpora providers
8	08.11.2022	Correspondence with corpus providers, start of transformation with help from Katja	Continuing the transformation of corpus, continue correspondence with corpus providers in order to receive more data
9	15.11.2022	Continuation of data transformation, struggling with given scripts, receiving help from Katja	Continuation of data transformation, start of project work documentation, correspondence with corpus providers, write Alex for follow up on how to upload the samples into BeSTT and run first test runs
10	22.11.2022	Finishing corpus transformation, start of project work documentation, Alex explained how to upload the samples into BeSTT and provided a script	Sending the transformed corpora to Katja for check, create benchmarks
11	29.11.2022	Mistakes from first transformation were reworked, definite corpus	Do first test runs, do big runs

		transformations are done, uploading all samples into BeSTT, test runs were done, benchmark created	
12	06.12.2022	Benchmarks were done, looking at first results, continue project work documentation	Continue with project work documentation, create plots with results
13	13.12.2022	Continuation of data analysis	Recalculate WER with inputs of Mark and Alex, finish project work documentation
14	20.12.2022	Project work substantially done, finishing touch ups, discussing results with Mark	Finish project work