FACULTY
OF MATHEMATICS
AND PHYSICS
Charles University

DOCTORAL THESIS

Dominik Macháček

Multi-Source Simultaneous Speech Translation

Institute of Formal and Applied Linguistics

Supervisor: doc. RNDr. Ondřej Bojar, Ph.D.

Study Program: Computational Linguistics

Prague 2024

I declare that I carried out this doctoral thesis independently, and only with the cited sources, literature and other professional sources.

I understand that my work relates to the rights and obligations under the Act No. 121/2000 Coll., the Copyright Act, as amended, in particular the fact that Charles University has the right to conclude a license agreement on the use of this work as a school work pursuant to Section 60 paragraph 1 of the Copyright Act.

Prague, 29. 2. 2024 Dominik Macháček

Title: Multi-Source Simultaneous Speech Translation

Author: Dominik Macháček

Department: Institute of Formal and Applied Linguistics

Supervisor: doc. RNDr. Ondřej Bojar, Ph.D.,
Institute of Formal and Applied Linguistics

Abstract:

Neural machine translation has the capability to translate from several parallel inputs into different languages. Current simultaneous speech translation sometimes faces issues with quality, especially when the source is noisy. We investigate the opportunity to use multiple parallel speech signals — the original and simultaneous interpreting — as sources for translation to achieve higher quality. We create an evaluation set ESIC (Europarl Simultaneous Interpreting Corpus). We analyze the challenges of simultaneous interpreting when used as an additional parallel source. Then, we investigate the robustness of multi-sourcing to transcription errors and assess the reliability of machine translation metrics when evaluating simultaneous speech translation. Last but not least, we implement Whisper-Streaming, a tool that enables real-time processing of large offline speech-to-text models and demonstrates the state of the art.

Keywords: simultaneous speech translation, speech translation, speech-to-text translation, machine translation, multilinguality, multi-sourcing, simultaneous interpreting, natural language processing

Název práce: Simultánní překlad řeči z více zdrojů

Autor: Dominik Macháček

Pracoviště: Ústav formální a aplikované lingvistiky

Vedoucí práce: doc. RNDr. Ondřej Bojar, Ph.D.,
Ústav formální a aplikované lingvistiky

Abstrakt:

Neuronový strojový překlad má schopnost překládat z několika paralelních vstupů v různých jazycích. Kvalita současného automatického simultánního překladu řeči trpí zejména když je zdroj zašumněný. Zkoumáme možnost využití více paralelních řečových signálů — originál a simultánní tlumočení — jako zdroje překladu s cílem dosáhnout vyšší kvality. Proto jsme vytvořili evaluační dataset ESIC (Europarl Simultaneous Interpreting Corpus). Dále analyzujeme aspekty simultánního tlumočení jako doplňkového paralelního zdroje. Poté zkoumáme odolnost vícezdrojového překladu proti chybám v přepisu a hodnotíme spolehlivost metrik strojového překladu na hodnocení simultánního překladu řeči. V neposlední řadě implementujeme Whisper-Streaming, nástroj na simultánní režim velkých offline modelů pro převod z řeči na text, který demonstruje současný stav poznání.

Klíčová slova: simultánní překlad řeči, překlad řeči, překlad z řeči do textu, strojový překlad, vícejazyčnost, vícezdrojovost, simultánní tlumočení, zpracování přirozeného jazyka

Acknowledgements

I want to thank my supervisor, Ondřej Bojar, for his guidance, passion and creating a motivating environment, and to my consultant and mentor, Raj Dabre, for his advice, welcoming in Japan, and fruitful collaboration. I am also very thankful for the excellent collaboration, especially for Peter Polák, Dávid Javorský, Matúš Žilinec, and many others, especially from ÚFAL, NICT, and from the ELITR project.

I am grateful to Ondřej Plátek, Martin Popel, Tomáš Musil, Mariia Anisimova, Danni Liu, Věra Kloudová, Dávid Javorský, Jindřich Libovický, Zdeněk Kasner, and prof. Ivana Čeňková for providing altruistic and very useful feedback on this dissertation and on my research.

Besides, I thank all the colleagues at ÚFAL, including those who provide administrative and technical support, fellow students, office mates, fellow researchers and others for creating a positive and supporting workplace that made my work, studies, trips, and my internship possible, pleasant, and memorable. I also want to thank the guarantors of the Ph.D. program at ÚFAL and to our ancestors for making the program smooth and excellent.

Lastly, I am grateful for the funding provided by the following projects, which directly and indirectly supported the research in this thesis: H2020 ELITR (European Live Translator; Grant ID: 825460), GAUK Strojový překlad tlumočených projevů (Machine translation of interpreted speech; Grant ID: GAUK 398120), Babel Octopus: Robust Multi-Source Speech Translation (START Programme of Charles University; Grant ID: START/SCI/089), GA ČR Mnohojazyčný strojový překlad (Multi-Lingual Machine Translation; Grant ID: 18-24210S), GA ČR EXPRO NEUREM3 (Neural Representations in Multi-modal and Multi-lingual Modeling; Grant ID: 19-26934X, RIV: GX19-26934X), LINDAT/CLARIAH-CZ (Language Resources and Digital Arts and Humanities Research Infrastructure; Grant ID: LM2023062), and SVV 260 575 (Teoretické základy informatiky a výpočetní lingvistiky).

Contents

1 Introduction

Neural machine translation (NMT) has the capability of handling more source or target languages at once (Firat et al., 2016a,b; Zoph and Knight, 2016; Johnson et al., 2017; Dabre et al., 2020; Kocmi et al., 2021). The desired effect of multilinguality are gains in translation quality, efficiency, or flexibility in comparison to bilingual NMT with one source and one target language (Kocmi et al., 2021, Section 2.2).

In this thesis, we focus on simultaneous multi-source speech translation from the original speech and simultaneous interpreting. We primarily focus on the use case of conferences with long-form monologue speeches where the participants and the speaker do not have a common language. It often happens in international multilingual organizations such as United Nations (UN), European Union (EU), or in European Organisation of Supreme Audit Institutions (EUROSAI). These organizations have several official languages that the speakers at the meeting can use, and there is simultaneous interpreting into the other official languages.

However, the set of official languages is often smaller than needed. The official languages usually include only the most spoken languages in the world, such as English, Arabic, Chinese, French, Spanish and Russian in case of UN, and English, German, French, Spanish, and Russian in EUROSAI. There may be participants who need assistance with understanding in e.g. Czech, Polish, Hungarian, Ukrainian, Turkish, Romani, or other languages that are not among the official ones. However, although the EU has 24 official languages that cover all the official languages of the member states and most of their population understands at least one of them, the exhaustive language support is usually available only at official meetings such as at the Plenary Sessions of the European Parliament. The language support is often limited by capacity reasons.

Another challenge is assistance with understanding the foreign language variety. When the speaker is asked to speak one of the official languages, it is likely that he or she is not entirely comfortable with that language and understanding may be difficult for those who are not familiar with the accent and grammar. Imagine, for example, that a usual Czech L1[1] user with LX knowledge of French and Russian rarely meets an Afghan citizen speaking LX Russian or Cambodian using LX French. Understanding the LX accents may be difficult. It may significantly complicate the mutual understanding.

Last but not least is the challenge of terminology. The meeting participants may need assistance with understanding e.g. specialized vocabulary, or foreign culture concepts. The need for assistance with understanding is therefore not limited to those who have no or little knowledge of foreign languages.

We see an opportunity that a natural language processing (NLP) technology may assist with understanding. Simultaneous speech translation (SST) is able to translate speech audio signal in the source language into text in the target language with a small additive latency (Müller et al., 2016; Niehues et al., 2018; Polák et al., 2023; Fukuda et al., 2023) of, e.g., 2 seconds on average. However, at multi-lingual meetings with simultaneous interpreting, there is a question, what should be the source language: Should we machine translate from the original or from the language of interpreting? Machines, in contrast to humans, can follow multiple speech signals at once. In text-to-text machine translation (MT), it was shown (Zoph and Knight, 2016; Firat et al., 2016b; Dabre et al., 2017) that using multiple parallel language text sources can lead to better translation quality, e.g. due to word sense disambiguation. In speech translation, we face the added challenge of speech recognition (Ruiz and Federico, 2014; Ruiz et al., 2017; Xue et al., 2020; Martucci et al., 2021), and multi-sourcing can be beneficial. Multiple parallel language sources may have complementary speech recognition errors. Therefore, we see an opportunity to investigate the methods of using multiple parallel language speech sources for SST. Figure 1.1 illustrates this use case and setup.

[1] Dewaele (2017) suggests the terms “L1 and LX users” for denoting “native and non-native speakers.”

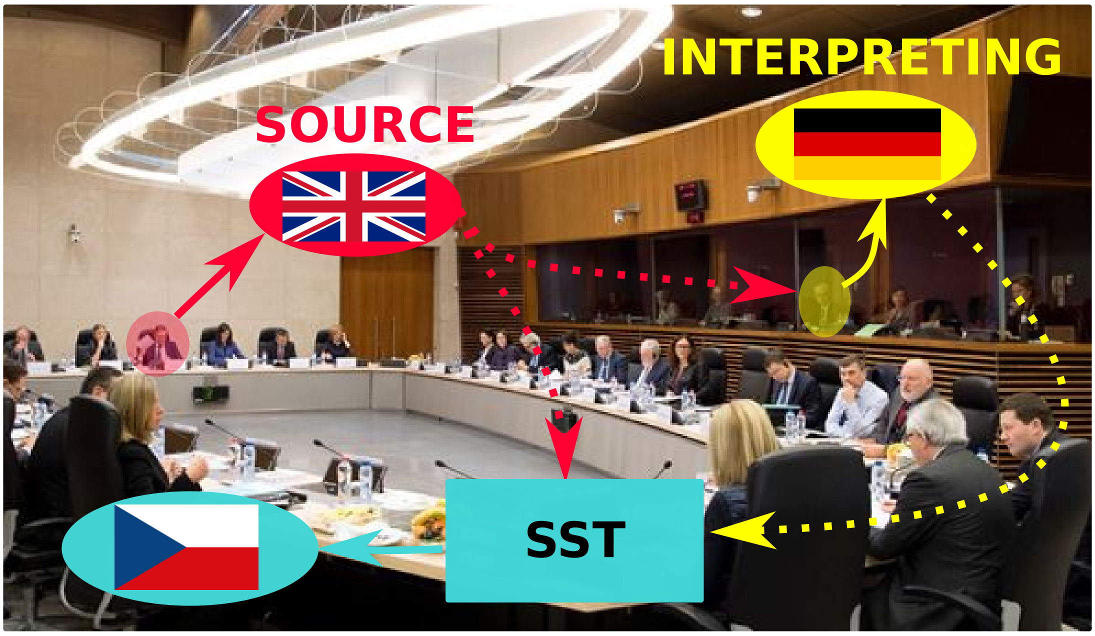


Figure 1.1: Illustration of the use case and setup of multi-source simultaneous speech translation. Imagine a multi-lingual meeting like the one in the picture. The original speaker is using English, which is simultaneously interpreted into German. Czech is another target language that is not covered by human interpreting for capacity reasons. Automatic simultaneous speech translation (SST) could provide Czech and could combine the two parallel language sources. (The photo was taken from `https://ec.europa.eu/education/knowledge-centre-interpretation/conference-interpreting/conference-interpreting-explained_en`.)

1.1 Long Story Short

We summarize our thesis into the following brief points:

Main finding Multi-source simultaneous speech translation from the original speech and parallel simultaneous interpreting may bring quality gains in certain situations, especially when the speech recognition quality of the sources is similar (Chapter 6 and Chapter 8, Macháček et al., 2023c).

Main contributions

1. ESIC – an evaluation corpus for multi-source SST with simultaneous interpreting from the European Parliament (Chapter 4, Macháček et al., 2021).
2. Analyses of simultaneous interpreting as a source of SST (Chapter 5, Macháček et al., 2021).
3. Experiments showing that multi-sourcing leads to robustness to ASR errors (Chapter 6, Macháček et al., 2023c).

4. Evidence that the offline text-to-text MT metrics are reliable in simultaneous mode (Chapter 7, Macháček et al., 2023a).

5. Whisper-Streaming – a practical tool that makes large offline ASR model work in simultaneous mode (Chapter 8, Macháček et al., 2023b).

6. We thoroughly describe the motivation, challenges, considered options, state of the art, and our experiments with details for reproductions (in this whole thesis, Chapters 1-9) to inspire others as much as possible.

Specification

- We primarily focus on research, and not on development. We propose methods for creating and using a practical SST system, and we experimentally evaluate the methods.

- The methods that we investigate and propose are language and domain-independent. However, we primarily experiment with one example set of languages and domain – English and German sources and Czech as a target, and speeches at the European Parliament (Chapter 4).

- We primarily focus our research on the text-to-text MT component of the cascaded SST system. We assume there are underlying ASR systems that produce text for the MT input. However, our methods and contributions are applicable also to direct speech-to-text SST (Chapter 3).

Future work We advanced the state of the art and set foundations for future research; however, more research needs to be done to apply multi-source SST to real-life use case (Chapter 8).

1.2 Publications

We elaborated this thesis in a 4-year Ph.D. program between October 2019 and October 2023. We continuously published the preliminary results of our research, mostly as articles at peer-reviewed conferences, but also in one book.

We briefly describe the publications and their relation to this thesis. We order them chronologically, and we introduce them with a brief "nickname." or the colloquial phrase we use to refer to them.

1. “AntreCorp” (Macháček et al., 2019) – shortly before our Ph.D. program, we were involved in the collection and creation of automatic speech recognition (ASR) test set that was later expanded by translation and simultaneous interpreting. We propose it as a possible option and do not use it in our experiments; however, AntreCorp was repeatedly used in the IWSLT Simultaneous Speech Translation Task (Ansari et al., 2020; Anastasopoulos et al., 2021, 2022).

2. “Subtitler” (Macháček and Bojar, 2020) – we got familiar with latency of retranslating SST, and we first used Continuous Rating, a method for collecting human rating of SST.

3. The ELITR system submission for IWSLT 2020 (Macháček et al., 2020) – we got familiar with a realistic cascaded SST system and with all the components and subtasks in the cascade. We used this ELITR system in our further work during the next year.

4. The “ESIC paper” (Macháček et al., 2021) – describes the creation of ESIC corpus and analyses of simultaneous interpreting.

5. “The Reality of Multi-Ling. MT” (Kocmi et al., 2021) – we co-authored a book where we wrote about our research before the Ph.D. program, but also our practical experience with ELITR live speech translation.

6. “Subtitler user study” (Javorský et al., 2022) – we collaborated on a study where we showed that Continuous Rating is a reliable method for collecting human feedback of SST.

7. “Robustness” (Macháček et al., 2023c) – this paper contains one of the central components of our dissertation. We create a multi-source model and test its robustness to speech recognition errors in both sources.

8. “MT metrics correlation” (Macháček et al., 2023a) – we analyze Continuous Rating in contrast to offline text-to-text MT metrics, and document that the metrics can be used reliably. Our motivation was to inspect the SST evaluation method for our further experiments.

9. Whisper-Streaming (Macháček et al., 2023b) – in order to be able to use a state-of-the-art simultaneous ASR as the source in multi-sourcing, we implemented simultaneous mode decoding for large offline ASR model Whisper, using the method from the recent SST competition at IWSLT. The tool was very innovative and effective, so we published it as a system demonstration.

Table 1.1 summarizes the titles of the papers, venues where the papers were published, references to the Bibliography section, and chapters or sections in this thesis where we use the texts that we wrote and previously published in four highlighted papers. We wrote the highlighted papers primarily on our own, with consultations and reviews of our co-authors, mostly Ondřej Bojar and Raj Dabre. We mark the published texts in this dissertation, and we modify and expand them. We also mark the pieces of work that were primarily elaborated by our colleagues Peter Polák and Matúš Žilinec.

Less related to our dissertation are four publications which we were involved in. There is a publication about the ELITR project (Bojar et al., 2020), and two publications about the ELITR complex and distributed system for live speech translation (Franceschini et al., 2020; Bojar et al., 2021a). We also co-authored the machine translation component in the CUNI system for IWSLT 2020 (Polák et al., 2020).

1.	A Speech Test Set of Practice Business Presentations with Additional Relevant Texts – Macháček, Kratochvíl, Vojtěchová and Bojar (2019) – Presented at SLSP 2019
2.	Presenting Simultaneous Translation in Limited Space – Macháček and Bojar (2020) – Presented at ITAT 2020
3.	ELITR Non-Native Speech Translation at IWSLT 2020 – Macháček, Kratochvíl, Sagar, Žilinec, Bojar, Nguyen, Schneider, Williams and Yao (2020) – Presented at IWSLT 2020
4.	**Lost in Interpreting: Speech Translation from Source or Interpreter?** – Macháček, Žilinec and Bojar (2021) – Included in Chapter 4 and Chapter 5 – Presented at INTERSPEECH 2021
5.	The Reality of Multi-Lingual Machine Translation – Kocmi, Macháček and Bojar (2021) – book
6.	Continuous Rating as Reliable Human Evaluation of Simultaneous Speech Translation – Javorský, Macháček and Bojar (2022) – Presented at WMT 2022
7.	**Robustness of Multi-Source MT to Transcription Errors** – Macháček, Polák, Bojar and Dabre (2023c) – Included in Chapter 6 – Published in Findings ACL 2023
8.	**MT Metrics Correlate with Human Ratings of Simultaneous Speech Translation** – Macháček, Bojar and Dabre (2023a) – Included in Chapter 7 – Presented at IWSLT 2023
9.	**Turning Whisper into Real-Time Transcription System** – Macháček, Dabre and Bojar (2023b) – Included in Section 8.4 – Presented at IJCNLP-AACL 2023 as system demonstration

Table 1.1: List of our publications relevant to this thesis or containing significant results reported in this thesis (highlighted in bold). The text that we previously published in the highlighted publications is included into this thesis, modified and extended.

1.3 Thesis Overview

Let us very briefly overview this dissertation thesis:

Chapters 1-3 introduce and explain the motivation and our primary focus. Chapters 4-8 describe our original research, and Chapter 9 concludes the thesis.

This dissertation is about the task of multi-source simultaneous speech translation from the original and simultaneous interpreting. This is a list of the chapters and questions they answer:

- Chapter 1: Introduction – Why is this task useful? Why to read this dissertation?
- Chapter 2: Motivation – The task is interesting and not thoroughly investigated yet.
- Chapter 3: Focus – What exactly is the task like? What do we focus on? What are the tasks we rely on?
- Chapter 4: Evaluation Data – How can we measure the progress in our research? Why and how we created the ESIC evaluation corpus?
- Chapter 5: Interpreting Analysis – What are the challenges of using simultaneous interpreting as a source in SST?
- Chapter 6: Multi-Sourcing and Robustness – Can multi-sourcing lead to a higher robustness to ASR errors?
- Chapter 7: Evaluation Questions – How to evaluate SST systems? Are MT metrics reliable in simultaneous mode? How to use them?
- Chapter 8: Multi-Sourcing in Reality – How to create a realistic multi-source SST system? What should be the next research directions? Why and how can we use off-the-shelf offline models in simultaneous mode?
- Chapter 9: Conclusion.

2 Motivation

In this chapter, we present and elaborate on the reasons why we primarily focus on the task of multi-source simultaneous speech translation (SST). We argue that multi-source SST is:

I. **Potentially useful.** As we mentioned in Chapter 1, the task has the potential to help to overcome the language barrier in a large number of practical situations, at multi-lingual conferences with simultaneous interpreting.

II. **Not thoroughly investigated yet.** As far as we know, the technology for multi-source SST is at a low maturity level, as we further elaborate in Section 2.1. We survey the most related work in Section 2.2.

III. **Challenging.** Multi-source SST is a challenging research task. Multi-sourcing may bring benefits of quality gains, but there are risks that it may not work in practice for various reasons. We elaborate this in Section 2.3.

IV. **Ready to work on.** There is a solid background of prerequisites that are necessary or useful for research in this area, and we have a good experience and access to them. More specifically, we have access to baseline automatic speech recognition (ASR), SST and data from the ELITR project, we have experience with machine translation tools and frameworks, we have close connections to research community that helps with consulting the issues, ideas and challenges, we have a strong institutional background and access to computer cluster with excelent IT support, etc. We can also use many natural language processing (NLP) tools and solutions that solve the basic subproblems, not needing to develop them first. We are also inspired by many previous and contemporary works and publications.

V. **Ethically eligible.** As far as we can tell, the only ethical reservations to our technology research are small and manageable. First, as with any technology, there is a risk of misuse, but this responsibility is on the users. The next risk is that the speech translation technology may bias some social groups, e.g. because those whose typical way of speaking is represented in training data will be served with higher quality than the others. This risk is manageable by the technology provider. Another risk arises from the fact that we publish some data during our research. While we respect valid regulations when publishing new data, subsequent regulations may prefer stricter conditions. Last, but not least, we discuss the potential social implications of new technology in Section 2.4.

2.1 Technology Readiness Level

Technology Readiness Level (TRL) is a standardized concept for describing the maturity level of technology.[1] It can be used to indicate the progress of technology research, which is our main contribution in this thesis.

The technology we focus on is the multi-source SST from the original and the interpreter. As far as we know, in 2019, when we started to work on this thesis, its TRL was 1 – basic principles observed. In the survey of multi-source and multi-target neural machine translation (NMT), Dabre et al. (2020) briefly mention a possible application of multi-sourcing to speech translation of the original speaker and the simultaneous interpreter at meetings of multi-lingual organizations. The expected quality gains were anticipated for the same reasons as in multi-source text-to-text NMT, primarily due to meaning disambiguation. In the survey, it was presented only as a possible idea, without any reference to technological concepts or description of experiments.

The low technology readiness level of multi-source SST offered us a good chance to focus on it.

2.2 Related Work

As far as we know, there is no previous work that investigates the combination of multiple sources, the original and the interpreter, in SST. However, there is some previous similar work, but missing some of the key features. We survey it in this section, ordered from the most to the least relevant.

[1] We use the TRL definition of the EU research and innovation programme HORIZON 2020: https://ec.europa.eu/research/participants/data/ref/h2020/wp/2014_2015/annexes/h2020-wp1415-annex-g-trl_en.pdf.

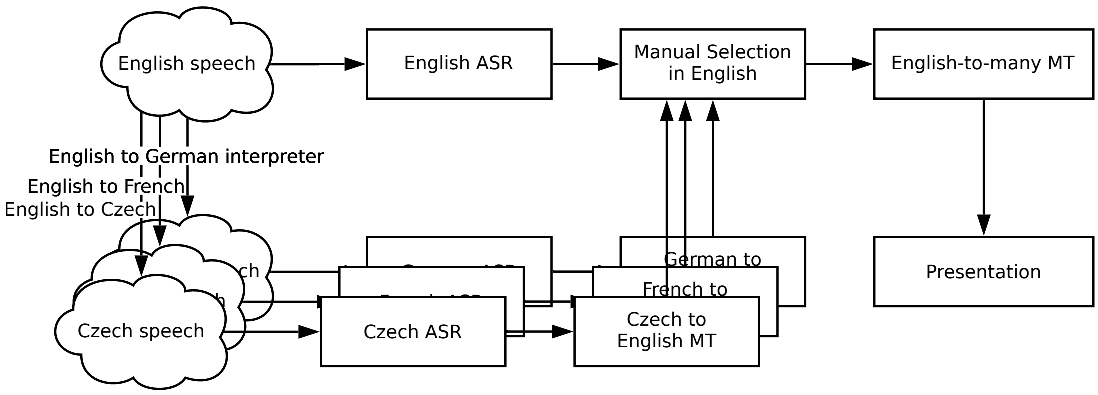


Figure 2.1: Illustration of ELITR multi-source system for translation from 4 alternative language sources (English original, or simultaneous interpreting into German, French, or Czech) using pivoting through English and manual selection of the best English candidate for multi-target translation. Figure reproduced from slide presentation of Bojar et al. (2021b).

2.2.1 ELITR: Multi-Sourcing by Manual Selection

Bojar et al. (2021b) describe a simple approach to multi-sourcing. If there are multiple source options, such as the original speech, or parallel simultaneous interpreting into four languages, such as in the ELITR setup at EUROSAI Congress in April 2021, one of the five parallel sources is manually selected as the source for single-source multi-target NMT. See the illustration in Figure 2.1.

However, the manual selection has obvious limitations. A human operator has to follow five text sources at once, which is very demanding and risky in simultaneous mode. Ideally, the latency between the time when the speaker utters a word and translation appears should be two seconds.[2] In such a short time, it is not easy to reliably detect which source is currently better. The ELITR setup uses pivoting, which means that the 4 languages are first translated into English, and then the best English candidate stream is selected, to be translated into 42 languages. This manual selection is achievable by one person who knows English, however, the best source could be different for each target language. There should be ideally one human operator for each target language. Alternatively, pivoting could be replaced by direct machine translation (MT), but then the human operator must have good knowledge of all the candidate languages, and that is usually not feasible for a high number of languages.

[2] IWSLT 2023 shared task on simultaneous translation requires average 2 seconds delay when not counting computational delay. Human simultaneous interpreting has an average delay of 4 seconds in English-Czech in European Parliament data (Macháček et al., 2021).

Furthermore, the ELITR system is not robust to fast switching between the sources. Since the interpreting tracks have different delays, it often happens that a part of the original content is repeated or missed because of the switch. Last, but not least, selecting the one single source does not allow combining the parallel sentences to exploit the information from multiple language sources, such as spelling of acronyms and proper names, word sense disambiguation, or correction of speech recognition errors.

Despite these limitations, we consider this system as an example of current state-of-the-art for multi-source SST from the original and interpreter. As far as we know, there is currently no other system prototype or reliable method that solves any of the mentioned issues.

2.2.2 Simultaneous Multi-Pivot NMT

Dabre et al. (2021) (a pre-print, not peer-reviewed) propose *simultaneous multi-pivot NMT*. The intended use is for the SST of a monolingual speech. First, the speech is automatically translated into multiple pivot languages by bilingual simultaneous NMT. Then, a multi-source simultaneous NMT translates it into the target language.

The paper contains the results of an experiment showing that multi-pivoting is more effective than pivoting through a single language. However, they experiment only with small multi-parallel training data, 200 000 training sentences from United Nations (UN) corpus (Ziemski et al., 2016). Their experiments are also limited to the simulation of simultaneity from the text-to-text test set, not considering realistic speech recognition errors or a realistic delay of the multiple pivots.

2.2.3 Offline Multi-Source ST

To the best of our knowledge, Wang et al. (2020) is the only work where any authors publish results of parallel multi-source translation of the offline speech. They create multi-parallel speech-to-text corpus CoVoST, from 11 languages into English. It contains isolated sentences that volunteers read and recorded for a massively multilingual Common Voice corpus (Ardila et al., 2020). The sentences are optimally segmented (one sentence per recording) and aligned, even in the test set. The usage of this model in practice is limited, the user has to record an isolated sentence twice, in at least two languages.

The authors report an improvement of the double-source model over the single-source, however, they do not publish any details on the double-sourcing architecture; they describe it only as “baseline multi-source.” Also, their training data are limited. They do not analyze the reasons for the benefit. As Kocmi et al. (2021) observe also in other situations, the benefits may not be coming from the multi-lingual source, but may arise because the longer input makes the encoder spend twice as many steps which allows for better encoding.

2.2.4 Multi-Source MT

Multi-source MT was first intended for text-to-text translation of individual sentences. It could be applied, e.g., in a large multi-lingual organization where many documents have to be translated in high quality into many target languages in a short time. If the text has already been translated into some languages and revised by human translators, all the revised parallel language variants may be used to increase the translation quality into other target languages.

Zoph and Knight (2016); Firat et al. (2016b); Dabre et al. (2017); Nishimura et al. (2018) and others propose multi-source NMT model architectures and training methods. They showed the benefits of multi-sourcing especially in setups with low amounts of training data. The architectures and training methods are inspiring and we apply some of them in our research, namely early and late averaging by Firat et al. (2016b) in Chapter 6.

Och and Ney (2001) describe multi-sourcing in text-to-text statistical machine translation (SMT).

2.2.5 Applications with Parallel Speech

Some publications describe solutions for various NLP tasks using combinations of multiple sources where at least one source is speech parallel to other sources, either speech or text. However, none of them involves simultaneous speech translation.

Speech enhanced CAT Khadivi and Ney (2008) propose a speech enhanced computer assisted translation (CAT) tool for translating text to text with human interaction. It records a human translator who says the translation out loud. The speech is then automatically transcribed and processed by a multi-source MT.

Multi-Source ASR Paulik et al. (2005); Miranda et al. (2012); Soky et al. (2022) use multiple parallel speech sources to enhance the ASR. The last one uses the most up-to-date neural architecture. Their expected and tested use case is the ASR of a low resource language, e.g. Khmer, that is interpreted into a high resource language, e.g. at court proceedings.

Punctuation restoration Miranda et al. (2013) use multiple parallel speech sources for punctuation restoration.

2.2.6 Interpreting in Training Data

As far as we know, the only use of simultaneous interpreting in MT that was published in previous works is using interpreting as an alternative source of parallel data. In some cases, automatic transcripts of simultaneous interpreting may be more accessible in large volumes than parallel translations.

Paulik and Waibel (2009) use ASR transcripts of simultaneous interpreting as training data for single-source phrase-based machine translation (PBMT) (Koehn et al., 2003). The core component in PBMT is a phrase table that is used to compute the statistics for alignment and reordering of source and target phrases. Since the source and target in PBMT are clean texts, and the interpreting data were used only in aggregation for statistics, they do not have to be properly aligned to parallel segments. They use overlapping time-aligned windows that do not have to be accurate on the margins. They also do not clean the data from the ASR errors because the noise has only a negligible effect in the phrase table.

Similarly, Paulik and Waibel (2008) extract ASR n-gram hints or phrases from the interpreter, and apply a discount factor when they detect them in the source of the statistical speech translation (ST).

These two papers were published within the TC-STAR project ("Technology and Corpora for Speech to Speech Translation," `tcstar.org`).

2.3 Benefits and Risks

In this section, we describe the expected benefits and risks of multi-source SST from the original and simultaneous interpreter that motivate us for our research. We also describe our plan on how to exploit the benefits and avoid the risks.

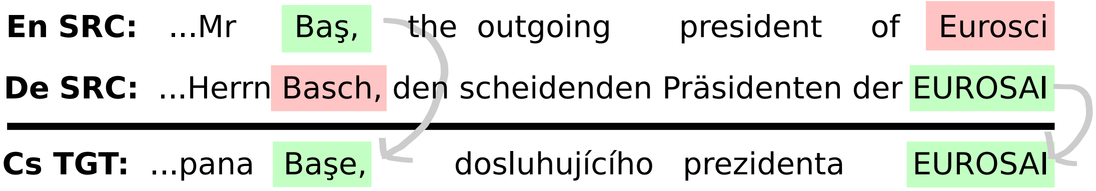

En SRC: ...Mr Baş, the outgoing president of Eurosci
De SRC: ...Herrn Basch, den scheidenden Präsidenten der EUROSAI
Cs TGT: ...pana Başe, dosluhujícího prezidenta EUROSAI

Figure 2.2: Example of how complementary speech recognition errors from two parallel language sources (English and German) can be used to benefit the target translation (into Czech). This example is selected artificially to illustrate the potential benefit.

2.3.1 Expected Benefits

Complementary speech recognition errors The most expected benefit of multi-source SST are quality gains due to the robustness to speech recognition errors. Speech recognition is one of the most challenging parts of speech translation (Ruiz and Federico, 2014; Ruiz et al., 2017; Xue et al., 2020; Martucci et al., 2021). As we illustrate in Figure 2.2, multiple parallel speech sources may have complementary errors, and a multi-source SST system can use all of them for better quality.

Disambiguation The second expected source of quality improvements is input disambiguation. For example, the Czech word “zámek” can be disambiguated when the word “lock” or “castle” in parallel English is available. Similarly, the English word “alien” can be disambiguated by the Czech parallel word “mimozemšťan” (extraterrestrial) or “cizinec” (foreigner). The two languages can complement each other. The disambiguation may be helpful and complementary at various language layers, not only in the lexical but also e.g. in morphology and syntax.

No human intervention needed With a fully automatic multi-source SST, no human intervention is necessary for selecting and switching the optimal source from several candidates.

Best from original and interpreting When translating from the original and simultaneous interpreter, it is possible to create a system that has the best features from both options. In Macháček et al. (2021), we found out that the translation from the source is more word-for-word, it may be more faithful than the interpreter. On the other hand, especially when the source speech is not completely fluent, e.g. due to hesitations, disfluencies, or L2 accent, it may be too complicated to follow. Similarly, the pace of the speech or limits on the receiver's end, such as small screen space can lead to unbearable speed requirements. In such cases, the fact that the interpreters summarize, reduce redundancies, use shorter and less complicated words and grammar, may be extremely useful.

Furthermore, the interpreters have a chance to adapt their interpreting to the target audience at the specific event. They may e.g. prefer or not prefer technical terms, or apply intercultural transfer. When present at the site, the interpreters may also easily handle any contextual references that the speaker may be using.

Last, but not least, the simultaneous interpreting is usually obtained in booths where good acoustic conditions may be easier to achieve than with the original speaker.

In more detail, we analyze interpreting in Chapter 5.

2.3.2 Risks

Latency In terms of translation latency, interpreting creates a delay. Translation from interpreting is more delayed than from the original, however, in Macháček et al. (2021) and in Chapter 5 we measured that the delay is feasible in low latency translation.

Risk of no improvement in practice We are aware of the risk that there may not be enough room where multi-sourcing can be meaningfully applied in practice, e.g. because the baseline – single source translation with simple heuristics for detecting the optimal source – can be very effective. Another possibility is that multi-sourcing may outperform single source when the speech recognition quality of all the sources is low, but in such cases, the speech translation quality may be also low and unusable.

In Chapter 6, we investigate whether there is room between "all too good" and "all too bad" sources where multi-sourcing could be beneficial. However, this room may be very small, including only a small fraction of all use cases. It is possible that developing and maintaining a special multi-sourcing system for rare cases may not be reasonable.

Challenges Multi-source SST consists of several challenging subtasks, e.g. simultaneous low-latency speech recognition, aligning the sources – original and interpreting, and segmentation to translation units. Simultaneous interpreting usually does not translate one source sentence into one target sentence, in contrast to text-to-text translation. Sentence segmentation is a problem.

The next challenge is to obtain training and evaluation data to train NMT for multi-sourcing with the original and interpreting. Another challenge is to make NMT working in the simultaneous mode, to translate incomplete sentence prefixes right at the time when the speaker is uttering them, in low, real-time latency.

Next, there are challenges of the speech source modality: ambiguity at all language layers from phonetics to syntax, noise, speaker adaptation, etc. Handling speech recognition errors is challenging.

Last, but not least challenge of multi-sourcing is to reconstruct the original meaning from the multiple noisy sources. The options are voting when we have at least three sources, confidence scores from the sources, or a neural network to detect the confidence by itself, from supervision and sufficient context.

Risk of high complexity Since the subtasks are in a sequence, they influence each other. It is possible that we substantially advance performance in our main subtask, but the entire solution fails because of low performance in another subtask beyond our scope. However, there is a simple mitigation strategy that we describe in the next section.

2.3.3 Risk Management

We have a plan to exploit the benefits and avoid the risks. We do our research in small subsequent steps, focusing on only some subtasks first, leaving the other ones to future work or to the other researchers.

We plan to document our progress in a simplified simulation in laboratory conditions. When working on a task that requires expected, but not yet existing solutions of the underlying tasks, we plan to simulate them artifically.

Task decomposition, prioritization, and a meaningful approximation of the out-of-scope conditions is an interesting and challenging part of research.

2.4 Social Implications

When we present the vision of simultaneous speech translation technology in front of somebody who is not in the field of technology research or development, a common reaction of that person is a question of whether the interpreters lose their jobs, whether they will be replaced by technology. We propose an answer by Fantinuoli (2019) from his paper "The Technological Turn in Interpreting: The Challenges That Lie Ahead."

Inevitable drives In the first part, Fantinuoli (2019) describes three inevitable drives that force people to adopt new technology. The first drive is anthropological. It is human nature to prefer reducing necessary labor with tools, technology, or efficient methods. The second is economic, the technology saves resources. The last drive is psychological. People are worried about being less productive than their competitors if they refuse to adopt technology that others adopt or could adopt. They would feel endangered, so they rather follow the trend and adopt the technology.

Our point of view is that our SST research does not simply take away anyone's job. We rather say that jobs will probably undergo a change related to new technology. We, the researchers, are motivated by the three drives to research new technology. At the same time, people are motivated to adopt it when it becomes available. We notice that people often express their fears of the technology that is unavailable and unknown to them.

Stages of automation Second, Fantinuoli (2019) describes four subsequent stages of automation that technologies go through depending on the level of development and adoption. He gives an example of aviation, but we see the same stages in e.g. NLP tasks.

The first stage is that the task is fully dependent on human labor, e.g. ear-to-ear simultaneous interpreting.

The second stage is human labor enhanced by tools, e.g. simultaneous interpreting (SI) using audio transmission through microphones and headphones. It enlarges the impact by reaching more people.

The third stage are tools that reduce the trivial, repetitive and labor-intensive subtasks ("autopilots"), but a human expert must supervise it and operate the subtasks that are not reliably automated yet. In SI, there are computer assisted interpreting (CAI) tools that e.g. detect and write down numbers and acronyms, or display and suggest vocabulary terms when necessary, but human interpreter performs the rest. In text-to-text translation, the third stage is automatic translation that needs post-editing.

The last, fourth stage of automation, is the fully autonomous, reliable, and widely adopted tool. In many cases, the ethical and legal aspects need to be resolved before reaching this stage, e.g. the liability of fully automatic SI that impacts court decisions.

In summary, our research primarily aims at the fourth stage, but we are aware that it is rather a long-term goal to which we contribute, but we do not reach it on our own. However, we assume that our results will be applicable in the second and third automation stages as well.

Low-end applications Fantinuoli (2019) mentions applications of technologies in the third stage of automation (tools under human supervision) that create new markets and job opportunities. They are usually applied to low-end segments where lower quality for low cost is acceptable. In SI, there are e.g. speech translation tools that assist with interlingual communication while travelling. If the tool makes an error, the users usually recognize it immediately and try again, or they try another way, e.g. paralinguics or phone assistance, or they accept the miscommunication. Such users would probably not travel without the technology and the opportunity to serve them would not exist. The higher-end segment where human SI is applied are either not affected by the third stage technology, or the human experts are more productive thanks to it.

Third stage technology in SI Next, Fantinuoli (2019) summarises and predicts new technology that assists humans in SI, e.g. the CAI tools, computer-assisted interpreter training, better quality ASR in the CAI tools, tools that extract and select information from the background and context data that interpreters need to study before an event, or interpreting management systems. Last, but not least, he summarizes the benefits and costs of remote interpreting.

Fourth stage challenges in SI Fantinuoli (2019) lists the reasons why it is very challenging to create automatic SI that could work reliably without human supervision. He claims that post-editing supervision is impossible in simultaneous mode, because of the low latency. We, on the contrary, consider it possible, although nowadays risky and slow. However, further research can advance it. The other large challenge in fully automatic SI is the enormously large complexity of communication that entails for more than just speech. We assume that in principle, the entire multi-modal communication is learnable by machines from data, and it is possible, although challenging, to collect such data.

Fantinuoli (2019) concludes with the statement that the new technology needs to be thoroughly evaluated before release. We agree.

3 Focus

In this chapter, we specify and explain the task of multi-source simultaneous speech translation (SST) from the original and interpreting, and describe and explain our primary focus: multi-source text-to-text machine translation (MT) component of cascaded SST for long-form monologues.

Then, we describe the typical SST system from sound acquisition to user interface, with all the tasks and components that lie before and after MT.

3.1 Task Specification

We primarily focus on simultaneous speech translation from multiple parallel language sources, from the original speaker, and parallel simultaneous interpreting. Moreover, we primarily focus on long-form monologue speech and the text-to-text machine translation component of speech translation that can be used within a cascaded SST system. We explain the mentioned concepts in the rest of this section.

3.1.1 Speech Translation

We use the term *speech translation (ST)* for denoting a task of translating speech in the source language into text or speech in the target language, without the functions that human interpreters typically provide in addition to translation. We consider speech translation as a subtask of interpreting.

There also exists a related term *spoken language translation (SLT)*. Some authors use it as a synonym for ST. However, in our work, we distinguish between ST and SLT. SLT is text-to-text machine translation of the spoken language domain, with standard normalized text on the input, while ST input is audio. The spoken language domain may cover e.g. texts prepared for spoken production or transcribed and normalized speech.

3.1.2 Simultaneity

We focus on *simultaneous* speech translation (Niehues et al., 2018), which is also called *online*, *real-time*, *low-latency*, *live*, etc. It means that the speech is translated at the same time as it is being produced and recorded, only with a small additive delay, e.g. 2 seconds. See the illustration in Figure 3.1.

The need for simultaneity arises when we want to enable the target audience to interact with the original speaker in real time, which is usually useful in multi-lingual meetings and conferences. Furthermore, the additive latency enables targeting many languages at once without blocking time for each language which would serve only part of the audience. The users receive the translations individually, without influencing others, either as text captions on their personal devices or as audio in their headphones.

In contrast to simultaneous translation, *offline* speech translation is applied to pre-recorded audio. In simultaneous speech translation (ST), there are two objectives, latency and translation quality, while in offline ST, the only objective is quality.

Streaming vs. re-translation

There exist two main approaches to simultaneous MT: streaming and re-translation (Niehues et al., 2018; Arivazhagan et al., 2020b). In both approaches, the system continuously receives an input text segmented to sentences, as produced by the speaker and the underlying automatic speech recognition (ASR). In simplified simulation it is assumed that after receiving every token, there is unlimited time for processing it before the next token arrives. It is an unrealistic, computationally unaware simulation. In reality, more tokens may arrive at once, or during the time when previous tokens were processed.

Re-translation Re-translation systems (usually) generate translation from the beginning of the sentence whenever a new part of the source arrives. Ideally, it should append new words to the end of the previously produced target, but often it also changes words in the middle of the sentence, as illustrated in Figure 3.2. The advantage of re-translating is that the outputs can be available fast, e.g. translated in 200

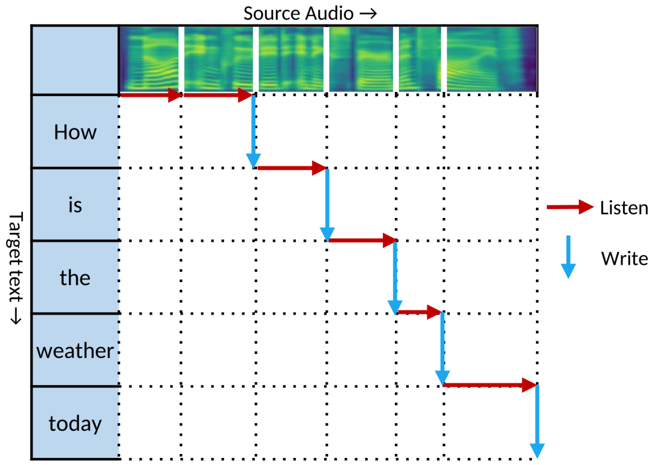


Figure 3.1: Illustration of *simultaneous* speech translation. The source audio (represented by spectrogram in the figure) arrives incrementally in time segments (columns). After receiving each segment (moving to the right, Listen operation), a SST system outputs zero or more target tokens (Write, moving down). The outputs are thus available incrementally and *simultaneously* with the inputs, which means *nearly at the same time*. Figure reprinted from Ren et al. (2020).

Source	Output								Erasure
1: Neue	New								-
2: Arzneimittel	New	Medicines							0
3: könnten	New	Medicines							0
4: Lungen-	New	drugs	may	be	lung				1
5: und	New	drugs	could	be	lung	and			3
6: Eierstockkrebs	New	drugs	may	be	lung	and	ovarian	cancer	4
7: verlangsamen	New	drugs	may	slow	lung	and	ovarian	cancer	5
Content Delay	1	4	6	7	7	7	7	7	

Figure 3.2: Illustration of re-translation MT reproduced from Arivazhagan et al. (2020b). Erasure is a measure of stability. There is a sequence of re-translation updates, top to bottom. Erasure indicates how many target tokens from the end of the previous update need to be erased before appending tokens for the current update. The fewer erasures, the better.

milliseconds when using Marian neural machine translation (NMT) as in our ELITR submission at IWSLT 2020 (Macháček et al., 2020). However, the outputs may be unstable; the user may have a chance to read the preliminary version, and then need to read it again, with the final version. The outputs may also change ("flicker") so frequently that they become unreadable. The flickering may be alleviated by finetuning NMT on source-target prefixes (Niehues et al., 2018).

The advantage of re-translation is the fact that the final hypothesis is generated from the whole sequence that uses all possible source context, not only the prefix, so the quality is identical to offline MT.

Re-translation was used in the ELITR project. The motivation for it was that the users who have no knowledge of the source language need high quality and tolerate longer waiting. The users who have some but limited knowledge of the source language need to look at the automatic translations occasionally when they do not understand. They need low latency and tolerate lower quality of the earlier hypotheses.

In Macháček et al. (2020), we created a tool MT-Wrapper that enables using any offline sentence-level NMT model in re-translation mode. We validated the practical usability of this tool on many real-life events. It allows translating of multiple subsequent sentences in one batch, caching because re-translation may revert the previously translated inputs, and skipping the updates that became outdated during processing the previous ones so that MT-Wrapper always catches up the most recent inputs as soon as possible.

Streaming The streaming[1] systems either produce one or more target tokens, or decide to read the next input token to have more context for translation. The longer they wait, the more context is available and, usually, higher quality is produced. But, obviously, the cost for waiting has to be paid by the user. On the other hand, streaming achieves optimal stability because the new target words are only appended. We illustrated streaming simultaneous ST in Figure 3.1.

The goal of streaming simultaneous ST is to translate the input with high quality and low latency. Quality is measured on full sentences as in standard text-to-text MT, e.g. with BLEU or other metrics (more on this topic in Chapter 7). The standard latency measure applicable to streaming is Average Lagging (AL, Ma et al., 2019). It is the average number of tokens behind an "optimal" policy that generates the target proportionally with reading the source.

[1] We prefer this term, but the terminology is not yet settled universally. Some authors use the term "simultaneous" when they mean the streaming approach to simultaneous ST (Chang et al., 2022; Papi et al., 2023c), some others use "incremental" (Polák et al., 2023; Polák, 2023).

Our focus In this thesis, we do not focus on advancing the streaming or re-translation methods; we use the state of the art proposed by others.

We use re-translation in our work in Chapter 5, in Macháček and Bojar (2020); Macháček et al. (2020); Javorský et al. (2022), and in the ELITR project (Franceschini et al., 2020; Bojar et al., 2021a). More details on the re-translation system in this thesis are in Section 5.3.2.

Later, in Section 6.4 and in Whisper-Streaming (Section 8.4), we use the streaming approach because the most recent research findings also focus on streaming.

3.1.3 Long-Form Monologue

We primarily focus our research on long-form monologue speech. *Long-form* means that the speech is uninterrupted, consisting of multiple utterances in a sequence, with inter-sentence coherence, and without any explicit marking of sentence boundaries. The alternative is "segmented speech," which is long-form with provided segmentation to sentences (or sentence-like units). However, the sequence of segments typically comes from a long-form coherent speech, so the inter-sentence context can be useful. Segmented speech is often used in research, e.g. at IWSLT shared tasks, in order to evaluate speech translation systems independently on the speech segmentation tool.

The other alternative to long-form speech is single utterances, e.g. in speech translation applications for travelers such as VoiceTra (Misu, 2010; Matsuda et al., 2013) where the users are requested to record one simple sentence for individual translation. No inter-sentence context is expected.

Monologue means that the speech is given by one speaker. The alternative is a dialogue of two or more speakers, possibly with speakers overlap. We focus on monologues because they contain all the linguistic challenges as dialogues, including references to previously spoken content. As a baseline, dialogues can be translated with a monologue translation system using simple adaptation, e.g. as a monologue, only highlighting the speaker turns in output, or as many subsequent monologues, each consisting of one speaker turn.

Source quality challenges We must be aware that the source speech might not be of an expected quality. For example, a speech might be read or spontaneous, while ideally, only the latter case should be a case for SST. If the speech is available as text to be read, the text can be used for translation. Next, speech may be very smooth and fluent, or it might contain disfluencies (false starts, repetitions, hesitations, filler

Figure 3.3: Illustration for explaining the difference between interpreting and translation. Reprinted from the American Translators Association website (`https://www.atanet.org/client-assistance/translator-vs-interpreter/`).

words), pauses at random places, without connection to syntax or meaning, interruptions by other speakers and non-linguistic sounds (applause, laughter, cough) etc. The speech might contain code-switching (insertions of other language), and the speaker might have a specific, or non-native accent.

However, we simplify our focus on cases where these challenges do not appear or are resolved by external tools, e.g. voice activity detection (Silero, 2021) to filter out non-voice sounds, speech reconstruction that detects and removes disfluencies (Češka, 2009; Chen et al., 2020), etc.

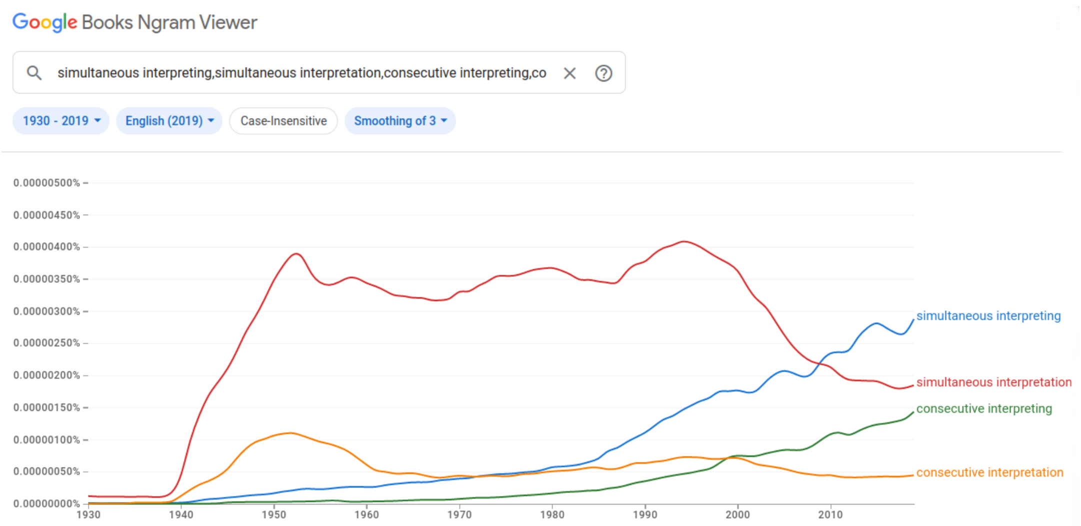
3.1.4 Interpreting

Since we focus on SST from two parallel multi-lingual sources, the original, and simultaneous interpreting, let us explain the term *interpreting*, in contrast to translation.

Professional translators and interpreters (for example Ešnerová, 2019) distinguish two tasks that are usually performed by human experts: *Translation* is processing text in the source language into text in the target language. *Interpreting* is processing speech in the source language directly into speech in the target language, to mediate the communication between the speaker and the audience. See a nice and brief explanation by the American Translators Association (ATA),[2] and illustration in Figure 3.3.

Interpreting involves more than just translating words. The interpreters provide also the inter-cultural transfer (explaining concepts that may not be known in the culture associated with the target language), they explain the background that was not uttered, but the audience might not be aware of it and might need it to under-

[2] `https://www.atanet.org/client-assistance/translator-vs-interpreter/`



simultaneous interpreting,simultaneous interpretation,consecutive interpreting,co

1930 - 2019 ▾ English (2019) ▾ Case-Insensitive Smoothing of 3 ▾

Figure 3.4: Google NGram Viewer displaying frequencies of bigrams "simultaneous interpreting, simultaneous interpretation, consecutive interpreting, consecutive interpretation" in English books published in the years 1930's to 2019. We observe that the word interpreting has becoming more frequent since 2000's in collocation "consecutive interpreting," and since 2010's in "simultaneous interpreting."

stand. Furthermore, the interpreters handle inappropriate words and offenses in a suitable way, they comment on the actions on the stage and provide organizational comments, when necessary. They use not only speech on their input but complete audio-visual information (paralinguistics, e.g. who is addressed by a gesture, etc.), and meta-information, such as current time, location, the event schedule, slides and other relevant documents, etc.

Interpreting or Interpretation?

Interpreting and *interpretation* are synonyms. Using one or the other is arbitrary, both are used by active interpreters and in research literature in the context of mediating foreign language speech. However, we decided to prefer the term interpreting because we noticed it is more frequent in recent research publications.

Moreover, we found an evidence in Google NGram Viewer (Michel et al., 2011) that simultaneous (and consecutive) interpreting term is becoming more frequently used than interpretation. See Figure 3.4. The terms were compared in corpus denoted as English 2019. It is described as "Books predominantly in the English language published in any country." The same trend is in the American English corpus, but not in the British. However, we prefer to follow the global and recent trend.

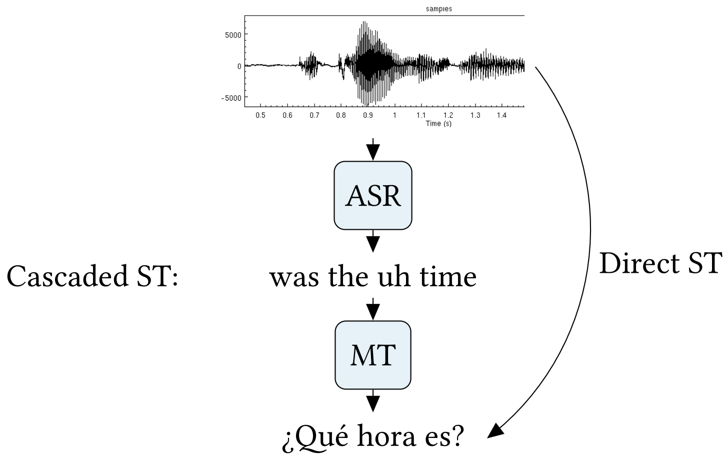


Figure 3.5: Illustration of cascaded vs. direct ST. The source is an audio signal at the top. The cascaded system consists of ASR that transcribes the audio in the source language, in this case without any punctuation and casing, including hesitation "uh" that is supposed to be dropped, and including errors – the intended message was "What's the time?" Then, a text-to-text MT system translates it into the target language Spanish, ideally by correcting the ASR errors. A direct system translates audio into the target language directly. Figure reprinted from ELITR project, originally by Barry Haddow.

3.2 Focus: Text-to-Text MT in SST

The automatic speech translation systems can be *cascaded*, or *direct* (also called *end-to-end*). Sperber and Paulik (2020) overview and compare them, and also describe their hybrids. The cascade is a pipeline of individual systems for processing intermediate tasks, e.g.: (i) ASR, (ii) normalization, which may include removing disfluencies, expanding or compressing acronyms, digit normalization, inserting punctuation, and truecasing, and (iii) machine translation (MT). See illustration in Figure 3.5.

The advantage of cascaded ST is the possibility of distributed development and easier training. There are more labeled training data for ASR and MT separately than for speech-to-text in another language (ST). The ASR and MT systems can be trained separately. The disadvantage of cascaded ST is error propagation between the sub-systems. In the alternative direct approach, the ASR and MT are provided by one compact neural network that may reduce the error propagation, due to unsupervised information flow between the sub-tasks. On the other hand, the direct speech-to-text translation training data for supervised training may be small, and the training for high quality is therefore challenging.

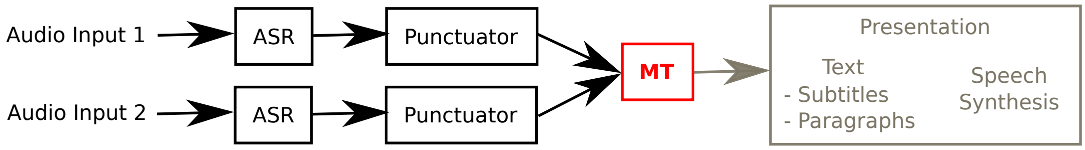


Figure 3.6: A simple scheme of cascaded multi-source SST system from two independent language sources. The multi-source text-to-text MT component, on which we primarily focus, is highlighted in red. The presentation options of MT outputs are unspecified, we draw them in gray.

Although Bentivogli et al. (2021) claim that the performance gap between cascade and direct ST is negligible, they observe it in three language directions from English. We suppose that this is not the case for other language pairs, especially low-resource ones. Moreover, we aim to study parallel multi-sourcing methods. It is easier to implement them in the text-to-text MT than in speech-to-text ST, however, we suppose that they can be later adapted to direct ST as well. Therefore, we primarily focus our research on the multi-sourcing methods in the MT component of SST, as we illustrate in Figure 3.6. We assume that our multi-source MT receives punctuated text transcripts from underlying ASR systems. The MT outputs are then presented with an unspecified method, e.g. either long text in paragraphs or short subtitles, or as synthesized voice. We explain all the typical cascade components in Section 3.3.

3.3 Live Speech Translation Service

In this section, we highlight that we primarily focus on research for advancing the quality of SST, and more specifically, its MT component. The MT is supposed to be deployed as a component of service for live speech translation of multi-lingual meetings or conferences. There are other upstream and downstream components that lie before and after MT in the cascade, and that may influence the overall quality of the service. We suppose they are of sufficient quality, but we want to highlight that the other components than MT are not under our control. On the other hand, all of them, with some reservations to ASR, are highly developed and are not a research challenge anymore, but rather standardized tools. We overview them in this section.

A typical live speech translation service, e.g. Google Translate simultaneous speech-to-text, ELITR (Bojar et al., 2021a; Franceschini et al., 2020; Bojar et al., 2021b), KIT Lecture Translator (Cho et al., 2013; Dessloch et al., 2018), etc. is composed of the following components:

1. Human operator – a person that operates the service, i.e. installing the hardware, starting and setting up the software, and instructing the target users to access and understand the outputs. The operating person must be knowledgeable of the service requirements and limitations.

2. Sound acquisition. The very first technical component is sound acquisition, or recording the speaker's voice in sufficient acoustic conditions for further digital processing. It consists of the following steps:

   (a) Microphone – a sound recording device. There are many types, the selected microphone should be suitable for the specific conditions. Features to consider are e.g. sensitivity to voice sound frequencies, directionality – capturing sound only from short distances, or from long distances, including noise. Practical aspects are also important, e.g. wiring, wireless connection and batteries, instructing the speaker to use it correctly, etc.

   (b) Digitization. An electronic device typically transforms acoustic analog signal into digital stream. It involves sampling the sound into short time segments (e.g. 16 000 per second), and representing the sound wave energy in each time segment as a digital number, e.g. 16-bit low-endian signed integer. This process is also called PCM – pulse-coding modulation.

   (c) Encoding. The digitized audio stream may be encoded into a specific digital format for further processing. The formats typically differ by compression method, bit length – accuracy of digitization, and sampling frequency.

   (d) (Optional) noise cancellation – some systems involve noise canceling algorithms. They are included especially in teleconferencing applications where a laptop microphone captures the human voice at the same time as noise, e.g. produced by a laptop fan, or the laptop loudspeakers playing the sound from the other side of the call. On one hand, the resulting recording consists only of the relevant part – local voice, and prevents echo, on the other hand, may sound unnatural to humans and affect their perception (Newman and Schwarz, 2018).

   (e) Transmission – the digital signal has to be transmitted to the next components of the system, e.g. through a computer network. The transmission should be fast and reliable.

   (f) (Optional) transmission error correction. E.g. speed up a short speech segment to catch up a short connection break, as it is sometimes applied in teleconferencing applications.

3. Language identification. The SST often requires one or more specific languages on the source. There may be a component that detects the language of the speech (Bartz et al., 2017; Valk and Alumäe, 2021), and then triggers a corresponding setup of further components. If there is no automatic component in the service, ensuring correct language on the source relies on the human operator.

4. (Optional) speaker diarization. In some cases, it may be useful to detect when speaks who, e.g. to mark it in the outputs, or to set up the downstream components accordingly. Speaker diarization methods (Park et al., 2022) can be applied.

5. SST – a simultaneous speech translation can be decomposed into multiple subsequent subtasks. An SST system can be either a cascade of independent tools or a direct neural model that processes multiple or all following tasks at once.

    (a) Simultaneous processor – every independent simultaneous tool in the cascade needs a simultaneous processor. It usually ensures incremental processing and implements a specific simultaneous protocol, e.g. retranslation or streaming (Arivazhagan et al., 2020b).

    (b) (Optional) voice activity detection (VAD) – to filter out non-voice sounds and silence (Veysov and Voronin, 2022; Silero, 2021) from the ASR input.

    (c) ASR – consists of:

        i. segmentation of source segments into processing units, e.g. 30-second segments including one full sentence. There are automatic tools, e.g. SHAS (Tsiamas et al., 2022).

        ii. transcription – the ASR models usually consist of acoustic and language modeling parts, either explicitly modeled, or implicitly. See an overview of neural ASR model architectures in Papastratis (2021).

        iii. casing, punctuation – old fashioned ASR models produced uncased and unpunctuated text, but casing and punctuation is often useful in downstream MT models and in presentation. Truecasing (Lita et al., 2003) and punctuation restoration (Chordia, 2021; Alam et al., 2020) can be applied. Recent ASR models (Radford et al., 2022; Pratap et al., 2023) usually produce cased and punctuated transcripts.

iv. (optional) normalization – sometimes, the text format outputed by ASR has to be normalized before MT. It may include digit normalization, e.g. transforming words to digits, removing disfluencies (false starts, corrections, hesitations), currency, and sometimes handling special sound marks detected by ASR, such as laughter, noise, applause, silence, etc.

(d) MT

i. segmentation – a typical MT model processes individual sentences. For that, a sentence segmentation tool may be applied, e.g. ERSATZ (Wicks and Post, 2021), WtPSplit (Minixhofer et al., 2023), or a tool from Moses (Koehn et al., 2007).

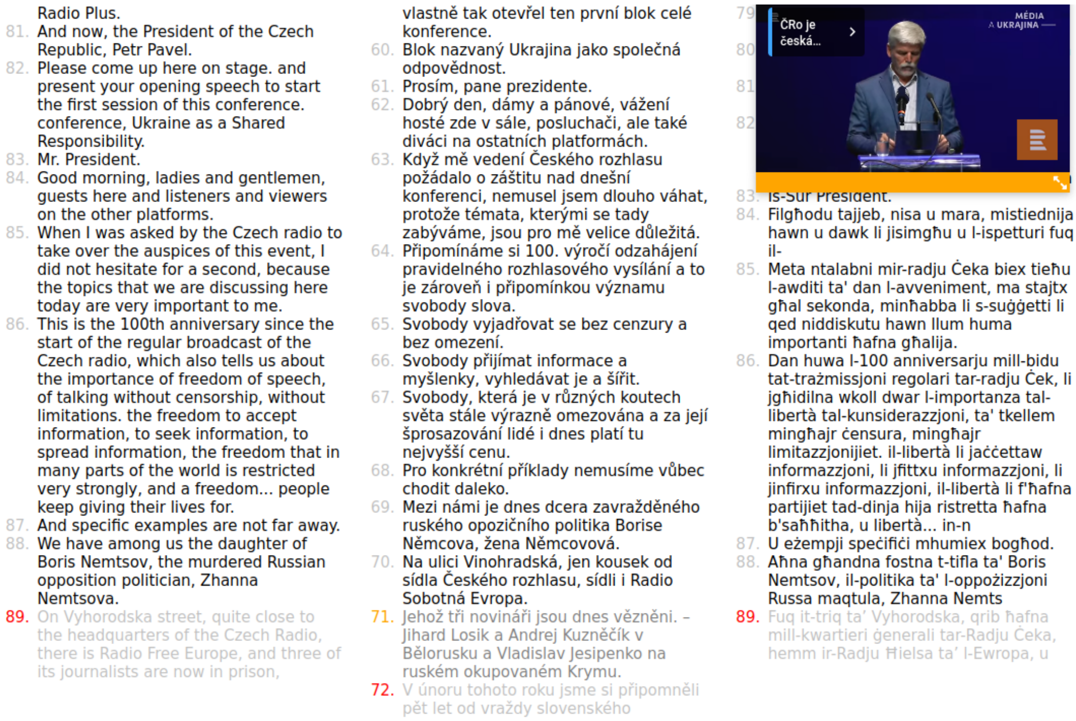
ii. translation – usually applying the NMT model, e.g. Transformer neural architecture consisting of an encoder and decoder. In SST, the NMT model may be adapted to simultaneous mode.

6. (optional) Moderation, post-editing – sometimes it is advisable to have a human operator monitor the outputs and moderate them, e.g. to abort or restart translation if a wrong and inappropriate word appears in outputs that would bring an undesired attention of the audience, as in the example in Figure 9.3 in Kocmi et al. (2021). In live television broadcasting, human post-editors are sometimes able to edit the live subtitles, e.g. in SubtitleNEXT editor.[3]

7. User interface – a platform where the users can access the translation, either in the form of text, or speech. The text form can be projected e.g. on a shared screen that is visible to the whole audience, or accessible on personal devices on a web page. Speech can be delivered to personal receivers with headphones. The reliability and robustness of the user interface are critical, however, in Javorský et al. (2022) we found that small alternations of presentation options are rather insignificant.

8. Presentation – it is always advisable to display as long text of translations as possible, i.e. whole paragraph view as in Figure 3.7. If the space is limited, e.g. because of the need to project video or presentation slides, and there can be several lines of subtitles (Macháček and Bojar, 2020).

[3] https://subtitlenext.com/subtitlenext-powerfully-accelerates-live-subtitling-with-google-asr-integration/



Radio Plus.
81. And now, the President of the Czech Republic, Petr Pavel.
82. Please come up here on stage. and present your opening speech to start the first session of this conference. conference, Ukraine as a Shared Responsibility.
83. Mr. President.
84. Good morning, ladies and gentlemen, guests here and listeners and viewers on the other platforms.
85. When I was asked by the Czech radio to take over the auspices of this event, I did not hesitate for a second, because the topics that we are discussing here today are very important to me.
86. This is the 100th anniversary since the start of the regular broadcast of the Czech radio, which also tells us about the importance of freedom of speech, of talking without censorship, without limitations. the freedom to accept information, to seek information, to spread information, the freedom that in many parts of the world is restricted very strongly, and a freedom… people keep giving their lives for.
87. And specific examples are not far away.
88. We have among us the daughter of Boris Nemtsov, the murdered Russian opposition politician, Zhanna Nemtsova.
89. On Vyhorodska street, quite close to the headquarters of the Czech Radio, there is Radio Free Europe, and three of its journalists are now in prison,

vlastně tak otevřel ten první blok celé konference.
60. Blok nazvaný Ukrajina jako společná odpovědnost.
61. Prosím, pane prezidente.
62. Dobrý den, dámy a pánové, vážení hosté zde v sále, posluchači, ale také diváci na ostatních platformách.
63. Když mě vedení Českého rozhlasu požádalo o záštitu nad dnešní konferenci, nemusel jsem dlouho váhat, protože témata, kterými se tady zabýváme, jsou pro mě velice důležitá.
64. Připomínáme si 100. výročí odzahájení pravidelného rozhlasového vysílání a to je zároveň i připomínkou významu svobody slova.
65. Svobody vyjadřovat se bez cenzury a bez omezení.
66. Svobody přijímat informace a myšlenky, vyhledávat je a šířit.
67. Svobody, která je v různých koutech světa stále výrazně omezována a za její šprosazování lidé i dnes platí tu nejvyšší cenu.
68. Pro konkrétní příklady nemusíme vůbec chodit daleko.
69. Mezi námi je dnes dcera zavražděného ruského opozičního politika Borise Němcova, žena Němcovová.
70. Na ulici Vinohradská, jen kousek od sídla Českého rozhlasu, sídlí i Radio Sobotná Evropa.
71. Jehož tři novináři jsou dnes vězněni. – Jihard Losik a Andrej Kuzněčík v Bělorusku a Vladislav Jesipenko na ruském okupovaném Krymu.
72. V únoru tohoto roku jsme si připomněli pět let od vraždy slovenského

79

80

81

82

83. Is-Sur President.
84. Filgħodu tajjeb, nisa u mara, mistiednija hawn u dawk li jisimgħu u l-ispetturi fuq il-
85. Meta ntalabni mir-radju Ċeka biex tieħu l-awditi ta' dan l-avveniment, ma stajtx għal sekonda, minħabba li s-suġġetti li qed niddiskutu hawn llum huma importanti ħafna għalija.
86. Dan huwa l-100 anniversarju mill-bidu tat-trażmissjoni regolari tar-radju Ċek, li jgħidilna wkoll dwar l-importanza tal-libertà tal-kunsiderazzjoni, ta' tkellem mingħajr ċensura, mingħajr limitazzjonijiet. il-libertà li jaċċettaw informazzjoni, li jfittxu informazzjoni, li jinfirxu informazzjoni, il-libertà li f'ħafna partijiet tad-dinja hija ristretta ħafna b'saħħitha, u libertà… in-n
87. U eżempji speċifiċi mhumiex bogħod.
88. Aħna għandna fostna t-tifla ta' Boris Nemtsov, il-politika ta' l-oppożizzjoni Russa maqtula, Zhanna Nemts
89. Fuq it-triq ta’ Vyhorodska, qrib ħafna mill-kwartieri ġenerali tar-Radju Ċeka, hemm ir-Radju Ħielsa ta’ l-Ewropa, u

Figure 3.7: A preview of presentation interface of simultaneous speech transcripts (middle column, in Czech) and translation (side columns, English and Maltese) in the form of paragraphs – a list of succeeding sentences that continuously appear from top to bottom. The gray color of the text at the bottom indicates recent text that can be edited by the system. The interface is from ELITR (Bojar et al., 2021b), and the speech is a part of Czech Radio conference Media and Ukraine, 22nd June 2023. It served also as an event for evaluation of Whisper-Streaming (Macháček et al., 2023b, Section 8.4 in this thesis), and for collecting data for further research (Section 4.4).

4
Evaluation Data

Evaluation data from an authentic use case are necessary for measuring the state of the art and detecting the progress of research.

In this chapter, we describe the requirements for an evaluation set for our task of multi-source simultaneous speech translation (SST) from the original and simultaneous interpreting. Then, we choose the languages for which we gather the data. Next, we consider the options, either using an existing dataset or creating a new one from an unprocessed resource. We create a new evaluation dataset and we name it ESIC – Europarliament Simultaneous Interpreting Corpus. ESIC is the first significant contribution of our research. We describe how we created ESIC and its two later extensions.

4.1 Requirements for an Evaluation Set

To evaluate a multi-source SST, we need an evaluation dataset. We have the following requirements and preferable features:

1. Authentic, i.e. recordings from a real multi-lingual event with simultaneous interpreting where multi-source SST could be used.

2. Long-form monologues (ref. Section 3.1.3), not single utterances. We need a recording of the whole speech, from the beginning to the end, not a set of isolated sentences. Authenticity is for us more important than cleanliness. We tolerate occasional deviations from monologues, e.g. short interruptions by another speaker.

3. Multi-parallel, in at least three languages – original speech in the original source language, parallel simultaneous interpreting in the second, alternative source language, and reference translation in the third, target language.

4. Audio recordings have to be available, to be able to realistically evaluate end-to-end, using real automatic speech recognition (ASR). Some corpora contain only the transcripts, which is not sufficient for our long-term plans.

5. Accessible – we want our results to be reproducible by other researchers, so we need data accessible by the public in the long term. This is achieved when the data are curated according to the FAIR principles[1] – findable, accessible, interoperable, and reusable.

6. Gold transcripts to evaluate the ASR quality – the word error rate (WER).

7. Reference translation for machine translation (MT) quality assessment using automatic MT metric.

8. Punctuation and casing. Some corpora do not contain punctuation and casing because it is easier to create transcripts without them. However, we assume that the authentic use case requires them in the target translation.

9. Word-level timestamps – for SST real-time simulation and measuring latency.

10. Sufficient size for statistically significant results. Standard MT test sets, e.g. WMT Newstest (Kocmi et al., 2023), typically have 3 000 sentences.

11. Preferable, but not critical features include:

    (a) Alignments of sources and target at the level of sentences (or parallel segments), or at the word level. It is useful for quality evaluation. It is not a critical feature because there are automatic tools we can use to create the alignments.

    (b) Validation and evaluation subsets.

    (c) Metadata, such as speakers' sociolinguistic characteristics, date, time, text summary of the speech, motivation, whether the speech is read or spontaneous, etc. All these details can be useful for a detailed evaluation analysis.

[1] `https://www.go-fair.org/fair-principles/`

(d) Marking “non-standard” phenomena, such as non-voice sounds (cough, laughter, applause, etc.), false starts, corrections, hesitations, or non-transcribable phenomena in the speech, such as foreign language, unrecorded segments from technical reasons, etc.

(e) Compatible with other standardized datasets. It is advisable to ensure that our evaluation data can be used by other researchers as much as possible. For example, they should not be included in any standard training data, otherwise the models trained on the standard training sets can not be evaluated on our test set. Usually, this is ensured by using unpublished and new data. The idea of “canary strings” of Big-Bench project (Srivastava et al., 2023) is also an interesting solution, but not much safer in practice.

(f) Representative and balanced – all the authentic language phenomena should be represented in the corpus. On the other hand, their frequency should be balanced, or easy to estimate. We assume that both goals can be reached by random sampling from authentic recordings, assuming that the future real data will not be remarkably different.

(g) Documented, extensible, reproducible, and correctable process of creating the corpus. In the future, the corpus could be useful in a way that we can not presume. It is advisable to enable easy further annotation or processing of the corpus, not requiring to repeat processes that were conducted during the initial corpus creation. For example, it is more advisable to publish a “noisy” version of the corpus with a filtering score together with an automatically cleaned version. A more advanced filtering can be applied in the future.

4.2 Choice of Languages

We will investigate language-independent methods, so we can be flexible in the choice of the primary languages in our experiments, however, we must be aware that frequent and continuous correctness checks are necessary when creating a system evaluation framework. In multi-lingual natural language processing (NLP), there are many opportunities to make a mistake that can be overseen by someone who does not have knowledge of the evaluated languages. We describe some examples from our experience from the ELITR project in Kocmi et al. (2021), Section 9.2. Since we can not have frequent consultations with foreign language speakers, and we build the evaluation framework by ourselves, we prefer to choose the languages that the main author speaks. They include Czech, English, and German. Fortunately, there are enough authentic SST use cases and available data in these languages. Therefore, we primarily focus our experiments on **English, German, and Czech**.

4.3 Existing Corpora as an Option

We surveyed and considered existing corpora for our evaluation, but none of them matched our requirements.

Interpreting corpora There are existing simultaneous interpreting corpora we considered reusing: EPTIC (Bernardini et al., 2016), EPIC (Sandrelli and Bendazzoli, 2006), and EPIC-Ghent (Defrancq, 2015) are small collections of transcribed interpretings from European Parliament created for analyses of interpreting. They contain only manual transcripts in selected languages, not including English, German and Czech. They do not contain timestamps and audios of interpreting, and their accessibility is restricted. The other corpora of simultaneous interpreting (Temnikova et al., 2017; Pan, 2019) focus on other languages. Therefore, we unfortunately can not easily reuse any existing interpreting corpus.

Too recent corpora Since we needed our evaluation corpus in the years 2020 and 2021, we created our own corpus ESIC (Macháček et al., 2021) and started using it. Several considerable corpora appeared later and we did not use them in our work, e.g. NAIST-SIC (Doi et al., 2021) for Japanese and English, Vox Populi (Wang et al., 2021) with all EU languages and data from the European Parliament, and ESIC UdS (Przybyl et al., 2022) containing English, German, and Spanish from the European Parliament. The last one (Przybyl et al., 2022) contains a nice summary of recent interpreting corpora.

AntreCorp Before working on this thesis, we were involved in creating an evaluation corpus of mock student business presentations that we unofficially name "AntreCorp," because we accessed the students thanks to collaboration with a company Antre, s.r.o. Originally, it is published as non-native English ASR evaluation corpus (Macháček et al., 2019). For IWSLT 2020 shared task (Ansari et al., 2020), Czech and German text translations[2] were added. English-to-German simultaneous

[2] https://github.com/ELITR/elitr-testset/tree/master/documents/iwslt2020-nonnative-slt/testset/antrecorp

interpreting and its transcripts were added at IWSLT 2022 (Anastasopoulos et al., 2022), unfortunately later than we needed. However, AntreCorp is a usable corpus for English-Czech SST using English-German simultaneous interpreting in multi-source, and we can recommend it for further work.

ST Corpora We also noticed and considered speech translation (ST) corpora, especially their evaluation subsets for our evaluation. There are e.g. CoVoST (Wang et al., 2020), MuST-C (Di Gangi et al., 2019) and Europarl-ST (Iranzo-Sánchez et al., 2020). However, they do not contain interpreting, and they are not usable for long-form speech evaluation. They were created for the training of contemporary ST models on speech, transcript, and translation triples. They contain individual sentences, and not document-level information (except MuST-C). In any case, these corpora do not include interpreting. Similarly, Vox Populi (Wang et al., 2021) and FLEURS (Conneau et al., 2023) contain only sentence-level data. Moreover, we could not use them because they are too recent, they were published after we created ESIC. However, FLEURS is multi-lingual multi-parallel evaluation corpus. We can use it in future work if the simplification on single utterance evaluation is meaningful.

4.4 Unprocessed Data Resources

There is no suitable corpus of speeches with simultaneous interpreting that meets our requirements for an evaluation set, but there are data resources that are potentially available for creating one.

European Parliament The most suitable resource for interpreting is the European Parliament that holds regular and long plenary sessions in all 24 EU languages including English, Czech and German with simultaneous interpreting into all EU languages. There was also a period (between 2008 and 2011) where both interpreting and text translations were published, which is a very useful resource for multi-sourcing. Furthermore, all the data are published on a web page, it is possible to download and process them into a comprehensive corpus. When we worked on this thesis in 2020, there were several corpora containing data from the European Parliament, namely Europarl for text-to-text MT (Koehn, 2005), Europarl-ST (Iranzo-Sánchez et al., 2020) for single source speech-to-text translation that does not include interpreting, and several small corpora for analysis of interpreting EPIC and EPTIC mentioned above in Section 4.3. Unfortunately, none of them was useful for multi-source SST. Therefore, we created one and we describe it in the next section.

However, we also considered the following data resources of simultaneous interpreting, but all of them are small, miss reference translation, or were available too late.

Interpreting school We started collaboration with the Institute of Translation Studies of the Faculty of Arts, Charles University and we were allowed to receive recordings from student mock interpreted conferences and seminars on simultaneous interpreting. We collected some data and our colleagues collected publishing authorization, but the corpus is not yet processed. The data are in 6 languages – Czech into English, German, French, Spanish and Russian, or from the 6 languages into Czech, or from any language through Czech into the other languages. There are often multiple parallel interpretations of the same speech into the same target language because there were multiple students who needed to practice at the same time. The mock conferences happen only several times in the summer semester. Their total duration is around three hours, there are seven audio tracks – main stage and six interpreting booths. Since there are so many language directions, the actual data size for each language pair is very limited, and therefore we have not processed them into a corpus yet. Similarly, we have some small unprocessed data from simultaneous interpreting seminars for some specific language pairs.

EUROSAI Congress Through the ELITR project, we have access to the congress of European Organisation of Supreme Audit Institutions (EUROSAI) that happened virtually in spring 2021, after rescheduling from fully on-site event in spring 2020 due to COVID-19 pandemic. However, there were only four to eight hours of meetings that were publishable, and the event happened too late to be usable in our research. The available data are mostly English speech with simultaneous interpreting into German, Spanish, French, and Russian.

Czech Radio conference Media and Ukraine In the summer of 2023, we received access to a one-day conference of the Czech Radio about media and Ukraine.[3] There was Czech, English and Ukrainian spoken, with simultaneous interpreting between these three languages. The data are again very small and they are yet to be processed.

[3] https://elitr.eu/subtitling-at-media-a-ukrajina/

Interpreting in TV For about three months in spring 2022, the daily news of the Czech Television were simultaneously interpreted from Czech into Ukrainian.[4] There are 111 episodes, each approximately 54 minutes long, which sums to 92.5 hours. These data could be available for research purposes, however, we have not used them because they became available too late, because they miss reference translations, and because they are not in our primary set of languages.

Last but not least, we mention the interpreting school, EUROSAI and Czech Radio conference because we partially contributed to data collection or processing. The data are planned to be further processed and released.

4.5 ESIC Creation

In this section, we describe how we created ESIC – European Parliament Simultaneous Interpreting Corpus, a manually transcribed corpus of authentic speeches from the European Parliament in English with simultaneous interpreting into Czech and German and with parallel text translations. The corpus consists of 10 hours, 370 speeches, and is suitable mostly for SST evaluation, but can serve many other purposes.

In this section, we largely extend the description of ESIC creation that we previously published in our paper "Lost in Interpreting: Speech Translation from Source or Interpreter?" (Macháček et al., 2021).

4.5.1 European Parliament as Data Resource

European Parliament is a very suitable resource of authentic multi-parallel data of speeches with parallel simultaneous interpreting, and also with parallel text transcripts and translations. We focused on the period 2008 to 2011 when both translations and interpretings were published.

[4] `https://www.ceskatelevize.cz/porady/14876111606-udalosti-tlumocene-do-ukrajinstiny/222411033280520/`

European Parliament is a legislative body of the European Union, which consisted of 27 member states in 2008-2011.[5] The Members of the European Parliament (MEP) are elected representatives of all the member states, with approximately proportional distribution to the states' population. There were around 750 MEPs in 2008-2011, elected every 5 years.[6] The MEPs have therefore very specific and diverse sociolinguistic backgrounds.

The Plenary Sessions of the European Parliament are regular meetings that happen once a month except August for four days in Strasbourg or Brussels.[7] The sessions in 2008-2011 were held simultaneously in 23 official EU languages,[8] and sometimes, a language of a candidate state or non-EU world languages such as Russian, Arabic, Chinese, Japanese, etc. are added.

The speakers at the plenary sessions are typically the MEPs, the president or vice-president who is chairing the session, commissioners, or other guests. A typical MEP's speech is 90 seconds to 2 minutes. The speaker can speak in any of the languages that are supported at the time. Usually, the speakers choose their native language, or English, which is the most common lingua franca in the European Parliament. Because of it, and because there were many MEPs from English-speaking countries (United Kingdom, Ireland, and partially Malta) in 2008-2011, English is the most frequent language in the European Parliament.[9] However, the English speeches in the European Parliament are often non-native and accented.

4.5.2 Speech Processing in European Parliament

All the official speeches in the European Parliament Plenary Sessions are recorded, transcribed, normalized, and translated, and then connected with metadata about the speaker (name, surname, ID, portrait photo, web link) and speech (language, agenda item, date, time, etc.) and sometimes with video. If the video is available, then it is with many parallel audio tracks, in the original language, and with simultaneous

[5] The list of EU member states was different than today in 2023. Croatia joined in 2013 and the United Kingdom left in 2020.

[6] The evolution of the number of MEPs by states over the years is in `https://en.wikipedia.org/wiki/Apportionment_in_the_European_Parliament`.

[7] According to `https://www.europarl.europa.eu/about-parliament/en/organisation-and-rules/how-plenary-works`.

[8] There are 24 EU languages today in 2023. Croatian was added in 2013. English is still an official EU language as it is an official language of Ireland, an EU member state.

[9] `https://www.theguardian.com/education/datablog/2014/may/21/european-parliament-english-language-official-debates-data`

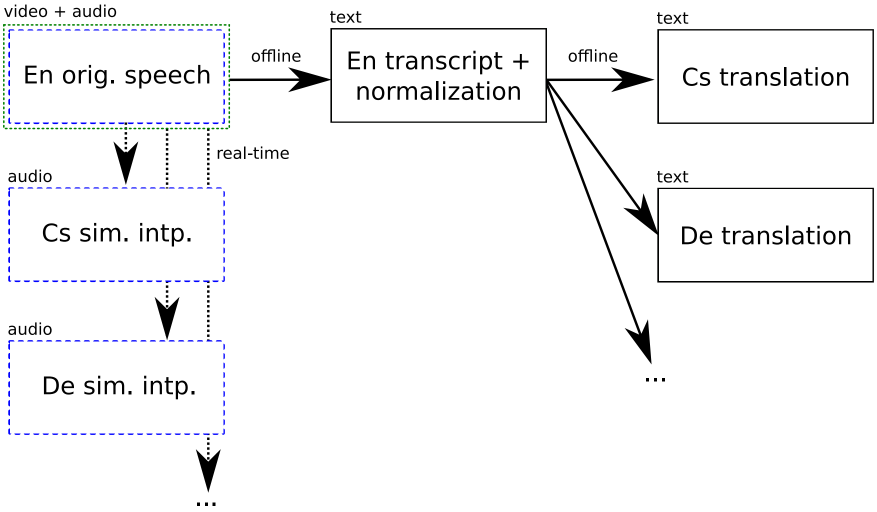


Figure 4.1: Scheme of data creation and processing at the European Parliament.

interpreting into all EU languages. The data are published on the web of European Parliament.[10] This processing is done by the staff of the European Parliament after the session. During the session, simultaneous interpreting is created. We illustrate the process in Figure 4.1.

The transcript and translation texts are available as "verbatim reports," and "minutes." Verbatim is actually normalized word-for-word transcript of what was spoken. Minutes are summaries or actions, such as the results of voting. We are interested in verbatim.

Normalization of verbatim transcripts is an adaptation of what was actually spoken during the live session for reading on the web. It makes better readability and coherence across all the speeches. The example is in Figure 4.2. The beginning salutations are changed to uniform phrase, and concluding "Thank you" is dropped. Disfluencies are removed, as well as side, organizational, and unintended comments. There are also grammatical and stylistic changes.

4.5.3 Downloading

We implemented an automatic tool to download data from the web page of the European Parliament. We did this work in 2020 when there was no other work we could reuse to download interpreting data from the European Parliament. We faced technical challenges, including:

1. Downloading and parsing texts.

2. Downloading videos was especially challenging.

[10] https://www.europarl.europa.eu/plenary/en/

orig	...you very much Mrs President, Mr Commissioner
norm	∅ Madam President, ∅
orig	The Stabilisation and Association Agreement with Serbia represents another important step in the process of integration towards the EU **[urder-]** undertaken by Serbia.
norm	the Stabilisation and Association Agreement with Serbia represents another important step in the process of integration towards the EU undertaken by Serbia.
orig	We are talking about a country whose progress in the recent years has already been impressive.
norm	We are talking about a country whose progress in ∅ recent years has already been impressive,
orig	And I think that further political and economic integration on the basis of the SAA will give the final boost to Serbia EU path.
norm	and I think that further political and economic integration on the basis of the SAA will give the final boost that Serbia needs on its path towards the EU.
orig	But considering the important role **[the Se-]** Serbia plays in the Western Balkans, the SAA will have a positive influence not only on the EU and Serbia as such.
norm	In view of the important role that Serbia plays in the Western Balkans, the SAA will have a positive influence not only on the EU and on Serbia ∅,
orig	But also on the whole region by facilitating its security, stability and development, as well as posing solid foundation for the enlargement process in the Western Balkans.
norm	but also on the region as a whole by enhancing its security, stability and development, as well as setting solid foundations for the enlargement process in the Western Balkans.
orig	I hope that, after it ∅ the green light of the European Parliament, the SAA agreement process can be concluded as soon as possible.
norm	I hope that, after it has received the green light from the European Parliament, the SAA process can be concluded as soon as possible.
orig	I would therefore ask the Member States to ensure that the ratification process can run in the smoothest and rapid manner.
norm	I would therefore ask the Member States to ensure that the ratification process takes place as smoothly and rapidly as possible.
orig	Thank you very much.
norm	∅

Figure 4.2: Example of the normalization ("norm") of the original speech ("orig") into the "verbatim transcripts" that are published on the web of the European Parliament. The changes are highlighted: deletions at the beginning and the end in red, stylistic alternations and grammar corrections in orange, and **[disfluencies]** in brackets and bold red. This example is a read speech by the MEP Boştinaru and is indexed in ESIC corpus as dev/20110118/005_031_EN_Boştinaru.

(a) The video links from the transcripts web page were often not available or not working. We found them on a different page of the web.

(b) We were interested in English speeches with Czech and German simultaneous interpreting. We soon discovered that downloading only the segments with English speeches was not possible because metadata that would indicate where English was spoken were inaccurate or missing. The only way was to download everything and then search and segment.

(c) The data we needed to download were very large. There were 4 years of our interest (2008-2011), 4-day meetings were held 11 times per year (monthly except August), for 8 hours per day (9 AM to 5 PM). It sums to 1 408 hours of recordings in each of three languages, original English, plus two interpreting tracks, Czech and German. Such massive download is challenging because of server load, network transmission, data storage, compression format, indexing and data organization, etc.

(d) The web structure that made the videos from 2008-2011 accessible was changing during our work.

  - First, there were links to video files in mp4 format that we could download easily, but not for all meetings.
  - Then, we found a web form where we could insert meeting date, time span, language, and our email, and after some waiting we received a temporary download link with the video. Since we needed nearly all the videos, we created an automatic tool that submitted thousands of requests for downloading, but we did not get any response on most of them.
  - In the end, we found a web streaming player and managed to download videos by an automatic tool that acted like a person watching videos through the browser.

(e) Limiting access and communication – after we submitted the requests for emailing the link for downloading videos, we were contacted by web administrator that large amounts of requests are not allowed. We apologized and kindly asked for support or enabling access to the data by other way, but we were replied that it was not possible. We were also unable to contact anyone at the European Parliament who could support us.

period	meeting days	transcripts	translations	SI videos
2008/01/14 – 2008/07/10	39	yes	yes	no
2008/09/01 – 2011/07/04	**164**	**yes**	**yes**	**yes**
2011/07/04 – 2012/12/13	78	yes	no	yes

Table 4.1: Availability of types of data from the European Parliament Plenary Session in different periods and number of meeting days we downloaded. Transcripts are in the original language of the speeches, and translations are parallel texts in all other EU languages, "SI videos" stands for videos with the original language and simultaneous interpreting into other EU languages. We further focus on the **bold-highlighted** middle period where all these data are available (3-times yes).

(f) Blocking access – although we attempted to reduce the download rate, we got blocked access to the European Parliament video web service because we accidentally sent requests over the limit. It happened because finding a working and not overloading way was difficult. Later, we implemented an option for downloading through parallel proxy servers which avoided blocking. We also reduced the downloading rate even more.

3. Metadata – the video streams were equipped with a sequence of metadata that contained speaker information, a time span of each speech, and a title and action item. Unfortunately, the timing information appeared to be unreliable for direct segmentation of the videos to individual speeches, but we could use the order of speakers and compare it to the order in text transcripts.

4.5.4 Usable Meetings

We downloaded texts and videos from all the days on which plenary session data were available between the years 2008 and 2012. Then, we found out on which dates the translations and videos with simultaneous interpreting were available. It is the period from 2008/09/01 to 2011/07/04, the summary is in Table 4.1. In total, we downloaded 164 meeting days, which is more than the regular amount we expected (4-day sessions every month except August). Maybe this number includes days and half days of meetings that were listed individually in the video streaming service, or there were some extra irregular meetings.

Language Statistics Table 4.2 summarizes the speeches by languages in the period in which both interpreting and translations were available. This summary contains the information from metadata, which can be partially incorrect, such as longer time spans for speeches, misspelled names, wrong language tags, etc.

language tag	language	speeches	speakers	duration	English words
BG	Bulgarian	283	25	7h 16m	70 220
CS	Czech	587	38	15h 55m	167 790
DA	Danish	353	22	8h 39m	99 585
DE	German	3 878	184	110h 44m	1 123 561
EL	Greek	941	56	24h 13m	218 222
EN	English	7 404	396	239h 50m	2 272 520
ES	Spanish	1 302	95	41h 7m	399 147
ET	Estonian	92	7	1h 51m	17 566
FI	Finnish	456	21	9h 33m	94 404
FR	French	2 674	193	100h 55m	1 026 783
GA	Irish	118	9	2h 8m	18 554
HU	Hungarian	841	48	22h 6m	211 309
IT	Italian	2 081	117	53h 39m	495 125
LT	Lithuanian	288	18	5h 43m	54 237
LV	Latvian	124	12	2h 54m	27 098
MT	Maltese	82	6	2h 38m	22 011
NL	Dutch	1 173	60	31h 50m	319 709
PL	Polish	1 787	93	41h 54m	403 246
PT	Portuguese	930	38	25h 52m	235 801
RO	Romanian	1 275	48	27h 31m	245 186
SK	Slovak	777	21	15h 47m	151 648
SL	Slovenian	221	13	5h 16m	50 426
SV	Swedish	576	37	17h 55m	191 228
XM	other/unspec.	242	101	13h 54m	145 367

Table 4.2: Statistics of downloaded and parsed speeches in period 2008/09/01 – 2011/07/04 by language tag in metadata, in alphabetical order. Language tag “XM” stands for other than EU or unspecified language. “English words” stands for the number of words in English translation (or transcript) of the speech.

4.5.5 Matching Text and Audio

We processed matching text transcripts and video segments including the speech. For that, we used the metadata that we downloaded with videos. We needed to resolve the following challenges:

1. Removing President. The chairperson (called "President," although he or she can be one of 14 vice presidents) is often recorded in video and included in metadata when he or she speaks to give floor to a speaker. However, this brief and very technical speech is not included in the normalized transcript. For simplicity in matching the other speakers, we excluded all occurrences of President's speeches from the data, no matter if he or she gives an actual informative speech, or not.

2. Misspellings in surnames. In video metadata, the surnames were sometimes misspelled, which made it a challenge to match them automatically with the transcripts. We created a list and simple rules for corrections. We also excluded some unmatched speeches.

3. Texts without speech. Sometimes, a transcript records a speech that was not uttered in the session. They are often marked as "in scribo," which means that the speaker was not giving the speech at the session, but delivered the text of the speech to the President to be included in the record. We excluded them.

4. Speech segmentation. We downloaded long videos (and corresponding audios with interpreting) of about 4 to 8 hours that contained all the speeches in a sequence. The individual speeches listed in metadata were supposed to be located in the long audio by time spans included in metadata. However, we discovered that the time spans were very inaccurate, starting too late or covering several minutes before and after, which included several more speeches. We needed to develop a more precise segmentation method. We considered multiple options:

    - Speaker diarization. We used the speaker-diarization tool LIUM_SpkDiarization[11] (Rouvier et al., 2013). It is supposed to label the time segments in audio when which speaker speaks. It works better when the number of speakers is known in advance, otherwise, it runs a clustering analysis to find it out automatically. We applied it to the interval marked in metadata extended by a margin. We selected the longest single-speaker

[11] https://projets-lium.univ-lemans.fr/spkdiarization/

segment as the speech to segment, assuming it is the whole speech without any interruptions. Unfortunately, this method is not totally precise. The assumption does not always hold, and the tool makes errors. Sometimes, a pause makes the tool mark two segments of one speaker as two speakers, sometimes an interruption by the chairpersons leads to an error, etc. However, we used this method as the most suitable in the end, before we had golden truth data.

- ASR – we created automatic transcripts for English, Czech and German audio tracks using ASR systems from the ELITR project (Cho et al., 2013; Povey et al., 2011; Kratochvíl et al., 2020). Unfortunately, we could not find any reliable automatic way to align them to the verbatim transcripts because the existing alignment tools (fast_align, Dyer et al., 2013; hunalign, Varga et al., 2005; etc.) expect parallel texts that have no prefix and suffix that has to be dropped. They also primarily work on clean bilingual texts, not on unpunctuated, uncased and inaccurate ASR transcripts that were available for us. These days, in 2023, we could consider much better performing Whisper ASR (Radford et al., 2022) and BertAlign (Liu and Zhu, 2022).
- Forced alignment – we attempted to automatically align the normalized verbatim transcripts to speech using the automatic forced alignment tool MAUS (Kisler et al., 2017). It works in two steps, first grapheme-to-phoneme conversion of the transcripts, and then searching for the optimal time alignment of phonemes to audio using an acoustic model, a component of a hybrid ASR system. Unfortunately, the normalized transcripts differed from the speech in salutation etc., and the tool was not reliable on not properly parallel transcript and audio that had some extra prefix and suffix.
- Constant offset to metadata – we could segment the speeches by the time spans indicated in metadata, with some constant offset that we would estimate from some subset of the data. It is a very simple, but inaccurate option.

In the end, we used the approximate speaker diarization, and manually revised the segmentation of the selected subset of speeches.

criterion	speeches
all matched speeches and texts	28 486
only English	7 404
with interpreting into Czech	7 030
with translations into Czech	4 452
filter out 1% of the longest	4 407
and 5% of the shortest – min 27s	4 187
filter out wrong and missing ASR	4 127
10 hours for ESIC	370

Table 4.3: Summary of selection of speeches into ESIC evaluation set.

4.5.6 Selection

After matching the texts with speech, we had 28 486 speeches in all languages. From this large set, we selected 370 originally English speeches, in total 10 hours, that we manually revised and included into an evaluation corpus ESIC.

Table 4.3 summarizes the selection process. There were 7 404 speeches (26%) in English, according to the label in metadata. Only 7 030 had simultaneous interpreting into Czech and German. We used language identification tool langid[12] (Lui and Baldwin, 2012) on the texts and filtered out those whose Czech translations were not detected as the Czech language, i.e. because it was labeled as Czech, but actually it was English original. Then, we excluded the very long and short outliers. We filtered out 1% of speeches with the longest duration and 5% of the shortest. The minimum length of the remaining ones is 27 seconds. Then, we processed ASR on the English, Czech, and German audio tracks, and excluded the speeches that were left unprocessed by ASR. The reason for ASR failure included wrong language in audio or technical issues with the ASR system. In any way, this criterion excluded only 60 speeches from the 4 187 examined, which is negligible.

Then, we decided to select 10 hours of speeches for the evaluation corpus ESIC, in order to have a feasible size for significant results, and not too big to avoid costly manual revisions. Since the same speeches were already in existing Europarl-ST corpus (Iranzo-Sánchez et al., 2020) that has training, evaluation, and development subsets, we wanted to enable safe use of ESIC evaluation on systems trained on Europarl-ST. The evaluation and development subsets of Europarl-ST consist of a set of speakers that are not included in the training set. It is a common and useful practice in ASR and ST to avoid overfitting to some subset of voices. The set of speakers is expected to be balanced and representative, e.g. include both male and female voices, native and non-native speakers, all age groups, etc. We therefore selected the same

[12] `https://github.com/saffsd/langid.py`

speakers that are in Europarl-ST evaluation and development sets. We found all their speeches in our collection and inserted them into the ESIC evaluation and development set (test and dev). Since they did not sum to 10 hours, we selected 28 additional speeches randomly, regardless of the speaker. These speeches can have an overlap with Europarl-ST training set. To make all potential users aware, we inserted them into a subset labeled as “dev2.”

4.5.7 Manual Revision

Annotators We hired and paid annotators to work on the manual revisions. We had several groups of annotators depending on the task. For revising segmentation, we hired lay speakers of English, German and Czech. For manual transcription of speech, we hired professional translators who have sufficient expertise with English, German, and Czech listening and writing and who offered suitable price range and availability. We selected them from approximately 30 candidates that we enquired.

Working with annotators We instructed the annotators through a shared document with very detailed instructions and guidelines. They were instructed to use the Git versioning system and to submit plain text files that they edited. We split the task into packages which they first claimed in a shared document, so that no files were edited by two persons at once, and then processed and submitted. When a new annotator joined, we revised their first revisions and gave them feedback, so that they could apply it to their next revisions.

Correcting speech segmentation First, our annotators manually revised and corrected the automatic segmentation into individual speeches in all three tracks (English source, Czech and German interpreting) because the automatic diarization was inaccurate at beginnings and ends. Furthermore, detecting beginnings and ends of the speech in simultaneous interpreting was not possible automatically because there is usually one coherent sequence of one interpreter speaking on behalf of many speakers in a row, so that the speaker turns or diarization do not help the segmentation. We considered several options, such as ASR followed by automatic alignment, but it would not be of sufficient quality. We therefore instructed our annotators to listen to the original and interpreting, and use language knowledge to mark the boundaries.

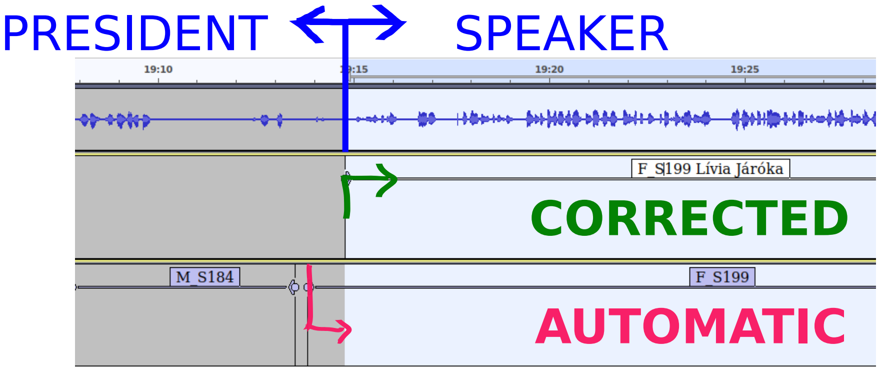


Figure 4.3: Preview of working interface for manual corrections of speech segmentation. There is an audio track at the top visualized as an oscillogram: the horizontal axis is time (left to right), and the sound intensity is displayed on the vertical axis (blue peaks and bulbs). The pauses and pause-delimited phrases can be observed visually. Under that, there is a label track where the annotator is supposed to correct the main speaker's time segment. The automatically detected sex ("F" as female), speaker id ("S199") from LIUM SpkDiarization tool, and name and surname ("Lívia Járóka") are labeled on the track. The annotators can use them as a clue, but primarily they should rely on the sound. The bottom track is another clue, it shows the automatically detected sex and speaker IDs. In the visible area, there is the President speaking on the left, followed by the main speaker.

We instructed the annotators to work with the Audacity[13] sound editing software. It enables listening and navigation in multiple parallel audio tracks, such as original and simultaneous interpreting, marking and labeling time segments. We gave the annotators an automatically created segmentation hypothesis for correction. Then they were instructed to listen to the track and shift the boundaries to correct positions. For that, they could use automatically generated clues that could make the task easier: diarization, and words of automatic transcripts attached to time segments when they were uttered. A preview of the working interface and clues is in Figure 4.3.

Transcribing interpreting In the next steps, our annotators manually transcribed the English-German and English-Czech interpreters following fixed annotation guidelines that we created. Our annotators marked false starts, unintelligible words, short insertions in different languages, and voice turns, e.g. swapping the interpreters, so that ESIC users can decide to handle them in a particular way. They transcribed and marked the segments which could not be easily transferred from

[13] `https://audacityteam.org`

	read		spont.		unmarked	
dev	141	79%	35	20%	3	1%
test	134	70%	54	28%	3	2%

Table 4.4: Number and percentage of speeches in ESIC subsets that were marked as mostly read, spontaneous, or not marked at all.

orthography to verbatim, e.g. the non-canonical forms of numerals, dates, loaned named entities, and acronyms. They inserted orthographic punctuation and spelling but did not make any changes in syntax, even when the interpreter's syntax could be considered as ungrammatical. Hesitations were not marked.

Transcribing original The transcripts of English sources were revised in the same way as the transcripts of interpreters' speeches, but the annotator re-used the transcripts from the web, which were manually revised and normalized by EP staff for comfortable reading, recall Section 4.5.2 and Figure 4.2. The annotator thus reverted the normalization back into the verbatim transcript to favour the match with the original speech rather than the grammatical or stylistic qualities.

Spontaneous or read Furthermore, our annotator marked, with the use of the video-recording, whether the speech was spontaneous, or read because we believe it has a big impact on the grammar, style, and complexity of translation. Table 4.4 gives a summary of the categories.

4.5.8 Post-Processing

Quality control After the annotators completed transcriptions, they were asked to randomly select a subset and revise it to correct typographical and grammatical errors, using primarily only the text. They used grammar correction tools for that. We repeated the process until we reached sufficient quality.

Then, we applied regular expressions and other similar methods to detect and correct frequent problems in the manually revised data, such as wrong brackets for the tag markers. We also needed to normalize digit transcriptions and sentence segmentation – one sentence per line.

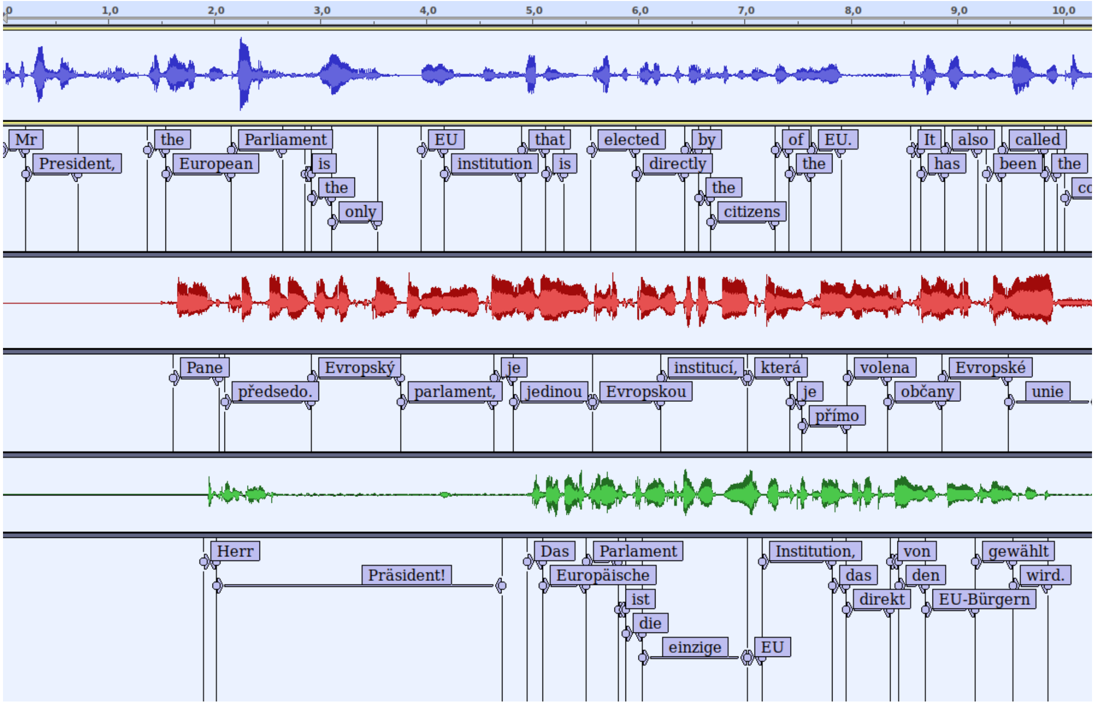


Figure 4.4: Preview of word-level timestamps attached to audio in ESIC. We can see the original English audio track at the top (depicted as a blue oscillogram), followed by words of the manual transcripts (in light violet rectangles) that are automatically aligned to the time segments (black vertical lines with round "knobs") when the word was uttered in the audio. The rectangles sometimes overlap the segments, but it is only a matter of visualization. In the middle and at the bottom, there are parallel audio and timestamped transcript tracks of Czech and German simultaneous interpreting. They start with several seconds of silence when the interpreters either wait for a meaningful translation unit or are occupied by the previous speech.

Timestamps Finally, we used MAUS forced aligner (Kisler et al., 2017) for English, German and Czech to insert word-level timestamps into the data. It works in two steps, grapheme-to-phoneme conversion, and phoneme-to-audio alignment. We dropped the phoneme-to-audio alignments because we do not presume we would need it. However, we can add them easily, upon request. Figure 4.4 displays the word timestamps attached to audio tracks.

Versions In sum, we ended up with the following preliminary working versions in each of the three languages – English, German, and Czech:

- Format of manual revisions ("man"), a human-readable plain text, one sentence per line, with special marks and tags.

beg.	end	verbatim	orthophic	sent.	tags
0.12	0.52	Thank	Thank	1	
0.52	0.60	you	you	1	
0.60	0.77	very	very	1	
0.77	1.01	much	much	1	
1.01	1.33	mister	Mr	1	
1.33	1.88	President	President,	1	
1.88	2.34	misses	Mrs	1	
2.34	3.01	Commissioner	Commissioner	1	
3.01	3.54	Kuneva	Kuneva.	1	
			…		
10.54	11.11	the eighth of	8	2	
			…		
18.19	18.19	<unk>	(??)	2	unclear=True,token_type=unknown_word
18.19	18.34	de	[de-]	2	is_corrected=True
18.46	19.19	devastation	devastation	2	

Figure 4.5: Illustration of the timestamped vertical rich format ("vert+ts"). The first two columns (from the left) are the beginning and end timestamps of the word on that line, in seconds. Word transcript follows, in the verbatim form in the third column followed by an orthographic transcript in the fourth. For many words, the verbatim form is identical to the orthographic without punctuation. The exceptions are abbreviations ("Mrs" – "misses"), numerals ("8" – "the eighth of"), canonically non-transcribable words (i.e. a declined numeral in Czech) and not transcribable words (i.e. unclearly pronounced). The fifth column is sentence ID, followed by a column for tags explaining non-transcribable words or types of disfluencies.

- Vertical rich format ("vert") which we generate from "man." It is an intermediate format for easy computer processing. Each word has assigned its orthographic and verbatim form (such as "2008" and "two thousand and eight"), sentence id, and special markers.

- Timestamped vertical rich format ("vert+ts"). It is "vert" with timestamps generated by automatic forced alignment. This format is illustrated in Figure 4.5

From the timestamped vertical rich format ("vert+ts"), we generate two alternative transcript versions that we propose for public use:

- "Verbatim," which does not include any punctuation, but does include false starts.

- Orthographic version ("Ortho") with punctuation and without false starts.

man	**<READ>**
man	Thank you very much **Mr** **{mister}** President, **Mrs** **{misses}** Commissioner Kuneva.
orto	Thank you very much **Mr** President, **Mrs** Commissioner Kuneva.
verb	thank you very much **mister** president **misses** commissioner kuneva
rev	**Mr President,**
man	First of all, please let me express my sincere sorrow for the huge tragedy that hit Taiwan on **8** **{the eighth of}** August and in particular for all the people that were killed by the incredible energy of the **(??)** **[de-]** devastation of this major disaster.
orto	First of all, please let me express my sincere sorrow for the huge tragedy that hit Taiwan on **8** August and in particular for all the people that were killed by the incredible energy of the devastation of this major disaster.
verb	first of all please let me express my sincere sorrow for the huge tragedy that hit taiwan on **the eighth of** august and in particular for all the people that were killed by the incredible energy of the **de** devastation of this major disaster
rev	**f**irst of all, please let me express my sincere sorrow for the huge tragedy that hit Taiwan on **8** August and in particular for all the people that were killed by the incredible energy of the devastation from this major disaster.

Figure 4.6: Illustration of four transcript versions of ESIC speech dev 20090917/008_003_EN_Boştinaru. The "man" version includes a tag with the type of speech (read) at the beginning, and some special markers. Orange highlights the expansion of abbreviations and not canonically transcribable digits, red are disfluencies. The "orto" version includes orthographical forms and excludes disfluencies, "verb" is verbatim form without punctuation and casing including disfluencies, and "rev" is the revised and normalized transcript. Black bold highlighting indicates the words where the versions differ (except casing and punctuation of "verb").

Furthermore, there is a version "Revised" which includes normalized translations (not interpreting), as downloaded from the web. The four text versions are illustrated in Figure 4.6. We encourage ESIC users to generate their own version from "vert+ts," e.g. verbatim with punctuation and casing and without disfluencies, etc., depending on their needs, e.g. for evaluating verbatim ASR that produces false starts, or disfluency removing ASR.

The size statistics of the three final versions and duration of the audio tracks are in Table 4.5.

		Source	**Interpreting into**	
		English	**German**	**Czech**
Dev	Revised	2019 44986	2015 42969	2019 37017
	Verbatim	179 47478	179 38956	179 33863
	Ortho	2772 45862	2818 38482	2736 33163
	Duration	5h8m38s	5h9m17s	5h10m30s
Test	Revised	1997 45068	1991 42347	1997 36600
	Verbatim	191 47331	191 39115	191 34464
	Ortho	2693 45640	2900 38738	2720 33747
	Duration	5h3m54s	5h2m23s	5h6m16s

Table 4.5: Size statistics of ESIC corpus. The two numbers in each cell are the number of sentences (or documents, in the row of Verbatim transcription), and number of words.

Edits versioning We have a Git repository that stores all edits of ESIC revisions, both the manual and automatically post-processed. If we make an error correction, e.g. a typographical change, we can simply edit a text file, and then reprocess the underlying automatic post-processing, such as digit normalization, removing disfluencies to clean orthographic versions, adding word-level timestamps, etc. We can therefore easily create a new publishable version of ESIC.

4.5.9 Publication

The following tasks were necessary to resolve before publishing ESIC dataset.

Authorisation Since we did not create the data in ESIC but we are reusing them from the public web page of the European Parliament, we need authorization to do so. We communicated with a responsible person at the language service department of the European Parliament (Directorate-General for Logistics and Interpretation for Conferences, DG LINC). We received authorization to repackage and publish the texts and audios of the original speakers and the text transcripts of interpreting. Since interpreters' voices are personal data, they can not be repackaged without permission of the persons who were recorded back in 2008-2011. It is practically impossible to reach them and ask for permission now (in 2020-2023) because we do not have their contact information. However, their voice recordings stay public on the European Parliament website. Therefore, we create and publish an automatic tool that every ESIC corpus user can use to download, segment, and save the interpreting audio locally on their computer.

Publication We published the corpus ESIC 1.0 as a package in a persistent repository LINDAT. It is available at `http://hdl.handle.net/11234/1-3719`.

We and our co-authors Matúš Žilinec and Ondřej Bojar also wrote and published a paper that describes the corpus composition and some analyses (Macháček et al., 2021 and Chapter 5). The paper titled "Lost in Interpreting: Speech Translation from Source or Interpreter?" was published and presented at the conference INTERSPEECH 2021.

4.6 ESIC Segment Alignments

We published the initial ESIC 1.0 version without a precise alignment of the parallel sentences because we initially did not need them. For our initial experiments and analyses that used ESIC in 2021, the document-level alignment was sufficient. However, later, in 2022 while working on the robustness analysis of multi-sourcing (Chapter 6), we needed reliable tri-parallel sentence alignments of the revised transcripts and translations. And even later, in 2023, we needed parallel alignment of English original and German interpreting.

We considered automatic alignment methods that we summarized in Section 4.6.1. However, in the end, we realize that they are not sufficient and not available, and it would be too ineffective to implement them on our own only to apply them once on a set of 5 465 sentences. Therefore, we manually aligned the revised texts (Section 4.6.2) and original and interpreting (Section 4.6.3). We release them as ESIC 1.1, a new extended version of ESIC.

4.6.1 Automatic Methods

There are automatic methods to align parallel sequences of segments, such as Gale-Church algorithm (Gale and Church, 1993). It was originally proposed to extract parallel sentence pairs (or M-to-N chunks, M sentences corresponding to N sentences) from translated texts that are segmented into sentences. Very roughly, the n-th sentence on one side may be a translation of the n-th sentence on the other side. However, this is not always true in translation, we must assume that succeeding sentences can be joined, long sentences can be split, and a sentence can be skipped. However, we assume no reordering, the order of sentences in the target stays the same as in the original.

We formulate *parallel segment alignment* as a task of multi-sequence labeling. The labels are assigned to segments in the sequences and must have a non-decreasing order in the sequences. The multi-subsequences with the same label, called "chunks," are parallel in meaning. The goal is to find the smallest chunks – it should not be possible to split them into smaller chunks that are still parallel in meaning.

Gale-Church algorithm is implemented e.g. in hunalign[14] (Varga et al., 2005). Hunalign uses dynamic programming and the assumption of similar lengths of parallel sentences. It also uses a dictionary of parallel words that are found by the tool itself, or inserted on the source.

We could not use hunalign for aligning three language variants, English, German and Czech revised texts in ESIC. We could apply it only to three pairs, English-German, English-Czech, and German-Czech.

We also tried hunalign on the transcripts of the original English and German simultaneous interpreting. However, we observed that the quality of the automatic alignment was very low. Interpreting is a very loose translation of the original, the similar length assumption is probably not working, and interpreting often joins multiple source sentences into one sentence in interpreting, often skipping some sentences. The beginning and concluding parts of the speeches are also very loosely included in interpreting, and it complicates the alignment.

These days, there exists BertAlign (Liu and Zhu, 2022) that is a considerable alternative to hunalign. We did not use it because it was not available when we needed it.

4.6.2 Tri-Parallel Texts

We aligned tri-parallel revised texts in ESIC, the English normalized transcripts, and their translation into Czech and German. We proceeded with the following process:

1. First, we excluded two documents out of 370 because they had German translations missing.

2. Then, we run hunalign on English-Czech and German-English. If the English side of the two pairs was identical, we created a triple. We also created an automatic script that found if two sentences span over one, and we aligned them automatically.

3. Then, we manually browsed the automatically pre-aligned triples with bare eyes and revised the ones where the length of one sentence did not match the others in the triple. We used our language knowledge of English, German, and Czech.

4. Then, we applied heuristics to find and fix full stops that were not followed by space because they often caused hunalign to make an error.

[14] https://github.com/danielvarga/hunalign

	sent.	doc.	En words	De words	Cs words
dev	2002	179	44866	43323	38347
test	1963	189	44273	42491	37695

Table 4.6: Size statistics of tri-parallel sentence-aligned “revised translations.” English is original, German and Czech are translations.

5. Then, we applied dual cross-entropy scores to each pair of sentences in the triples, and averaged the two scores (English-German and English-Czech). Then, we ordered the documents from the worst score to the best, and we gradually revised the alignments of documents in this order.

In total, manual revisions in the last step took us around two working days, which is surely more efficient than re-implementing hunalign for triples.

The size statistics are in Table 4.6.

Non-standard Slovakisms in Czech During the manual revisions, we found some imperfections in the Czech translations. We found two traces of the Slovak language in Czech that we consider as non-standard: *“ve Výbore” instead of “ve Výboru,” and *“po prvé... po druhé... po třetí” instead of ”za prvé... za druhé... za třetí,” which means “first,... second,... third” in enumerating reasons. Especially the second case does look more like a systematic error than a small mistype.

We have two possible explanations. First, it is possible that the Czech translation was not created directly from English, but from Slovak, and a mistake was overlooked. The Slovak-Czech language pair is much closer than the English-Czech, so the translator could prefer to work on the first as it is easier. The second possible reason is the possibility that the translator was less experienced in English-Czech translation, and was more used to the Slovak language than to Czech. Maybe the translator was a native Slovak-Czech bilingual and was not noticing it.

In any case, we keep these non-grammatical occurrences in the texts for authenticity. We still consider the whole ESIC as an authentic corpus that can serve as a reliable reference for translation evaluation.

4.6.3 Original and Interpreting

Later, in 2023 when we worked on Chapter 8, we needed segment alignment of the original English transcripts and of the transcripts of German simultaneous interpreting. We do not need to manually align the Czech reference side because there are automatic metrics that work reliably with automatic alignments, e.g. those provided by mWERSegmenter (Matusov et al., 2005), or we can concatenate the documents into a single sequence, as we analyze in Chapter 7.

We applied hunalign on the transcript of English original speech and German interpreting, but the results were of rather low quality, especially when interpreting was not a loose translation of the original. Therefore, we manually revised and corrected the sentence alignments using our knowledge of English and German.

4.6.4 ESIC 1.1 Release

We publish the manually revised sentence alignments as ESIC 1.1, a new version that extends ESIC 1.0 with the plain text files containing the two kinds of segment alignments. We released the new version in a persistent repository at the following link: `http://hdl.handle.net/11234/1-5415`.

4.7 Summary

In this chapter, we described the evaluation data for multi-source SST from original and simultaneous interpreting, and our choice of primarily focused languages, English, German and Czech. Then we created and published ESIC – a 10-hour evaluation corpus of speeches, simultaneous interpretings, and parallel translations from the European Parliament. It is our first significant contribution.

We created ESIC primarily for evaluation of latency and quality of multi-source SST and simultaneous ASR. However, ESIC contains very detailed manual annotations and metadata and can be therefore used for many other purposes. E.g. Purchartová (2023) uses ESIC for linguistic analysis of interpreting, Ryšlink (2022) for studying translation units in SST, and we analyze the latency, shortening, complexity, and comparison to translation in Chapter 5 of this thesis. ESIC can also be useful to many other tasks, e.g. analysis of disfluencies, read versus spontaneous speech, non-native English, analysis of English-German versus English-Czech interpreting, etc.

We also downloaded and pre-processed 28 thousands of speeches, approximately 800 hours, from the European Parliament in 23 EU languages with simultaneous interpreting. These data can be used in further research, e.g. in training of SST or even for the future task of automatic interpreting that goes beyond SST. They can also be used for the analysis of interpreting between many language directions that can be based on very big data. Although we downloaded data from 2008-2011, our downloading process can be applied also to much larger newer data, similarly to VoxPopuli corpus (Wang et al., 2021).

5
Interpreting Analysis

In this chapter, we analyze simultaneous interpreting (SI) to understand the challenges of SI that we will face when using SI in multi-source simultaneous speech translation (SST), but also the features of SI that could be considered beneficial.

We observe four challenges. The first is the fact that SI is not entirely parallel translation of the original. There are two main reasons: (i) the interpreting strategies that we describe in Section 5.1, and (ii) the issue of quality described in Section 5.2.

The second challenge is latency. SI is delayed by a varying time offset behind the original. In Section 5.3, we overview the latency of SI and analyze whether the delay caused by SI may be feasible in multi-source SST.

The next challenge we mention is the fact that SI consists of more tasks than just speech translation and we are not sure what is the optimum, whether speech translation that is usually more literal, or SI that tends to be brief and simpler, while occasionaly providing useful inter-cultural transfer. We analyze this question in more detail in Chapter 7.

The next challenge is that simultaneous interpreters often and on purpose segment their speech into sentences another way than it was in the original. It complicates aligning the parallel segments of the original and interpreter. We leave this challenge to future work, together with other potential challenges.

This chapter contains the original research findings of us and of our colleagues, and our survey of the literature. It is the second big contribution of this thesis. We summarize this chapter in Section 5.4.

5.1 Interpreting Strategies

There are two kinds of strategies that the interpreters tend to follow (Ešnerová, 2019). The first are *offline* strategies that include tasks that are performed before the interpreting event. The most typical is studying materials provided by organizers, but also anticipating and studying what could be said during the event, although not in the materials. Next is e.g. staying up to date with the source and target language culture, e.g. by the following news, etc.

The last part of offline strategies is ensuring that all the practical requirements are met, e.g. that the sitting chair has sufficient height so that the interpreter can see the speaker, making sure that the interpreter can collaborate with a partner and take breaks, etc. There are guidelines[1] and norms that help to set the minimum standards, e.g. the size of the interpreting booth. Remote interpreting, as it is a relatively new practice, is thoroughly described in guidelines by interpreting association.[2]

The *online* interpreting strategies are applied during interpreting. Their purpose is to cope with the practical constraints of time, memory, knowledge, and attention, and with the source quality. The strategies are described e.g. in Gile (1995); Jones (2002) and in textbooks for students of interpreting (Čeňková, 2008).

Economy The principle of language economy advises to prefer shorter synonyms and less complicated grammar constructions. Segmentation into simple sentences is preferred, to avoid long-range dependencies.

The means of language economy are e.g. shortening, simplification, generalization, and reducing redundancies. These are also four tasks that make SI not totally parallel to the original, and we should be aware of them when designing multi-source SST.

See examples of the economy strategies in Table 5.1 and Figure 5.1. We made an analysis on ESIC corpus and measure that SI is around 20% shorter than translation (Section 5.1.1) and uses a simpler vocabulary (Section 5.1.2). We also analyze information preserved in SI compared to direct translation (Section 5.2.3).

[1] https://aiic.org/site/world/about/profession/guidelines

[2] https://aiic.org/document/4418/AIIC%20Guidelines%20for%20Distance%20Interpreting%20(Version%201.0)%20-%20ENG.pdf

original	**generalization**	**reason**
cats and dogs	pets	short
a carp	a freshwater fish	when forgot the translation
Hallwang	some village	redundant proper name, inter-cultural transfer

Table 5.1: Examples of generalization of terms in SI reasoned by language economy means in SI.

Source (En)	Interpreting (En→Cs)	Gloss to Interpreting
And we try to compare the municipalities with the class of municipalities with the same size,	Zde máme srovnání obcí které mají srovnatelnou velikost.	Here we-have a-comparison of-municipalities, which have a-comparable size.
so we are not comparing Vienna to Hallwang, so we are trying to find similar municipalities **so em** so it will be a fair **compare,** comparison.	Nesrovnáváme tedy nějakou vesnici s Vídní kupříkladu, aby to bylo spravedlivé.	We-are-not-comparing thus some village with Vienna for-instance, so-that it was fair.

Figure 5.1: Example of interpreting where we observe interpreting strategies: segmentation to simple sentences, simpler syntax construction than in original ("compare … with class … of the same size" → "which have comparable size"), removing redundant repetitions and disfluencies (**red highlighted**), and inter-cultural transfer in case of blue higlighted word "Hallwang" that is also redundant as a proper name. It is also possible that the interpreter misheard or forgot this proper name as it is relatively infrequent and difficult to remember. This example origins in the presentation of the Austrian Supreme Audit Office representative and we reproduce it from Ondřej Bojar's keynote at WMT 2022.

Inter-cultural transfer An inter-cultural transfer is explaining concepts from the source language culture that the SI users may not know. For example, "November 1989" in Czech refers to the Velvet Revolution that ended the Communism era (Ešnerová, 2019). An inter-cultural transfer is a case when SI may be actually longer than the original. It may be combined with other strategies as in the "Hallwang" example in Table 5.1 and Figure 5.1.

Ear-voice span The interpreters are advised to keep an optimal *ear-voice span*, which means the delay behind the original speaker (Gile, 1995). It should not be too long because they could forget what was said, and not too short because the later words could disambiguate the source in other ways than what they translated. Optimal interpreting latency should vary depending on the content, complexity, current cognitive load, etc. We analyze latency in more detail in Section 5.3.

5.1.1 Shortening in ESIC

In Macháček et al. (2021),[3] we analyzed the properties of translation from the original English directly into Czech, versus translation from English-to-German SI into Czech. We performed this analysis on ESIC evaluation corpus using state-of-the-art Transformer neural machine translation (NMT) models that we cite in Macháček et al. (2021), Section 4. We use an English-Czech NMT model (EN→CS) that was trained for shortening, and default German-Czech model (DE→CS). We also compare the shortening of SI vs. reference translation.

Syllables are units independent of the orthography and phonemic inventory of the languages, and they are capable of expressing the shortening rate of translation into multiple languages. Therefore, we used grapheme-to-phoneme and syllabification tool (Reichel, 2014) for estimating the number of syllables in the original English source, and in Czech and German interpreting and translation. The results are in Table 5.2. We also demonstrate that German uses more characters per syllable than Czech, due to its smaller character inventory. This fact has to be considered especially in speech-to-text translation.

The results show that there is nearly no difference in the target length of interpreting, indirect DE-INT+DE→CS, and of our shortening model for direct speech translation (EN→CS). On average, one English syllable is translated into one Czech syllable. The revised text translation CS-REF is longer than the source, there is 1.19 syllables for 1 source syllable. The first reason might be that it is manually revised and adapted for reading. Shortening and simplification are not desirable in translation, while they are necessary in interpreting. The second possible reason is that interpreting might be unreliable. It may contain outages, and therefore be short.

[3] In this Section 5.1.1 and following Section 5.1.2, Section 5.2.3, and Section 5.3.2, we reproduce text and tables that we already published in our paper "Lost in Interpreting: Translation from Source or Interpreter?" (Macháček et al., 2021). Our colleague Matúš Žilinec contributed to this work with the shortening EN→CS MT model, with the study of vocabulary complexity (Section 5.1.2) which we helped to review and analyze, and with the evaluation interface in Section 5.2.3.

	Syllables	Characters
CS-REF	1.19 ± 0.12	0.93 ± 0.09
CS-INT	1.03 ± 0.17	0.80 ± 0.13
EN→CS (shortening)	1.03 ± 0.10	0.82 ± 0.04
DE-INT+DE→CS	1.01 ± 0.16	0.79 ± 0.12
DE-INT	1.01 ± 0.15	0.99 ± 0.14

Table 5.2: Length rate of source to target of ESIC test set. For example, CS-REF has 1.19 times more syllables than the English source. We report the average and standard deviation on all test documents.

	avg	±	std	words
CS-INT	6.15	±	2.83	32 992
DE→CS	6.16	±	2.85	32 703
CS-REF	6.32	±	2.93	37 182
EN→CS	6.42	±	2.89	32 488

Table 5.3: Mean and standard deviation of log word frequency ranks calculated from translations of the test set. The column "words" denotes the sample size (number of words in the translation). The proportion of out-of-vocabulary words is less than 0.5 % for each system.

5.1.2 Vocabulary Complexity in ESIC

Next, we compare the vocabulary complexity. We rank Czech words from the CzEng corpus (Bojar et al., 2016) by frequencies, such that the most common word has rank 1, and the least common word has the rank of the number of unique words. The "comma" and "full stop" characters were removed before the evaluation. Table 5.3 shows the mean and standard deviation of log ranks for each system across the documents in the test set.

We test whether the mean log rank of EN→CS is statistically equal to that of DE→CS. Using the two-sample Z-test, we reject this hypothesis with $p < 0.01$. Thus, we conclude that the translations EN→CS (machine) and CS-REF (human), which do not contain any interpreter component, use a more complex vocabulary than both setups involving an interpreter, CS-INT and DE→CS.

5.2 Quality of Interpreting

The second challenging reason why SI in multi-source SST can not be totally parallel to the original, is the issue of quality. Let us first focus on the reasons for low quality, and then on the question of how to define and detect the quality. Last, but not least, we present our study on content preservation of SI on ESIC.

5.2.1 Difficulties in SI

SI may not be completely reliable, it is "the art of possible" (Olsen, 2020; Ešnerová, 2019). The impossible or difficult situations in SI are:

Source quality Fast speech, noise, suboptimal sound conditions, non-standard accent or pronunciation, disfluencies or ungrammatical disfluent speech, etc. are difficult to understand, and therefore also difficult to interpret. The suboptimal conditions may lead to more stress, faster exhaustion, and errors. In the most difficult situations, the interpreters may abort a service, or focus more on decoding, relying mostly on presumptions. The coping strategies include preparation in advance, e.g. making sure that the practical setup meets the standards, studying the background materials to be able to presume the speech, or training for the speaker's accent.

Language Rare words (especially specific terminology) that the interpreter does not know or does not remember in time are difficult in SI. The coping strategies involve e.g. using a supernym or describing the term in other words.

Language direction Language direction also impacts the difficulty of interpreting. In general, the closer and more similar language pair, the easier it is for SI. The difference in typical word orders also impacts difficulty. The example is SI from English to Japanese, from SVO to SOV language. The verb in English is available before the object, so it has to be memorized, then an object has to be produced in Japanese, and then the verb must be placed at the end of the sentence in Japanese. He et al. (2016) describe that a common strategy to cope with the word order difference is to use grammar construction that eliminates it, such as passivization.

Read speech is typically more difficult to interpret than spontaneous. The interpreters are supposed to use a copy of the text during interpreting. If the copy is unavailable, it is difficult because written speech tends to have more complex language than spontaneous because the writer had a chance to invest more time and cognitive capacity in producing it. Furthermore, reading can be very fast.

Memory Numbers, acronyms, and long enumerations are usually difficult to memorize and interpret. Interpreters usually take notes while interpreting, and rely on the booth partner who writes them down. A possible strategy is also to elide items in enumeration, e.g. generalize them as "and others." However, even without the difficult content, the working memory of the interpreter has a limited capacity, and running out of the capacity may lead to errors. E.g. Daró (1997); Gabzdilová (2008) and others describe studies of memory in SI.

Non-standard situations Non-standard situations, e.g. with the practical setup, unavailable or wrong preparation, or non-standard language such as puns, mocking accent, poetry, etc. are difficult. The interpreters may resolve them by improvisation (Ešnerová, 2019).

Experience SI is a highly demanding and complex task that requires specific skills and experience. The most important, but not the only relevant skill is the source and target language proficiency. However, familiarity with the topic and vocabulary is also very important. Expert interpreters usually perform better than novices (Gieshoff, 2021). Several works explain it by e.g. memory capacity (Christoffels et al., 2006), cognitive flexibility (Yudes et al., 2011), etc.

Exhaustion The difficult situations may happen together or in a short time and lead to congestion of the mental capacity, which may lead to more errors. But even in standard situations, SI is a mental process that is prone to errors. Olsen (2020) presents Gile's effort model (Gile, 1995). There are mental processes of SI: listening, analysis, memory, production, and coordination. All these processes require effort, and the total effort capacity is limited. Exhausting the capacity leads to errors.

5.2.2 Definition of Quality

Ďoubalová (2020) summarizes the question of quality in SI. There is no comprehensive and exhaustive definition of quality. For example, the quality objective may be "the person receiving SI should have the same experience with participating in the communication as a person without any language barrier that receives the message in the original language."

The drawback of this definition is that it is practically impossible to meet it completely, and it does not answer how to compare two candidates that fulfill it partially. We therefore use the standard approach from SST research, which is reusing the standard methods for text-to-text machine translation (MT) quality assessment, or adapting them to SST. The quality in MT (e.g. in WMT tasks, Kocmi et al., 2022; Freitag et al., 2022) is judged by reliable human experts.

Another way of quality estimation is to define the intention why humans participate in communication and measure the success rate in meeting their intention while being assisted by the candidate system. TC-STAR evaluation report from 2007[4] compares SI to speech translation (ST) on comprehension questionnaires. Our Continuous Rating (Javorský et al., 2022) is an analogous and simplified method. See more in Chapter 7.

5.2.3 Content Preservation Study on ESIC

In Macháček et al. (2021), we performed a study to compare the difference in text simplification between machine translation and a human interpreter. We manually assess the amount of information from the source text preserved in the translation.

We employed two human annotators. They are both non-experts in the European Parliament debates, non-native speakers of English, and native speakers of Czech. The first one, a professional translator, worked 5 hours and annotated 107 sentences. The second one, a computational linguist, contributed 20 sentences (1 hour).

The annotators were provided with English revised transcripts of the whole document, and with six translation candidates. Four of them used the gold trancripts as the source, serving as the upper bound for speech translation quality. There was MT from the English original (EN src transc.+EN→CS) and from the German interpreting (DE-INT trans.+DE→CS), the gold transcript of Czech interpreting (CS-INT) and Czech reference translation (CS-REF). We contrast them with MT following authentic ASR transcripts of the English original (EN ASR+EN→CS) and of the German interpreting (DE ASR+DE→CS).

The candidate translations were all blinded and in random order. One random sentence from the source document was highlighted. The annotators were asked to express to what extent the information from the highlighted source sentence was preserved in the translation candidates, on a scale from 0 to 100. For comparability, they were asked to rate all 6 candidates at once.

[4] http://tcstar.org/documents/D30.pdf

System	avg ± std	avg ± std
CS-REF	0.77 ± 0.32	0.86 ± 0.11
EN src trans.+EN→CS	0.70 ± 0.33	0.89 ± 0.10
DE-INT trans.+DE→CS	0.49 ± 0.37	0.60 ± 0.29
CS-INT	0.47 ± 0.39	0.77 ± 0.20
EN ASR+EN→CS	0.38 ± 0.36	0.58 ± 0.28
DE ASR+DE→CS	0.19 ± 0.29	0.37 ± 0.27
Annotator	107 sent., 5h	20 sent., 1h

Table 5.4: Manual assessment of information preserved.

Table 5.4 indicates that EN→CS applied to the golden transcript preserves a similar amount of information as the manual translation. Involving any interpreter (DE→CS and CS-INT) leads to a considerable loss. ASR as the source for MT instead of gold transcripts significantly reduces translation quality and loses further information (EN ASR+EN→CS and DE ASR+DE→CS).

The aggregated scores of the two annotators are consistent. The second annotator (the computational linguist) reports that in many cases, the differences in non-ASR-based translations were subtle and probably unimportant for the intended audience at the live event. For example, there was a substitution of "president's office" and "the president," as a subject in the sentence, and such cases were penalized slightly. In some cases, the translation of the highlighted sentence could not be found in the target, probably due to interpreter overload, and was largely penalized. It explains the low scores of the interpreting-based systems. Future evaluations could be provided by domain experts capable of considering the importance factors of particular facts. Also, the frequency of interpreting outages can be estimated by a targeted evaluation.

Our evaluation process has limitations, e.g. the source being presented to the annotators only as English text, without audiovisual information. The gender of the speaker and addressed persons was thus often unclear, and its translation could not be evaluated. The interpreters use correct and consistent gender markers, while machine translation from English does not.

5.3 Latency of Simultaneous Interpreting

The latency of SI (also called ear-voice span in SI literature) is the time offset between the original speech and the parallel speech of the simultaneous interpreter. The lower bounds of latency are the practical and physical constraints. The first reason for latency is waiting for a translatable source segment. The next reason is that the mental and physical process of simultaneous interpreting (listening, analysis, memory, production and coordination, recall Gile, 1995) takes time. The last, but not least, is the occupation by interpreting the previous source segments.

Sometimes there is no possibility to interpret directly from the source language into target, but a *relay* through intermediate SI can be used. Waiting for the intermediate interpreting is another source of latency.

The upper bounds for the latency of SI are the fact that the listener is supposed to be engaged in communication and can tolerate average latency 4-5 seconds, according to studies mentioned by Sridhar et al. (2013). The next reason for the upper bound is that the interpreters are aware of the limited capacity of their short-term memory, and they choose a coping strategy that avoids keeping terms in memory for a long time (He et al., 2016). Empirical studies (Sridhar et al., 2013) show that more than 10-15 seconds latency is prone to errors due to memory capacity.

For our research of multi-source SST from the original and SI, we investigate whether the latency of such a system is feasible, and does not necessarily limit the overall usefulness of SST due to high latency. Therefore, we performed SI latency analysis on interpreting data in two steps. First in literature and on the large untranscribed and manually unprocessed data from the European Parliament (Section 5.3.1), and then on ESIC corpus (Section 5.3.2).

5.3.1 Initial Analysis

We learned the basic facts about latency from interpreting literature (Čeňková, 2008). The latency is not constant during interpreting, it varies during the speech depending on the language pair, interpreting strategy, complexity of the speech and language, etc. We also learned, e.g. from Lederer (1978), that the average latency is usually between 3 and 6 seconds. However, there is no reproducible background information about it. Therefore, we aimed to measure it on the real SI data.

In our initial analysis, we used the data that we downloaded from the European Parliament and described in Chapter 4. We used the 4 127 originally English speeches that had SI into Czech. There was 131 hours in total; recall the selection process in Section 4.5.6. We segmented the individual speeches approximately and automatically, without any rigorous measure of the error rate. It may happen that the beginning or end of the speech is wrongly segmented in the original, interpreting, or in both, and therefore not parallel. However, these errors are supposed to be smoothed by averaging over the large data. In any way, we matched the audio of the English source and Czech SI. Then we created automatic transcripts using automatic speech recognition (ASR) systems and we got word-level timing information using forced alignment of the automatic transcripts. For many speeches, we also had Czech translations and normalized English transcripts, but we did not use them. We did not measure the quality of the automatic transcripts or the word-level timing because we did not have any evaluation set at that time; the gold transcripts in ESIC were created later. The normalized transcripts that were available did not match the speeches word for word, so the scores would not be useful.

Latency Definition

A definition of SI latency is needed to measure it rigorously. We were searching for the exact definition in interpreting literature, e.g. in Čeňková (1988), but in the end, we created our own definition based on our understanding of latency. The definition follows.

Let us have two parallel audio streams, the original source S and its parallel SI I. The streams are encoded as sequences of segments, $S = \{s_1, \dots, s_n\}$ and $I = \{i_1, \dots, i_m\}$. The segments are words of transcripts that are produced at time $t(s)$, where t is a function that indicates time. Since S and I are in two different languages, there is a function $C : S \rightarrow I \cup \{\epsilon\}$ that maps corresponding segments in S and I, the ones that have parallel meanings. If $C(s) = \epsilon$, it indicates that there is no parallel segment in the interpreting stream.

The *interpreting latency* L of segment $s \in S$ for which $C(s) \neq \epsilon$ is $L(s) = t(C(s)) - t(s)$. As *average latency* of SI applied to individual speech, we mean an average of segment latencies. For a set of speeches, we usually report an average of speech averages, which assumes that each speech has an equal importance.

There is a natural expectation that latency is non-negative because the interpreters usually interpret after they hear the source. However, occasionally they can predict what is going to be said, which may result in negative latency.

Remark on automatic alignment In practical setups, it is typical to use automatic word alignment methods to detect parallel words. However, they are not totally accurate. An error in alignment can result in negative or extremely large positive latency. Therefore, the alignment error rate should ideally be reported together with latency, or at least with a warning that the result is an approximation with an unknown error.

Aligning Original and SI

For latency measure, it is necessary to align the original speech and SI at least on the level of phrases or sentences. Our dataset consists of 131 hours of 4 127 English speeches with SI into Czech. The challenge is that we have only approximate and inaccurate segmentation of the speeches, unpunctuated ASR transcripts with rather low, but not measured quality, and automatic word-level timestamps of the transcripts. We also have normalized English transcripts and Czech translations for some speeches.

We considered several approaches. Parallel audio sequence alignment could be possible if we implemented a new tool for that, e.g. by a combination of the word alignment (Tiedemann, 2000; Tamura et al., 2014; Li, 2022) and bioinformatics algorithms for parallel sequence alignment (Chen et al., 2006), for which we would need to resolve the issue that the sound sequences do not consist of a small set of characters. The character set could be created by quantization, or by automatic transcription to phonemes. Unfortunately, it would require lots of development work. We could not find any language-independent phonemic ASR model in 2021 when we worked on it. Zhao et al. (2021a) propose a dynamic programming algorithm for sentence alignment of the punctuated transcripts of the original and SI, however, it was available too late for us, they do not publish their implementation and our data are not punctuated.

The ideal, but unfortunately not possible approach back in 2021 was to automatically transcribe the sources with punctuation, and then use sentence alignment tool hunalign[5] (Varga et al., 2005) enhanced by matching parallel words. Unfortunately, the ASR achieved relatively low quality, and we did not have a punctuation tool that would be sufficiently reliable.

In the end, we used a standard word alignment tool fast_align (Dyer et al., 2013) on the whole unpunctuated ASR transcript sequences, each consisting of one full speech. We selected fast_align and not any neural model because they are typically trained on single punctuated sentences and clean texts, which is not our case.

Visualization

Figure 5.2 shows a preview and illustration of the alignment of the beginning of one document in the Audacity program. We highlight time segments that are delimited by aligned words, e.g. “one” in English, “jedna” in Czech. The words in the middle of the segments are not aligned. We can also observe missing punctuation and spelling errors due to low ASR quality, e.g. “well com” instead of “welcome.”

[5] `https://github.com/danielvarga/hunalign`

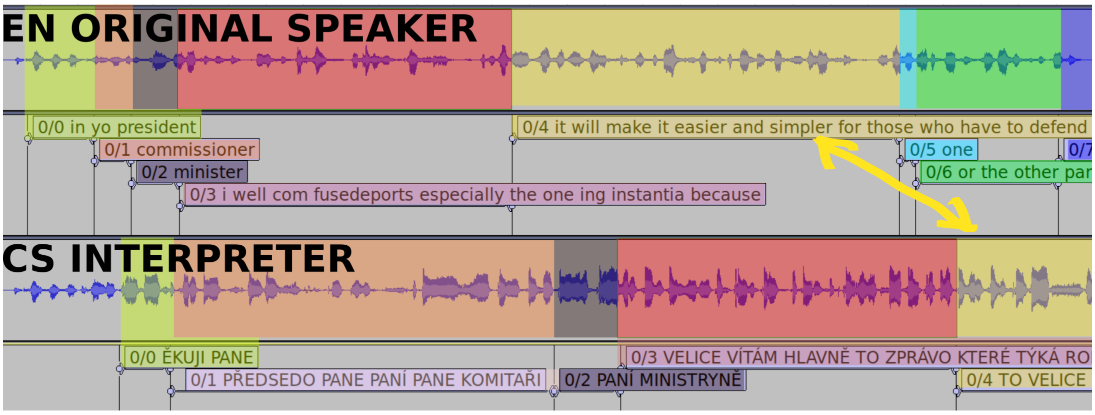

EN ORIGINAL SPEAKER
0/0 in yo president
0/1 commissioner
0/2 minister
0/3 i well com fusedeports especially the one ing instantia because
0/4 it will make it easier and simpler for those who have to defend
0/5 one
0/7
0/6 or the other par
CS INTERPRETER
0/0 ĚKUJI PANE
0/1 PŘEDSEDO PANE PANÍ PANE KOMITAŘI
0/2 PANÍ MINISTRYNĚ
0/3 VELICE VÍTÁM HLAVNĚ TO ZPRÁVO KTERÉ TÝKÁ RO
0/4 TO VELICE

Figure 5.2: Preview of alignment between the unpunctuated automatic transcripts of the English original and Czech SI. The horizontal axis signifies time. Aligned audio segments are highlighted by matching color. Automatic transcripts of the audio segments are placed under them in this preview, highlighted by matching color and prefixed by the same numbers in both tracks.

In Figure 5.3, we show latency progress during simultaneous interpreting session on several example speeches from our collection. We observe intervals with increasing and decreasing latency between 1 and 8 seconds, with one exception of latency of up to 16 seconds at the end of the speech, which could be caused by wrong detection of the speech end. We also observe the frequency and distribution of automatically aligned words between the unaligned words, as orange and blue data points.

Discussion

In summary, in this analysis, we got familiar with the practical aspects of the latency measure. We used our findings in our other analysis on the ESIC corpus.

Furthermore, we attempted to explain the reasons for peaks in latency within the individual interpreting sessions, but we did not observe any regularity. Then, we considered extracting the minimal translation units that the interpreters use to inspire segmentation for SST, but we realized that the exact unit boundaries are latent, not easy to determine exactly. In fact, we can detect the upper bounds of the translation units, the minimal ones appended by the source part that was emitted while the interpreter was occupied by producing the earlier units. However, the translation units of SI on ESIC corpus for inspiration of SST were later studied by Ryšlink (2022), and similarly by Liu et al. (2023) on English and Chinese.

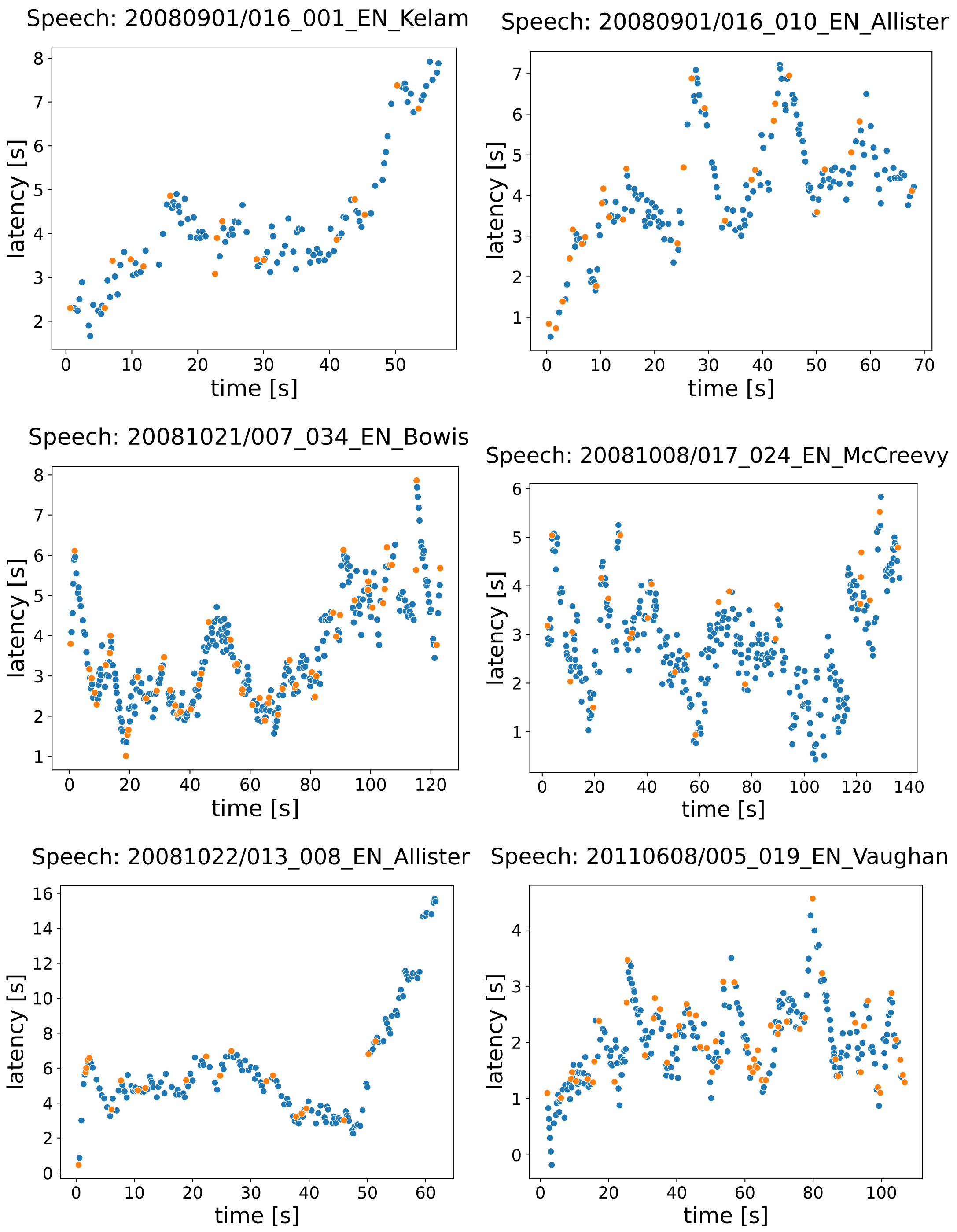


Figure 5.3: Progress of latency during example interpreting sessions. The data points are source words uttered at the time on the horizontal axis. The vertical axis indicates with what latency (how many seconds later) was a corresponding word uttered in interpreting. The orange data points indicate words whose corresponding word in interpreting was found by automatic word alignment. The blue points are aligned by length proportion. Note the different y-axis ranges.

Next, we attempted to analyze the relation between source speech pace and interpreting latency, to be able to find the translation units at least on the very slow speeches where interpreters could have comfortable speaking time. We did not complete this analysis as it is not our main focus, but we hypothesize that human short-term memory might have a “volume,” a capacity restricted by time durability and content size.

5.3.2 Latency Measured on ESIC

In our second analysis of latency, we could use the gold transcripts in the ESIC corpus on 10 hours of 370 speeches from our collection from European Parliament. The main reason for this analysis was to analyze, whether the latency caused by SI is feasible in SST. We therefore compare the latency of direct English-Czech SST, English-Czech SI, and English-German SI followed by German-Czech SST. We used two example SST systems that were among state of the art in 2021, as an upper bound of latency. Since our results are encouraging for SI in these SST systems from the latency point of view, we assume that it will hold for newer and faster systems as well.

Re-Translating MT

The MT systems that we use in this analysis are already mentioned in Section 5.1.1, and also described in Macháček et al. (2021). We use one English-Czech (EN→CS) NMT model trained to produce 20% shorter translations, and one German-Czech (DE→CS) model.

We use them in *re-translating* real-time mode in cascade with re-translating ASR system for English or German that were originally prepared for lectures (Cho et al., 2013). They emit partial hypotheses in real time and correct them as more context is available. The German ASR model (DE ASR) is a hybrid HMM-DNN model. The same system was also used by KIT Lecture Translator (Müller et al., 2016). English is neural sequence-to-sequence ASR (Nguyen et al., 2021a). They are connected in a cascade with a tool for removing disfluencies and inserting punctuation (Cho et al., 2012) and with the MT systems.

The re-translating MT systems receive updates from ASR segmented to individual punctuated sentences. MT Wrapper, our tool that wraps offline MT model and runs it in re-translating real-time mode, skips the updates that became outdated, and replaced by newer ones during the time when MT was blocked by processing previous updates. Then, it may happen that an update reverts a sentence version that was already translated. Therefore, a cache is consulted and the translation is retrieved immediately, or processed by a NMT model and saved in the cache.

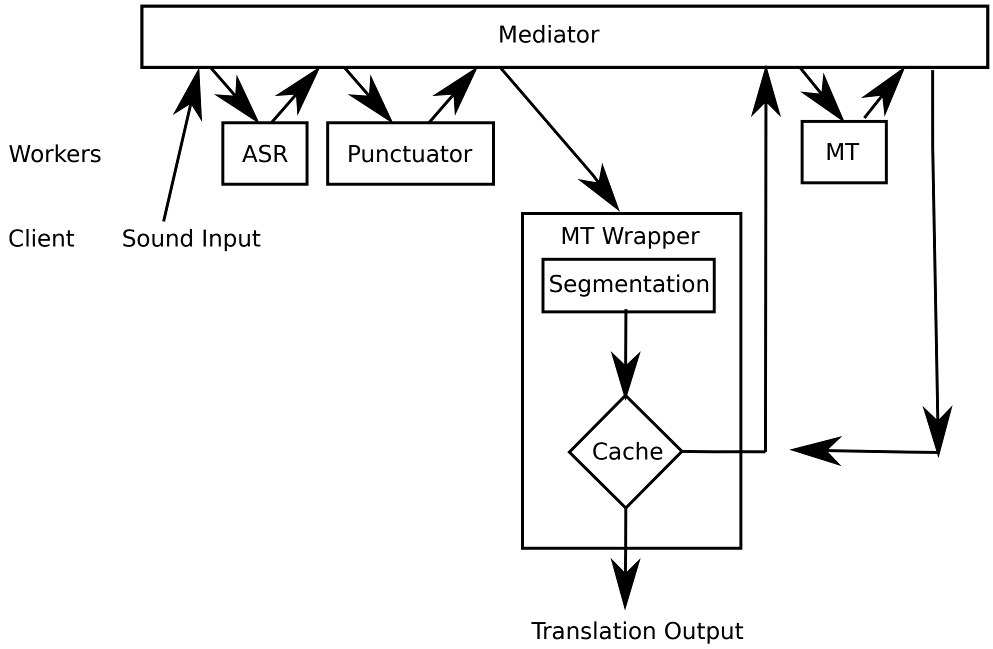


Figure 5.4: Scheme of re-translating cascade for real-time speech processing from ELITR project for IWSLT 2020 shared task (Macháček et al., 2020) that we use in our analysis. The mediator is a server that connects workers (distributed system components) through Internet network protocol. MT Wrapper is a software component at the client's side that resolves segmentation and caching. Figure reproduced from our slide presentation.

The cascade is the same that we used and evaluated in the ELITR project at IWSLT 2020 shared task (Macháček et al., 2020). Figures 5.4 and 5.5 illustrate the cascade and the intermediate tasks.

Production Time

To measure latency, first, we need to assess the time when each word in source, interpreting, and machine translation was produced. For the gold transcripts of source and interpreting in ESIC, we have word-based timestamps from the forced alignment tool. For the re-translating machine translation, we use the finalization time of a target word as in Arivazhagan et al. (2020a). It is the first time when the system produces the word, and the word and all its preceding words remain unchanged until the end of the session. This definition is rather harsh because it penalizes subtle, cosmetic changes in translation output the same way as meaning-altering re-translations. It is possible that a real user reads the translation earlier than at finalization time, and does not notice short flicker in previous words. However, the finalization time is an upper bound for the word production time.

	time	from	to	
ASR	1748	330	1040	miss
	1950	330	1390	mr press
	2181	330	1760	mr president
	3127	330	2480	mr president <human> for
Normalization	1853	330	1040	Miss...
	2055	330	1390	Mr. Press,...
	2298	330	1760	Mr. President...
	3307	330	2480	Mr. President for...
Segmentation	1854	100	101	Miss...
	2056	100	101	Mr. Press,...
	2298	100	101	Mr. President...
	3307	100	101	Mr. President for...
MT	3300	100	101	Miss...
	6000	100	101	Herr Präsident für...

Figure 5.5: Illustration of updates and processing between cascade components of the ELITR live speech translation pipeline from IWSLT 2020 (Macháček et al., 2020). The column "time" indicates time in milliseconds from the beginning of processing at which the row went through a component. The columns "from" and "to" are timestamps that are associated with the row. They correspond to the time segment in the source in ASR and Normalization. In Segmentation, they are replaced by a sentence index. Figure reproduced from our slide presentation (Macháček et al., 2020).

The latency is the difference of times of the source word and its corresponding word in the target (recall definition in Section 5.3.1). We assess the correspondence with automatic word alignment.

Alignments

As we figured out in our initial analysis, fast_align (Dyer et al., 2013) tool is sufficient to align the transcripts or translation. We used this tool after tokenizing (Koehn et al., 2007) and trimming the tokens to 5 characters as a trivial form of lemmatization. We processed all 370 ESIC documents, treating each as a single sequence. We added relevant sentence-aligned texts to fast_align training data, to expand the vocabulary: revised translations of Europarl (around 4 thousand documents from the same period) for interpreting, and the source and target sentence prefixes for machine translation. We obtained forward and backward alignments, and removed those going back in time, assuming that the interpreters do not risk predicting content. Finally, we intersected them. Table 5.5 shows that around 40% of source words were aligned to SI (CS-INT, DE-INT, and DE-INT followed by DE→CS MT), and 50% were aligned to the direct MT.

Based on a small manual check, the resulting word alignments were reasonably good, despite that fast_align is designed for individual sentences and our documents were much longer.

Figure 5.6 shows a preview of alignment on one example speech. In this example, we observe that the alignment is rather meaningful, matching parallel Czech and English words such as "bereme" and "taking," which is correct. The alignments tend to follow a line, a diagonal of a rectangle that is wider than taller because a parallel message in English consists of more words than in Czech, which is characteristics of the language pair and of SI. In this visualization, we see only the top left part of the rectangle, therefore the bottom part in Czech is not aligned in the visible area.

Furthermore, we observe in Figure 5.6 that there are not many alignments far from the line, which indicates that the alignment may be accurate. We also do not observe any long continuous segments of unaligned words which indicate that the Czech SI may be a relatively literal translation. Based on a small visual check, this is a case of many documents, but not of all. A rigorous expert evaluation is pending, however, system comparison is possible without it, with the assumption that the approximation error is uniform among the systems.

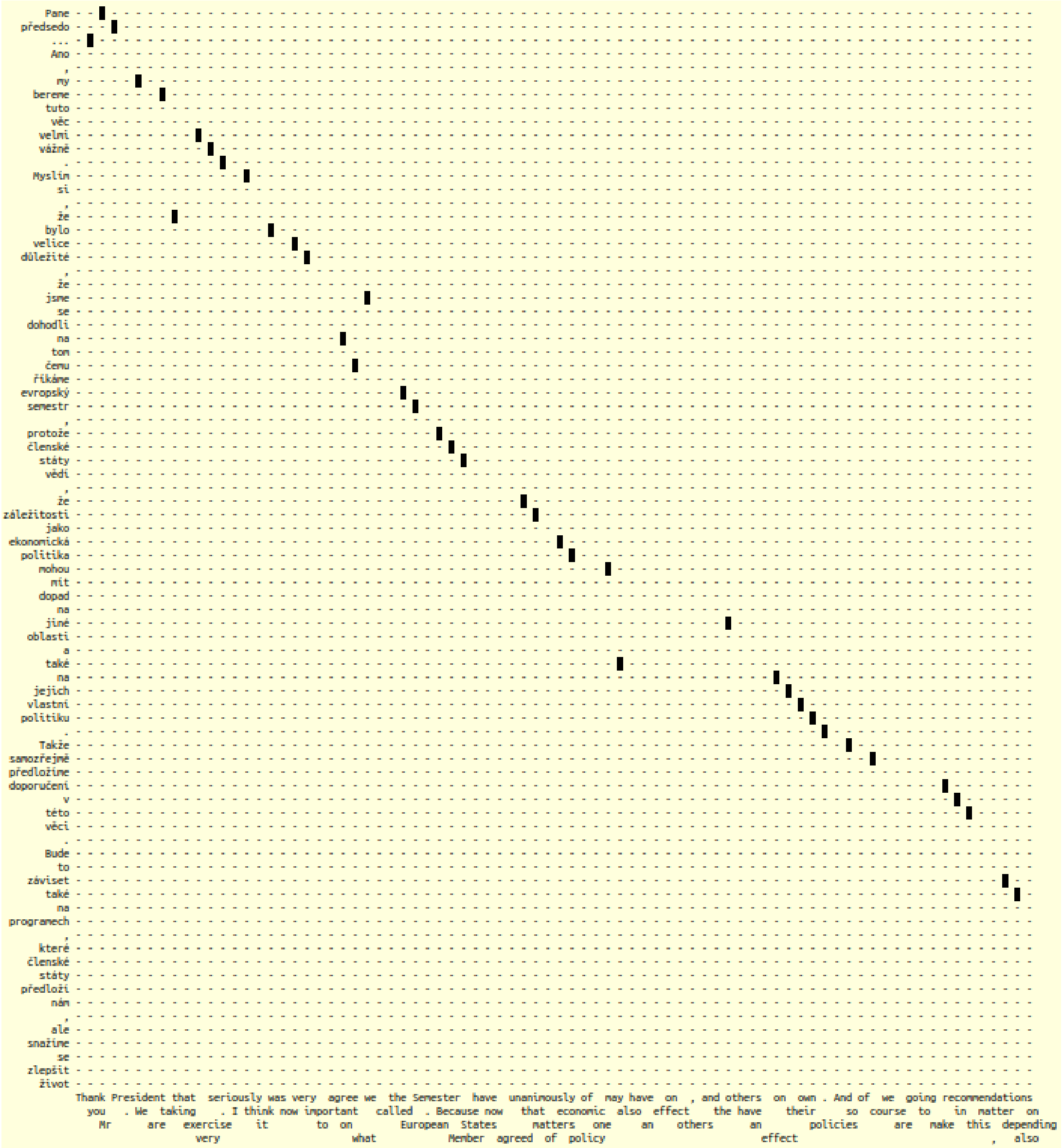

Pane
předsedo
...
Ano
,
my
bereme
tuto
věc
velmi
vážně
.
Myslím
si
,
že
bylo
velice
důležité
,
že
jsme
se
dohodli
na
tom
čemu
říkáme
evropský
semestr
,
protože
členské
státy
vědí
,
že
záležitosti
jako
ekonomická
politika
mohou
mít
dopad
na
jiné
oblasti
a
také
na
jejich
vlastní
politiku
.
Takže
samozřejmě
předložíme
doporučení
v
této
věci
.
Bude
to
záviset
také
na
programech
,
které
členské
státy
předloží
nám
,
ale
snažíme
se
zlepšit
život
Thank President that seriously was very agree we the Semester have unanimously of may have on , and others on own . And of we going recommendations
you . We taking . I think now important called . Because now that economic also effect the have their so course to in matter on
Mr are exercise it to on European States matters one an others an policies are make this depending
very what Member agreed of policy effect , also

Figure 5.6: Preview of automatic alignment in a simple text view. There is an 80-word prefix of the Czech target in rows and an 80-word prefix of the English source in columns. The English words associated with columns are placed under them, with the first character of the word right under the column. There is a dot grid for orientation. The black rectangles indicate the alignment of the source-target word pair.

	% of source words	in target	aligned
dev	CS-INT	75%	38%
	EN→CS	77%	51%
	DE-INT+DE→CS	82%	36%
	(DE-INT)	86%	40%
	(DE→CS)	95%	38%
test	CS-INT	76%	38%
	EN→CS	77%	50%
	DE-INT+DE→CS	83%	37%
	(DE-INT)	87%	41%
	(DE→CS)	96%	38%

Table 5.5: Percentage of aligned source words in contrast to length rate in number of words on ESIC corpus, on SI vs. MT translation candidates. I.e. in English-Czech SI (CS-INT), there is 75% of words compared to the English source, and 38% of the source words are aligned. The gray-backgrounded lines show the decomposition, only DE-INT or DE→CS.

Latency Comparison

We aim to compare the latency of interpreting and machine translation. We note that the comparison is inevitably limited by different output modalities: The interpreters produce speech, and the machine translation produces text. We disregard the perception effects of hearing versus reading. We consider the center of the time span when the interpreter was uttering a word as the word's production time.

The latency is summarized in Table 5.6. Both CS-INT and DE-INT have an average latency of around 4 seconds. In 90% of the source words that were aligned to any target word, the latency is below 7 seconds. In a small number of cases, at around 1%, the latency is larger than 23 seconds. It can be caused either by interpreters using such long translation units, or a rare error in the automatic alignment. The methodology is the same for all options, therefore we assume that the error rate is homogeneous, although unknown, so the results are comparable.

The machine translation systems used in our work have larger latency than interpreters: EN→CS around 7 seconds, DE→CS around 5 seconds. There are two reasons why their latencies differ, and why they are so large. First, EN→CS uses end-to-end ASR, which is approximately 1 second slower than the hybrid ASR of DE→CS. Second, both systems are used for re-translating growing system prefixes, despite they were trained on full sentences. The first word in the sentence is often finalized after the whole sentence is completed by the speaker. The English source speakers tend to make long sentences, sometimes even 30 seconds, while the DE-INT makes shorter ones. The systems thus translate much longer units than interpreters and therefore have larger latency.

			Percentile ≤		
		avg±std	50%	90%	99%
dev	CS-INT	4.17 ± 4.32	3.21	7.06	22.14
	EN→CS	7.56 ± 5.65	5.97	15.26	27.00
	DE-INT+DE→CS	9.90 ± 6.75	8.57	17.00	34.78
	(DE-INT)	4.26 ± 5.00	3.08	7.34	24.88
	(DE→CS)	4.92 ± 4.78	3.75	10.17	21.38
test	CS-INT	3.99 ± 4.38	3.00	6.77	22.23
	EN→CS	7.68 ± 6.28	5.98	15.17	30.38
	DE-INT+DE→CS	9.84 ± 7.16	8.43	17.08	36.70
	(DE-INT)	4.03 ± 4.70	3.02	6.64	23.27
	(DE→CS)	5.07 ± 4.89	3.90	10.56	20.95

Table 5.6: Latency of interpreting and machine translation from English to Czech (white background), based on automatic word alignments, in seconds. Gray rows break down the two intermediate components of the indirect translation: English-to-German interpreter and German-to-Czech translation. The percentile indicates that, e.g. 90% of aligned words fit under 7 seconds. The gray-backgrounded lines show the decomposition, only DE-INT or DE→CS.

Discussion The indirect DE-INT+DE→CS option has latency around 10 seconds between English and Czech, i.e. roughly twice larger than a single interpreter. This is comparable to relay interpreting via one intermediate pivot language. Relay interpreting is used in real-life settings, so real users might be accustomed to latencies of around 10 seconds. Therefore, we consider the indirect path of interpreter followed by machine translation as feasible from the latency point of view.

5.4 Summary

We overviewed the challenges of SI in multi-source SST: interpreting strategies and the issue of quality that makes SI not word-for-word parallel to the original source, and the latency.

On one hand, in our analysis of ESIC, we found that SI tends to be shorter and less complex than direct translation from the original, although not word-for-word and not preserving as much information as translation. On the other hand, shortening, simplification, and redundancy reduction could actually be benefitial, especially for disfluent or complex speeches that are difficult to understand. SI can also include inter-cultural transfer.

Latency is the disadvantage of SI in SST. More than 90% of the original words fit within 7 seconds of latency, even though the median is 3 seconds. We have also shown that an example SST from SI achieves 9 seconds median latency, which we consider feasible, similar to relay interpreting latency that is perceived as acceptable.

Although SI is not an entirely parallel source, we see an opportunity for enhancing SST quality with SI as the additional source. We assume it could be possible with multi-sequence to sequence NMT trained by supervised learning. The source can be a multi-sequence original and SI, possibly shifted because of the latency, and target the verbatim translation of the original. We assume that the training data can be authentic, e.g. the 121 hours of SI from the European Parliament that we downloaded and analyzed, but not included to ESIC, or they can be synthesized, e.g. using style transfer model as Zhao et al. (2021a). The latency can be also synthesized based on the distribution found on ESIC.

There are open questions left for the next chapters: Could SI be used for higher robustness to the ASR errors? What the reference for SST quality assessment should be like, translation, or interpreting? And last, but not least, how to design the SST system that would leverage both the original and SI?

6 Multi-Sourcing and Robustness

We split our main task, designing multi-source simultaneous speech translation (SST) from the original and simultaneous interpreting (SI), into two subsequent tasks that are easier to handle separately than together. First, in this chapter, we analyze whether SI can be used in multi-source SST for higher robustness to automatic speech recognition (ASR) errors. ASR errors negatively impact translation quality because of compounding speech recognition and translation errors (Ruiz et al., 2017; Sperber and Paulik, 2020), and thereby limiting the application of automatic speech translation in realistic settings. Later, if we discover that the robustness of multi-sourcing is sufficient, we focus on the second task, designing the multi-source model for a realistic setup. We leave it for the next chapters.

We investigate the robustness in two steps. First, we investigate the hypothesis that the ASR errors from two parallel language sources, the original and SI, are independent, and could be complementary. In the next step, we create a multi-source neural machine translation (NMT) model and investigate its robustness to ASR errors in the sources. For that, we simplify the setup. We use parallel, aligned, and synchronized translations from ESIC, and not original and SI that is not totally parallel, unaligned to parallel segments, and not synchronized. It is a less realistic use case than translating long speech documents without any sentence segmentation and alignment of the sources, but proving the robustness of multi-sourcing in this setting paves the way for its application in long speech document translation.

In this chapter, we use text that we already published in the paper "Robustness of Multi-Source MT to Transcription Errors" (Macháček et al., 2023c). Our original findings in this chapter are the third significant contribution of our dissertation.

subset	Cs interp.	De interp.	En original
dev	14.84	25.14	13.63
test	14.04	23.79	14.71

Table 6.1: Transcription WER on ESIC. There are 191 and 179 documents in dev and test subsets. The scores are weighted by the number of words in gold transcripts.

6.1 Independent Errors

We assume that a multi-source setting with the original speech and its simultaneously interpreted equivalent as the two sources will improve robustness to ASR errors if the errors in the two source streams complement each other. This is not obvious because, on the one hand, the ASRs work independently; they are deployed for different languages, trained on different data, and the processing is fully independent. On the other hand, the content of the speeches is parallel, almost identical. Interpreters' speech pacing also depends on the original speaker, and it may influence the quality of both ASRs in the same way. Therefore, in this section, we analyze the dependency of ASR errors in the source and interpreter, on 10-hour ESIC corpus (Macháček et al., 2021) to prove that the ASR errors are indeed independent.

Methodology First, we processed ASR for English original speakers and interpreters into Czech and German. We used the same systems from the ELITR project as in Chapter 5. For English, we used the low-latency neural ASR by Nguyen et al. (2021b). For German, we used an older hybrid HMM-DNN model trained using the Janus Recognition Toolkit, which features a single-pass decoder (Cho et al., 2013). For Czech, we used Kaldi (Povey et al., 2011) HMM-DNN model trained on Czech Parliament data (Kratochvíl et al., 2020). Table 6.1 summarizes the transcription quality on ESIC showing that the quality is low, but to the best of our knowledge it was the best one available for this domain in 2022 when we worked on this analysis.[1]

We then re-used the word alignments of gold transcripts between the original and interpreting as described in Section 5.3.2 and in Macháček et al. (2021). 38% of tokens were aligned between English and Czech interpreting, and 40% between English and German, see Table 6.2. It may be caused by the characteristics of the language pair (e.g. compound words in German vs. multi-word expressions in English), features of interpreting (non-verbatim translation, shortening) and by errors in automatic alignment. We only analyzed the aligned tokens further. Since there are many tokens left in two 5-hour subsets of the corpus, we consider further analysis as valid.

[1] We did not consider Whisper in this analysis because it was released later, on 21. 9. 2022.

	En tokens	En-Cs aligned	En-De aligned
dev	44,494	16,962 (38.12%)	17,809 (40.03%)
test	46,151	17,623 (38.19%)	19,280 (41.78%)

Table 6.2: Number and percentage of aligned tokens in gold transcripts between the original source (English [En]) and its interpretations (German [De] and Czech [Cs]).

En orig.		**Cs int.**		**De int.**	
		corr.	**incorr.**	**corr.**	**incorr.**
dev	**corr.**	13815	1497	7192	1561
	incorr.	1228	422	633	307
test	**corr.**	14204	1655	7895	1638
	incorr.	1344	420	692	336

Table 6.3: Contingency table of correctly and incorrectly recognized aligned tokens in English source (in rows) and interpretation into Czech and German (in columns), in dev and test subset of ESIC corpus. According to the χ^2 test of statistical independence, in all 4 cases, the parallel recognition is independent with $p < 0.01$.

Finally, we aligned gold and automatic transcripts using Levenshtein edit distance.[2] We classified each token in the ASR transcript as transcribed correctly or not, both for source and interpretings.

Results We made a contingency table (Table 6.3) and ran a χ^2 test (Pearson, 1900) of statistical independence. The results show that the **ASR systems applied on the parallel source and interpreting make errors independently** of each other with $p < 0.01$, for both pairs, English-Czech and English-German, for both dev and test subsets.

Severity of errors We drew 100 random pairs of aligned English-German ASR word transcripts where at least one is incorrect, and manually categorized the severity of errors. We qualified the following categories:

- correct – when there is an exact match of the gold and ASR transcript (i.e. the other language than the incorrect one)
- serious – a substitution that would have a serious impact on understanding in the source language, such as "produktion" instead of "evolution."
- small – a minor substitution that would have at most small impact on understanding in the source language, such as the wrong inflection form "stimmt" instead of "stimmen" (German "agree," 3rd person singular vs. plural).

[2] `https://pypi.org/project/edlib/`

lan.	**type**	**gold**	**ASR**
en	correct	evolutionary	evolutionary
de	serious	evolution	produktion
en	serious	commissioner	∅
de	correct	kommissarin	kommissarin
en	serious	communication	upcoming opening communication
de	serious	mitteilung	die kommen da kommt mitteilung
en	small	physiotherapy	physical therapy
de	correct	physiotherapie	physiotherapie
en	correct	vote	vote
de	small	stimmen	stimmt
en	normalization	travellers	travelers
de	correct	reisenden	reisenden
en	correct	2009	2009
de	normalization	2009	zweitausendneun
en	correct	one	one
de	disfluency	einer	ein einer

Table 6.4: Examples of errors severity categories in pairs of aligned words.

- normalization – when the non-exact match of the gold and ASR transcript is caused by non-matching orthography normalization, e.g. British and American spelling "traveller" versus "traveler," or numerals in digit versus word form.
- disfluency – a non-exact match that may be caused by wrong disfluency removal component of the ASR system, such as when the speaker utters "ein einer," the ASR transcribes it, but the gold transcript contains only "einer."

Examples are in Table 6.4, and the contingency table in Table 6.5. The results show that among the English errors, 71% of them are serious (49 serious errors out of 69 errors of any type). Among the German errors, 51% of them are serious. We want to highlight that these proportions are a property of the ASR systems that we used. Other systems could lead to different results.

However, although we have evidence on one test set, two language pairs and two pairs of ASR systems, the fact that most of the serious errors in English or German ASR are aligned with correct, and not serious error in the other language, allows us to believe that the ASR errors in two parallel language streams are indeed independent also when using other ASR systems.

		German					
		correct	serious	small	norm.	disfl.	sum
	correct	-	41	14	2	2	59
	serious	13	7	1	0	0	21
English	small	14	2	1	0	0	17
	normalization	3	0	0	0	0	3
	disfluency	0	0	0	0	0	0
	sum	30	49	16	2	2	100

Table 6.5: Number of error severity categories in a random sample of 100 aligned English-German pairs of words in ESIC.

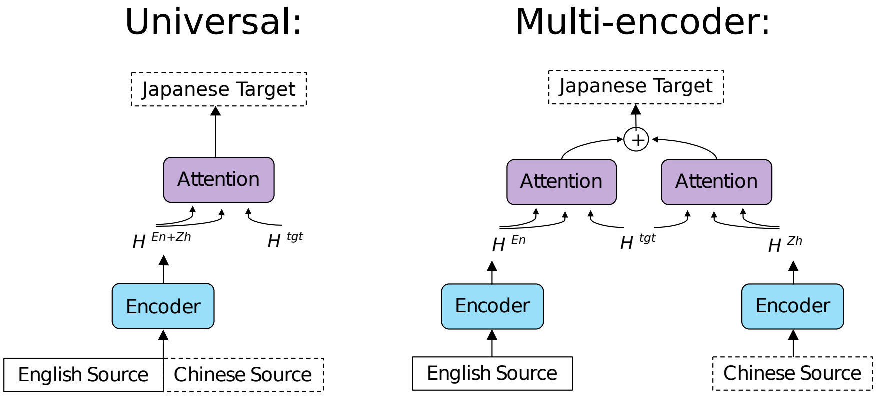


Figure 6.1: Illustration of joining the sources in universal and multi-encoder model. H^* stands for encoder representation, e.g. of the concatenation of English and Chinese sources ($En + Zh$), or separate English (En) or Chinese (Zh). H^{tgt} stands for previous target states. Figure reproduced from Xu et al. (2021).

6.2 Multi-Sourcing

Multi-lingual machine translation uses more than two languages within a single system (Dabre et al., 2020; Kocmi et al., 2021). It has been shown to improve translation quality, flexibility, or efficacy in various situations. In particular, multi-source machine translation (MT), as a subarea of multi-lingual MT that we primarily focus on, has a potential for improving translation quality.

In this section, we first review multi-sourcing NMT architectures that we considered and selected for our experiments. Then, we describe our implementation and training of multi-source NMT model, and present results with clean parallel sources.

6.2.1 Architectures

Combining sources There are two alternative methods to combine multiple source sequences in the NMT model. The first, multi-sequence method (Xu et al., 2021; Zoph and Knight, 2016; Dabre et al., 2017) handles multiple sequences similarly as one long sequence. The model can access all the elements in all sequences and learn to use them in any way needed. For example, the second, auxiliary source can be a very loose translation of the original source. Multi-sequence model can learn to use the auxiliary source for word disambiguation while translating the original source more literaly. Multi-sequence model can also complement information from two sources, or learn that a part of one source may be wrong and should be avoided. There are two options how to model multi-sequence model: concatenation in a single encoder, and single sequences in multiple encoders, as illustrated by Xu et al. (2021) in Figure 6.1.

The other method for combining multiple sources is averaging (Firat et al., 2016b). The source sequences are encoded and processed by the model individually, in parallel copies of the model, until a specific place inside the model where their representations are averaged. This method assumes that the sources are totally parallel in meaning, except for some noise that is smoothed by averaging. The second assumption is that the neural representations stemming from the language sources are compatible, e.g. nearly identical. This is ensured by model design, e.g. universal multi-lingual encoder, or sharing the target vocabulary.

The disadvantage of averaging is that it requires totally parallel sources. Every non-parallelity in meaning can lead to error.

These approaches differ in the complexity of required training data. Multi-sequence model requires multi-parallel training data that are often unavailable in large amounts. The averaging model can be trained with pairs of bilingual data, one source-target pair for each source. Parallel multi-sourcing is applied only in decoding.

Early and late averaging Firat et al. (2016b) propose early and late averaging of parallel sources. *Early* averaging requires shared training of a single encoder or multiple encoder models on multiple source languages and one shared decoder. In early averaging, the sources are encoded in two independent paths. The context vectors that are a result of encoding and attention are averaged and then decoded as a single source. *Late* averaging is similar to ensembling. Multiple independent models (or

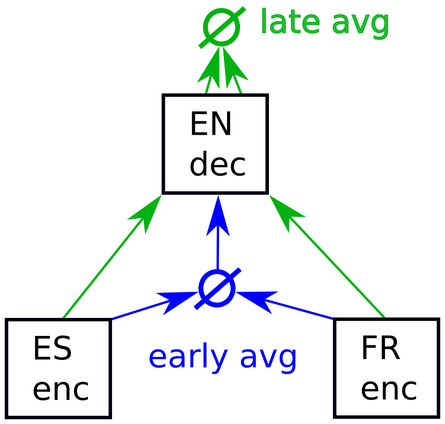


Figure 6.2: Illustration of multi-sourcing early and late averaging by Firat et al. (2016b). In "early" averaging, the context vectors from two independent encoders, one for each source language, are averaged, and then decoded as a single language. "Late" averages the distributions of target vocabulary from decoding two independent paths, one from each source language.

model checkpoints, or copies of the same model) that share the target vocabulary encode and decode the sources separately. The distributions over the output vocabulary are then averaged, just before choosing the output token. Figure 6.2 illustrate late and early averaging.

The advantage of late averaging over early averaging is the flexibility. There can be separately trained models or checkpoints, e.g. selected as optimal for each source. There can be also any number of models, while for early averaging, the encoders must correspond to the decoder, so they must be trained together.

The advantage of early averaging, besides efficiency because decoding is done only once, is the possibility of applying weighting based on word-level confidence scores, and then decode according to them. The disadvantage is the necessity of joint training of the encoders and decoders, the resulting checkpoint might not be optimal for all the sources.

Universal vs. multi-encoder There are two main types of multi-source model designs. The first is a universal model that has a single encoder and decoder for all the sources. It has the same shape and training method as the basic bilingual model. The multiple language sources are represented only in training and evaluation data, e.g. concatenated into one sequence with or without delimiter token, as illustrated in Figure 6.3. The advantage of this approach is better and more flexible generalization

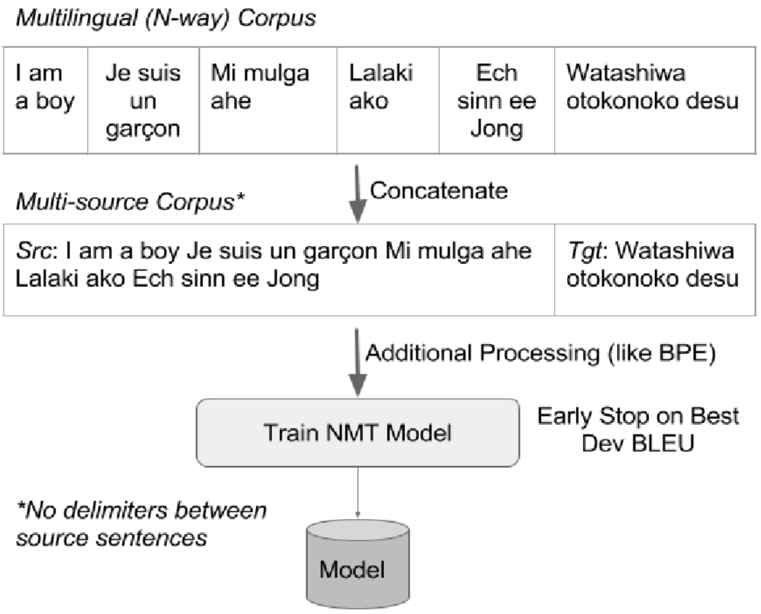

Multilingual (N-way) Corpus
I am a boy
Je suis un garçon
Mi mulga ahe
Lalaki ako
Ech sinn ee Jong
Watashiwa otokonoko desu
Concatenate
Multi-source Corpus*
Src: I am a boy Je suis un garçon Mi mulga ahe Lalaki ako Ech sinn ee Jong
Tgt: Watashiwa otokonoko desu
Additional Processing (like BPE)
Train NMT Model
Early Stop on Best Dev BLEU
*No delimiters between source sentences
Model

Figure 6.3: Multi-source NMT with concatenation by Dabre et al. (2017). The figure is reproduced.

across languages because internal sharing parameters for languages are learned by the model itself. The other approach is using multiple encoders, one for each source language. The advantage is the larger capacity for each language. The disadvantage is longer and more complicated training, and the risk of insufficient sharing.

Zoph and Knight (2016) and Firat et al. (2016b) propose multi-encoder models, while Dabre et al. (2017) experiments with both. The results in Dabre et al. (2017) show that the universal model performs better for related languages, while multi-encoder performs better for more distant languages.

Multi-way Multi-lingual multi-way NMT model allows translation from N source languages into M target languages. It is assumed that $N+M > 2$ because otherwise, it is a bi-lingual and not a multi-lingual model. Multi-way model, in contrast to parallel multi-source, can translate in only one of the multiple available language directions within a single decoding process.

Multi-way NMT is usually a universal single encoder-decoder model where the multilinguality is represented only in data. The source sequences, either sentences or other text units, are connected with a special token that prompts the desired translation direction.

Multi-way NMT was originally proposed by Ha et al. (2016), Firat et al. (2016a) and Johnson et al. (2017). It became a standardly used technique (Aharoni et al., 2019; Arivazhagan et al., 2019; Zhang et al., 2020a; NLLB Team et al., 2022; Huang et al., 2023).

In our experiments, we use multi-way for its flexibility, efficiency, and expected higher quality because of multilinguality. The next reason why we use the 2-to-1 multi-way model is the ability of early averaging as a multi-source decoding method.

6.2.2 Training Data

We investigate the robustness to ASR errors in the simplified setup with the universal multi-way model for English and German as the source languages, and Czech as the target. We train it using English-Czech and German-Czech bilingual, not multi-parallel training data. We use all English-Czech and German-Czech parallel corpora from OPUS collection (Tiedemann and Nygaard, 2004). We carry out the following data processing steps:

1. Downloading – we downloaded data from OPUS website[3] using OPUS-MT-train tool[4] (Tiedemann and Thottingal, 2020). We tried the tool because we supposed it would enable us creating the whole multi-way model from data downloading and cleaning to NMT training with one simple command. Unfortunately, we realized the tool did not work as simply as we expected. We had installation issues with some underlying software, and it did not work as flexibly as we needed. In the end, we used the tool only for downloading the corpora from the OPUS website.

2. Cleaning – we adapted the standard NMT training data cleaning process from the Bergamot project.[5] It uses tools from Moses that remove non-printing characters and normalize punctuation, and fasttext (Joulin et al., 2016b,a) language identification to filter out sentence pairs that are not detected as the desired languages.

[3] https://opus.nlpl.eu

[4] https://github.com/Helsinki-NLP/OPUS-MT-train

[5] https://github.com/browsermt/students/blob/master/train-student/clean/clean-corpus.sh

lan. pair	all clean	selected top 30M
En-Cs	148 440 352 (100%)	30 000 000 (20%)
De-Cs	49 737 361 (100%)	30 000 000 (60%)

Table 6.6: Number of sentence pairs of OPUS training data at the beginning and after selection of top 30 million by dual cross-entropy score, followed by the number of space-delimited words in each source and Czech target.

3. Removing test sets – in order to reliably assess the performance of our NMT model on standard MT test sets, we ensured that the training data do not overlap with the test sets. We removed all sentence pairs that could be found in any test set for English-Czech, German-Czech, or English-German MT, in the test sets from IWSLT, WMT, ESIC, Europarl-ST, FLORES, and MUST-C, in all versions until 2022.

4. Scoring parallelity – dual cross-entropy (Junczys-Dowmunt, 2018) is a method that scores sentence pairs by the extent to which they are translations of each other, and can be used to filter out the non-parallel sentences. For scoring, we use pairs of bilingual models for English-Czech and German-Czech from OPUS-MT (Tiedemann and Thottingal, 2020). These models are trained on the data that we score, however, we visually checked that the models do not overfit by scoring the largest corpora higher, so we assume the scoring is reliable.

5. Selection – we selected the 30 million sentence pairs for English-Czech and German-Czech that had the top dual cross-entropy score. We selected 30 million because it is 60% of all German-Czech data, and empirical results in Chen et al. (2021) suggest that the optimum is between 50 and 75%. Table 6.6 shows the size and proportion of data we selected. We also selected the same amount of training data for each language pair to prevent overfitting to the language, however, it is still possible because English-Czech may be of higher quality because they are selected from a larger amount.

6. Language id token – we prefixed every source sentence with source language identification token "<lang:en>" or "<lang:de>" that informs the model from which language to translate into Czech. We made sure that the tokens have a single entry in the vocabulary.

7. Shuffling – we merged English-Czech and German-Czech training data into one training set and randomly shuffled them, to enable multi-way training. Figure 6.4 shows an example of the resulting training data.

corpus	all clean	%	% in tests	top30M	%
CCMatrix	55 231 120	37.2	0.0	7 392 644	24.6
ParaCrawl	45 853 600	30.9	0.1	6 689 512	22.3
OpenSubtitles	30 013 267	20.2	1.6	11 352 773	37.8
CCAligned	8 616 268	5.8	0.6	2 010 850	6.7
DGT	2 899 443	2.0	0.2	1 084 289	3.6
XLEnt	1 117 937	0.8	0.4	342 836	1.1
JRC-Acquis	740 035	0.5	0.1	285 091	1.0
ELRC_2682	689 321	0.5	0.0	244 726	0.8
Europarl	627 294	0.4	2.6	122 867	0.4
WikiMatrix	503 125	0.3	0.0	44 305	0.1
QED	360 238	0.2	1.9	67 435	0.2
EUbookshop	342 820	0.2	0.1	60 778	0.2
ELITR-ECA	237 692	0.2	0.0	35 841	0.1
EMEA	237 274	0.2	0.2	104 930	0.3
Tanzil	215 456	0.1	0.0	13 145	0.0
News-Commentary	206 825	0.1	0.4	24 888	0.1
TED2020	159 535	0.1	6.5	23 800	0.1
wikimedia	79 683	0.1	0.1	13 473	0.0
KDE4	60 720	0.0	0.7	25 236	0.1
bible-uedin	59 920	0.0	0.0	3 783	0.0
ECB	46 168	0.0	0.1	14 617	0.0
Mozilla-I10n	38 140	0.0	0.7	11 047	0.0
WMT-News	32 488	0.0	100.0	0	0.0
Tatoeba	28 059	0.0	0.2	20 583	0.1
GlobalVoices	17 035	0.0	0.1	1 707	0.0
Wikipedia	7 200	0.0	0.1	3 206	0.0
EUconst	4 752	0.0	0.2	2 030	0.0
PHP	3 951	0.0	0.1	948	0.0
ELRC_3382	3 652	0.0	0.1	540	0.0
Ubuntu	3 518	0.0	0.2	1 356	0.0
TildeMODEL	2 363	0.0	1.2	639	0.0
ELRC_2922	1 073	0.0	0.0	74	0.0
ELRC_2923	294	0.0	0.3	9	0.0
GNOME	86	0.0	1.2	42	0.0
total	148 440 352	100.0		30 000 000	100.0

Table 6.7: English-Czech training corpora from OPUS. There is a number and the proportion of the clean sentence pairs, the percentage of their overlap to the test sets that we removed, and distribution in the finally selected top 30 million sentence pairs.

corpus	all clean	%	% in tests	top30M	%
CCMatrix	32 420 826	65.2	0.0	18 409 823	61.4
OpenSubtitles	11 528 172	23.2	1.1	7 580 887	25.3
DGT	2 771 597	5.6	0.3	2 074 643	6.9
JRC-Acquis	750 624	1.5	0.2	636 393	2.1
Europarl	552 309	1.1	2.4	378 738	1.3
EUbookshop	306 377	0.6	0.2	152 506	0.5
EMEA	231 845	0.5	0.2	195 249	0.7
WikiMatrix	230 133	0.5	0.1	75 339	0.3
News-Commentary	191 771	0.4	0.0	98 912	0.3
QED	190 014	0.4	2.9	95 521	0.3
TED2020	145 476	0.3	6.2	68 788	0.2
XLEnt	89 200	0.2	1.3	38 299	0.1
ECB	62 647	0.1	0.1	45 055	0.2
Tanzil	61 378	0.1	0.0	31 794	0.1
KDE4	57 812	0.1	1.2	39 197	0.1
ELITR-ECA	33 997	0.1	0.1	23 567	0.1
bible-uedin	30 165	0.1	0.0	15 751	0.1
Mozilla-I10n	28 067	0.1	0.9	16 712	0.1
WMT-News	20 098	0.0	100.0	1	0.0
wikimedia	13 069	0.0	0.3	7 620	0.0
EUconst	4 385	0.0	0.3	3 955	0.0
Tatoeba	4 375	0.0	0.4	4 219	0.0
GlobalVoices	3 698	0.0	0.0	1 448	0.0
PHP	3 612	0.0	0.1	1 661	0.0
Ubuntu	3 322	0.0	0.7	2 125	0.0
TildeMODEL	2 305	0.0	0.2	1 742	0.0
GNOME	87	0.0	3.4	55	0.0
total	49 737 361	100.0		30 000 000	100.0

Table 6.8: German-Czech training corpora from OPUS. Column description is the same as in Table 6.7.

English or German source	→	**Czech target**
<lang:de> Ja, die Physik funktioniert.	→	Je jasné, že fyzika působí.
<lang:en> Maybe, had I been blind, I'd have sung better.	→	Možná kdybych byl slepý, tak bych zpíval líp.
<lang:en> The Marketing Authorisation Holder provided supplementary information on 20 April 2005, 20 December 2005, 27 March 2006 and 9 May 2006.	→	Držitel rozhodnutí o registraci předložil doplňující informace ve dnech 20. dubna 2005, 20. prosince 2005, 27. března 2006 a 9. května 2006.
<lang:en> Hundreds of death threats.	→	Několik stovek vyhrůžek smrtí.
<lang:de> Du weißt nicht, wann du aufhören sollst, oder?	→	Nevíš kdy skončit, co?
<lang:de> Du kannst mir deine Antwort dann geben.	→	Potom mi můžeš odpovědět.

Figure 6.4: Example of training data.

Distribution of corpora Tables 6.7 and 6.8 show the distribution of corpora in English-Czech and German-Czech training data. We observe that English-Czech training corpus contains mostly three large copora. There is 38% of OpenSubtitles (Lison and Tiedemann, 2016), which is a collection of movies and series subtitles, 25% of CCMatrix (Schwenk et al., 2021), a collection of parallel texts from the Web, and 22% of ParaCrawl (Bañón et al., 2020), another Web collection. The remaining 5% are other small corpora.

The distribution of German-Czech (Table 6.8) is similar, with the exception that ParaCrawl is not available for this language pair. Therefore, there is a higher portion of CCMatrix – 61%, and 25% of OpenSubtitles. There is 6.9% of DGT,[6] translation memories of EU legislation. The remaining part is other small corpora.

Test sets overlaps The highest overlap with the test sets had WMT-News, which is a collection of WMT test sets and was therefore completely removed. Lower, but significant overlap had Europarl, probably because of overlap with ESIC and Europarl-ST, and TED2020, probably overlapping with MUST-C and IWSLT test sets. We can not explain the high overlap of GNOME and QED and the other overlaps. That could be due to the short sentences that were also in ESIC and TED talks in IWSLT tests.

[6] `https://joint-research-centre.ec.europa.eu/language-technology-resources/dgt-translation-memory_en`

Considered backtranslation We considered using larger backtranslated data, e.g. CzEng 2.0 (Kocmi et al., 2020) for English-Czech, and English-German backtranslation from WMT (Chen et al., 2021). In the end, we did not use them because we had a sufficient amount of authentic data from OPUS. Furthermore, German-Czech backtranslations are not available. We would need to create them on our own to have the same size of data for the two language directions, and it would require lots of work that is not within our main focus.

6.2.3 Creating Multi-Source Model

We trained a multi-way NMT model using Marian (Junczys-Dowmunt et al., 2018), a fast and effective toolkit for NMT training and evaluation. Then we implemented early and late averaging as multi-source decoding methods and evaluated the model. We describe the model creation in more detail.

Model The NMT model is Transformer Base (6 layers, 512 embedding size, 8 self-attention heads, 2048 filter size; Vaswani et al., 2017). We use two separate SentencePiece (Kudo and Richardson, 2018) vocabularies, both sizes of 16 000. The source vocabulary is joint for German and English and the target is only for Czech.

Training We train the model on 8 Quadro P5000 GPUs with 16 GB memory for 17 days, until convergence.

Validation and evaluation For NMT validation and evaluation, we use the "revised transcript and translations" from ESIC (Macháček et al., 2021, Section 4.5). These are the texts that were originally uttered in the European Parliament, transcribed, revised and normalized for reading and publication on the website, and then translated. They are analogous, but not identical, to the gold transcripts of the original and interpretations that we used in Section 6.1. We use the ESIC 1.1 version where we manually align the sentences in all three languages properly (ref. Section 4.6.2). Two documents were removed because they missed German translation. The corpus is of comparable size to a usual MT test set. The size statistics are in Table 4.6.

For a contrastive evaluation, we use Newstest11 (Callison-Burch et al., 2011). It contains 3003 sentences in five languages: English, German, Czech, French, and Spanish, the same amount in each. Newstest11 has references that were translated directly, not through an intermediate language. We also use three additional Czech references of Newstest11 that were translated from German (Bojar et al., 2012).

Checkpoint selection We validate all checkpoints (every 1 000 training steps, 15 minutes) on two single sources (English and German) and two multi-sourcing options: early averaging, and late averaging of a single checkpoint with two sources. Furthermore, after the training had ended, we selected the top 10 checkpoints that reached the highest BLEU scores for English and German single-source on the ESIC dev set. We evaluated all pairs of the top-performing checkpoints in late averaging multi-sourcing setup. The top-performing model from all validation and grid search options was selected as the final model. It is the late averaging model with a pair of distinct checkpoints, one for English-to-Czech and one for German-to-Czech. We also use these two checkpoints for single-source evaluation.

6.2.4 Multi-Sourcing Implementation

We convert Marian models to PyTorch to be used with the Hugging Face Transformers (Wolf et al., 2020) library, in which we implement late and early averaging. The advantage of Hugging Face over Marian is simple and flexible development. It is in Python programming language with PyTorch (Paszke et al., 2019) deep learning backend. The disadvantage is slow processing, due to many abstraction layers between the program and machine processing unit. We did not measure the processing time of our implementation because we did not need to optimize it for our experiments. We assume that if our multi-sourcing method appears to be useful, then it can be implemented in a way that supports fast real-time processing.

For both single- and multi-sourcing, we use greedy decoding because the beam search support is not implemented with multi-source.

Figure 6.5 shows a preview of validation scores on ESIC dev set for early and late averaging using different mixes of language sources, either one, two, or three sources, where two are identical and one is different ("de+de+en" or "en+en+de"). We used it primarily to inspect whether our implementation gives different results than two equally-weighted sources. The checkpoint in this evaluation is not the finally selected one, therefore the scores differ from the ones in the next section. We observe a reasonable trend that both averaging using the same single source (e.g. "en_early" vs. "en_late") achieve identical scores because averaging only one source is identical to single-source decoding. There is only one exception of 0.1 BLEU difference in the case of "de_early" and "de_late." It is probably caused by the fact that PyTorch does not guarantee deterministic and reproducible operation on various hardware

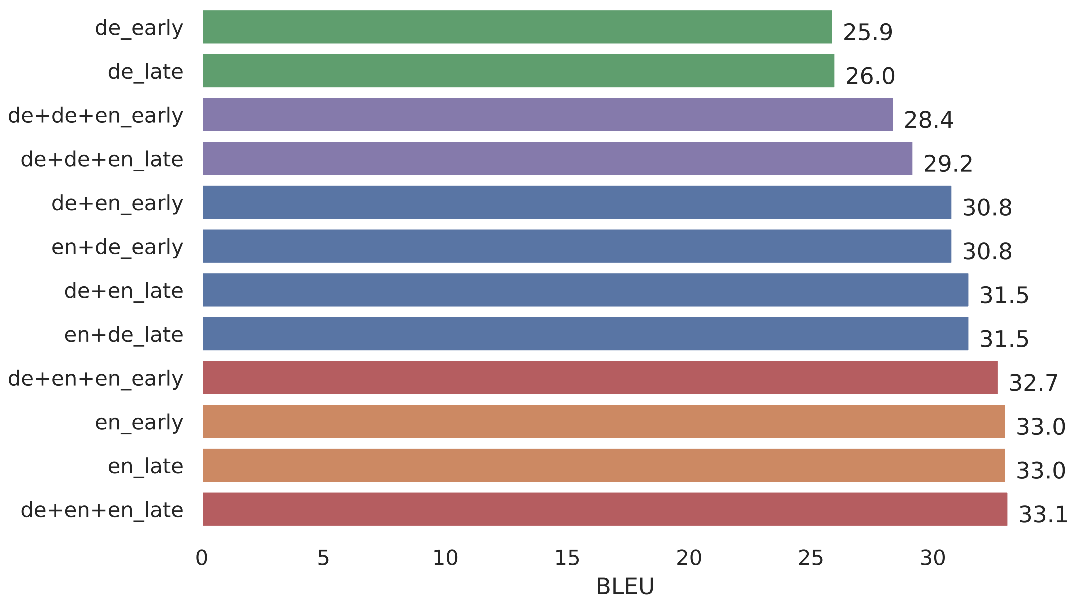


Figure 6.5: Preview of quality scores that we used to inspect the implementation of early and late averaging. We report the BLEU score on ESIC dev set on German ("de") and English ("en") sources. The colors of the bars represent the mix of sources, which is also included in the row labels, together with "early" or "late" averaging.

platforms.[7] We use a computer cluster with various GPU types, and we did not bind the GPU type to be identical between these evaluations. Another expected result in Figure 6.5 is that the order of sources ("en+de" vs. "de+en") does not influence the scores. Averaging is not affected by the order.

Averaging selection We observe on ESIC dev (Figure 6.5) and on various other checkpoints that late averaging achieves higher scores than early averaging for multi-sourcing English and German. We explain it by the fact that late averaging allows the model to use more capacity than early one. Decoding is processed twice independently, while in early averaging, decoding is processed once. These results are consistent with Firat et al. (2016b). Because of the higher quality, we further experiment only with late averaging.

[7] https://pytorch.org/docs/stable/notes/randomness.html

Weighting the sources As we observed in Figure 6.5, averaging two identical English sources with one German source gives a higher score than one English and one German ("de+en+en_late" 33.1 BLEU vs. "de+en_late" 31.5 BLEU). It is because the English source has a higher weight in averaging than the German source, the weight is 2/3 vs. 1/3. Another reason is that the English single source performs better, for reasons that we analyze in the next section.

In any way, we investigate weighting the sources. We implemented a parameter for multiplying the sources by given weights. For example in early averaging, the context vector c is averaged from the context vectors c_1, c_2 of the underlying sources:

$$c = \frac{c_1 + c_2}{2}$$

(Firat et al., 2016b). Given non-negative weights w_1 and w_2 for which $w_1 + w_2 > 0$, assuming they correspond to the sources with context vectors c_1 and c_2, *weighted early averaging* counts the averaged context vector as

$$c = \frac{w_1 c_1 + w_2 c_2}{w_1 + w_2}.$$

We implement support for any number of sources, not only for two. We apply analogical weighting also to late averaging.

Figure 6.6 shows results with different weights between 0 and 10 that sum to 10 on two sources, English and German. We observe that weighting German 2 or 3 and English 8 or 7 gives higher scores than the equal weights 5 and 5 (33.5 BLEU in bars 2 and 3 vs. 31.5 in bar 5 for late averaging). These weights are also higher than English single source (33.0 BLEU, bar 0).

We observe that single source decoding in the right-most and left-most bars do not give equal scores for late and early averaging. We assume it is again because of non-reproducible processing in PyTorch.

The weight parameter can be tuned on the validation set and then applied in inference. However, the fixed single weight can be improved. There can be a neural network that predicts the optimal weight for any sentence pair, and this weight predictor can be trained on training data. Furthermore, there do not have to be two networks, NMT model and the weight predictor. That would be prone to error propagation. There could be one network, a multi-sequence NMT model. We recommend the latest. Moreover, multi-sequence NMT can apply the weight within one sentence pair, which we assume is the most optimal solution.

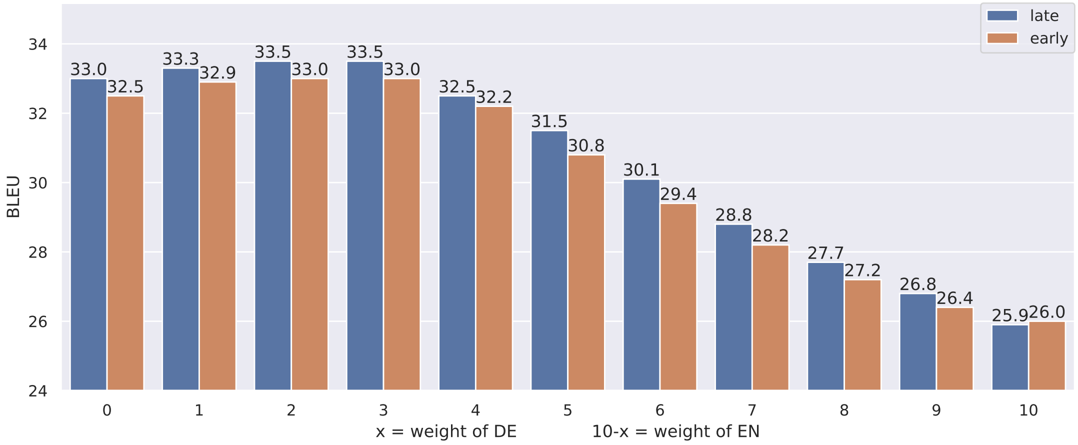


Figure 6.6: Results of weighted early and late averaging with two sources, English (en) and German (de). The vertical bars show the BLEU score on ESIC dev set, using a preliminarily selected model checkpoint. The horizontal axis shows the weight of the German (de) source. The weight of the other, English source, is 10 minus the number on the horizontal axis, below the bars. Therefore, the more to the left, the higher weight for English, and less weight for German. The left-most and right-most bars represent single-source decoding, the other source has 0 weight there.

In summary, weighted averaging is one improvement option of multi-sourcing, but definitely not the only one. Since our main goal is not to search for the most optimal multi-sourcing method but to show the robustness of multi-sourcing, we further use the basic, equal weights in our experiments. If we show that multi-sourcing is robust to noise with equal weights, we assume it stays robust with better tuned weights, and also with multi-sequence NMT.

6.2.5 Results with Clean Sources

First, we evaluate the quality with clean text sources. In the next sections, we investigate quality with ASR noise.

Evaluation metrics We estimate translation quality with BLEU (Papineni et al., 2002) and chrF2 (Popović, 2016) calculated by sacreBLEU[8] (Post, 2018). We also report the current state-of-the-art metric COMET[9] (Rei et al., 2020) that achieved the highest correlation with direct assessment as a kind of human judgments (Mathur et al., 2020) in that year. However, COMET requires one source on the input and is not suitable for multi-source. Therefore, we report it twice (En/De COMET) with two individual

[8] Metric signatures: BLEU|nrefs:1|case:mixed|eff:no|tok:13a|smooth:exp| version:2.2.1, chrF2|nrefs:1|case:mixed|eff:yes|nc:6|nw:0|space:no|version:2.2.1

[9] `wmt20-comet-da` model

sources. Note that En COMET scores assume English as the source and Czech as the target. Since ESIC is tri-parallel, even if the translation is obtained using German or English and German multi-source, we only use the English source as the input to the COMET model. De COMET scores are computed similarly.

Results Table 6.9 shows the results of multi-sourcing with clean inputs, without any speech recognition noise. One would be tempted to conclude that the translation from English is of a higher quality than the translation from German (e.g. 33 vs. 26 BLEU on ESIC dev set), but such a claim is risky. The metrics measure the match of the candidate translation with the reference sentence (and, in the case of COMET, also with the source), and it is conceivable that English served as the source for the human reference translation. The Czech reference thus may very well exhibit more traits of the English source than of the German source. While the chrF2 scores agree with BLEU, COMET scores seem to indicate that multi-sourcing is as good as, if not better than, using a single source. Since COMET is known to correlate with human judgments better than BLEU (Mathur et al., 2020), our results show that multi-sourcing is indeed a viable solution.

Impact of source language To further shed light on the impact of the source used for creating references, we evaluated the models with Newstest11 and computed the scores with three additional references that were translated only from German. The German single source achieves much higher BLEU than the English source (32.23 vs. 16.62 BLEU), with multi-sourcing in between (22.47 BLEU). Similar trends are observed in chrF2 and COMET scores. This is the opposite of ESIC scores, where the reference was obtained from English. It shows that the traits of the source language such as word order, and structure of clauses and terms are remarkable in automatic metrics when the reference is constructed from that source, but these effects may be negligible in human evaluation. Section 6.4.4 contains more details.

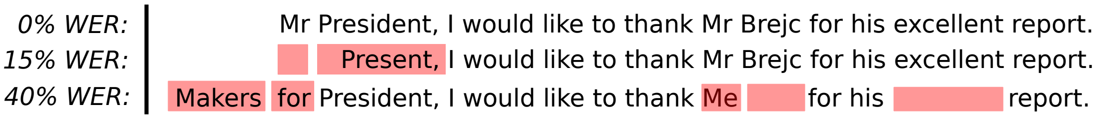
Finally, we consider a "balanced" scenario where an equal number of references comes from each of the source languages and this shows similar scores for both single sources (23.40 vs. 22.85 BLEU) with multi-sourcing outperforming them by 0.6 and 1.1 BLEU. We, therefore, conclude that our multi-source model should be well-prepared for content originating in any of the source languages, but the automatic evaluation metrics may not always capture this.

Set	**Metric**	**Model**		
ref. translation:		En	De	De+En
ESIC dev En→Cs	BLEU	***33.31**	26.13	*31.90
	chrF2	***60.17**	54.00	*58.59
	En COMET	$^\times$**0.920**	0.860	*0.919
	De COMET	$^\times$1.007	0.994	***1.022**
ESIC test En→Cs	BLEU	***33.63**	27.99	*32.57
	chrF2	***59.58**	54.75	*58.63
	En COMET	*0.906	0.871	$^\times$**0.912**
	De COMET	0.994	$^\times$1.006	***1.018**
news11 3×{De→Cs}	BLEU	16.62 ±0.29	**32.23** ±0.53	22.47 ±0.44
	chrF2	44.84 ±0.18	**58.81** ±0.38	49.72 ±0.27
	En COMET	0.528 ±0.002	**0.823** ±0.002	0.652 ±0.003
	De COMET	0.600 ±0.002	**0.967** ±0.001	0.757 ±0.003
news11 {De,En,Fr,Es}→Cs, Cs	BLEU	*23.40	22.85	***23.96**
	chrF2	$^\times$**51.00**	50.27	*50.83
	En COMET	0.627	***0.674**	*0.659
	De COMET	0.700	***0.832**	*0.766

Table 6.9: Evaluation scores with clean inputs (no ASR noise), machine-translated into Czech with single-sourcing English (En) or German (De), or multi-sourcing (De+En), on ESIC and Newstest11 (news11). Newstest is evaluated on a balanced reference that originates in 5 languages ({De,En,Fr,Es}→Cs translations and Cs original; 600 sentences each), and 3 times with additional references that were translated from German ("3×{De→Cs}"). We report avg±stddev for them. "En COMET" and "De COMET" are run with English and German source, respectively. Maximum scores are in bold. The symbol * means that there is a statistically significant difference ($p < 0.05$) from all the lower scores in the same row, $^\times$ means no significance (t-test for COMET, paired bootstrap resampling for BLEU and chrF2).

6.3 Modeling Transcription Noise

Multi-sourcing as a strategy brings various risks. The first is that multi-source SST may not be useful in practice because one source may be always good enough so that the translation quality from the single source could be sufficient and multiple additional sources could rather confuse the model than bring any improvement. Another opposite risk is that all the sources could be of poor quality, so that even in the multi-sourcing combination, the resulting SST quality could be very poor and practically unusable. We, therefore, investigate multi-sourcing quality with different quality levels of multiple sources, to inspect the area between "all too good and all too bad" sources.



0% WER: Mr President, I would like to thank Mr Brejc for his excellent report.
15% WER: Present, I would like to thank Mr Brejc for his excellent report.
40% WER: Makers for President, I would like to thank Me for his report.

Figure 6.7: Example of synthetic ASR errors in the clean text sources. The first line, 0% WER, is the correct, gold transcript. In the second line, there are two errors, deletion of "Mr" and substitution of "President" to "Present." There are 13 words in the gold transcript, the WER is therefore 2/13 = 15%.

In this analysis, we specify the source quality as word error rate (WER) of the underlying ASR systems. WER is a standard quality measure of ASR (Ali and Renals, 2018; Szymański et al., 2020). For simplicity, we focus only on recognition errors reflected in WER and we put aside other aspects that make additional source low quality, e.g. not parallel interpreting, and large latency.

We want to inspect multi-sourcing in various WER levels, but we do not have many different ASR systems with varying quality. Moreover, we do not have aligned original and simultaneous interpreting audio at the level of sentences, to input them into our multi-source MT. Therefore, we use the ESIC revised text translations, and we artificially insert ASR-like errors into the text sources, and evaluate our multi-source MT on various levels of the errors in sources. An example of various level of error levels is in Figure 6.7.

6.3.1 WER Noise Model

We adopt the lexical noise model by Martucci et al. (2021). The lexical noise model modifies the source by applying insertion, deletion, substitution, or copy operations on each word with a probability distribution that are learned from the ASR and gold transcript pairs. It thus may learn to realistically shuffle homonyms such as "eight" and "ate" and similarly sounding words such as "President" and "present" in Figure 6.7.

WER parameter We use the reimplementation and extension of the noise model by our colleague Peter Polák. He extended the model with a parameter for desired WER of the outputs. In the original lexical noise model by Martucci et al. (2021), the desired WER is bound to the performance of the given ASR system on which it is trained, and can not be changed. WER is defined as the number of incorrect words in the ASR transcript divided by the number of correct words in the gold transcript.

The errors are either insertions, deletions, or substitutions. We find a coefficient that we multiply with the parameters of the distributions that randomly select the rewrite operations. The details are in our paper (Macháček et al., 2023c). The coefficient is found as a root of a simplified polynomial.

Implementation The implementation of the noise model is released online.[10] The main author of this thesis contributed to this implementation by adding options to handle casing and punctuation. The parameters support copying the punctuation and casing from the clean input into the noised output targets, or random application of the rewrite rules, or removing punctuation and casing. We also added an option that capitalizes the first letter of the sequence, to match it with sentences that our NMT model is trained on, and adds a full stop if not added by the model.

We also added an option to produce a tag indicating how many characters are deleted by the model. It is useful for synchronizing noised sources in simultaneous mode.

Punctuation and casing We decided to do our analysis simply, with only one option that keeps the punctuation and casing the same as in the source. A better solution would be to investigate the quality levels of punctuation and casing so that we would have four systems to investigate the quality levels, two ASRs, and two punctuators. However, it would complicate reporting and analysis. We therefore assume that the punctuation and casing are optimal.

Training the noise model Our colleague Peter Polák trained the noise model using VoxPopuli (Wang et al., 2021) to retrieve around 100 000 audio and gold transcript sentences in English and 60 000 in German. They are from the same domain as ESIC, both corpora are from the European Parliament. He processed the audio with NVidia NeMo CTC ASRs[11] (Kuchaiev et al., 2019; Gulati et al., 2020). Then he trained the rules of the lexical noise model and we applied them on source data. Since the result is deterministic given the fixed random seed of the lexical noise model, we perform multi-sourcing using three different seeds and report average BLEU scores with standard deviation.

[10] `https://github.com/pe-trik/asr-errors-simulator`

[11] stt_de_quartznet15x5 and stt_en_conformer_ctc_large from `https://catalog.ngc.nvidia.com/models`

BLEU		**ESIC dev**	**En WER**								
		single-src.	0 %	5 %	10 %	15 %	20 %	25 %	30 %	35 %	40 %
s-src.			$33.3^{\pm0.0}$	$29.7^{\pm0.3}$	$26.3^{\pm0.4}$	$22.9^{\pm0.4}$	$20.4^{\pm0.5}$	$18.2^{\pm0.8}$	$15.8^{\pm0.1}$	$14.0^{\pm0.2}$	$12.1^{\pm0.1}$
De WER	0 %	$26.1^{\pm0.0}$	$31.9^{\pm0.0}$	$\underline{30.0^{\pm0.2}}$	$\underline{\mathbf{28.5^{\pm0.3}}}$	$\underline{26.6^{\pm0.1}}$	$25.2^{\pm0.4}$	$23.8^{\pm0.3}$	$21.9^{\pm0.3}$	$20.5^{\pm0.2}$	$19.3^{\pm0.3}$
	5 %	$23.5^{\pm0.0}$	$30.9^{\pm0.1}$	$29.1^{\pm0.2}$	$\underline{27.6^{\pm0.3}}$	$\underline{\mathbf{25.7^{\pm0.1}}}$	$\underline{24.2^{\pm0.4}}$	$22.8^{\pm0.4}$	$21.1^{\pm0.4}$	$19.6^{\pm0.2}$	$18.6^{\pm0.2}$
	10 %	$21.6^{\pm0.2}$	$30.0^{\pm0.2}$	$28.0^{\pm0.1}$	$\underline{26.6^{\pm0.4}}$	$\underline{24.6^{\pm0.3}}$	$\underline{23.4^{\pm0.2}}$	$\underline{21.9^{\pm0.4}}$	$20.2^{\pm0.1}$	$18.7^{\pm0.2}$	$17.5^{\pm0.5}$
	15 %	$19.0^{\pm0.3}$	$28.9^{\pm0.2}$	$27.1^{\pm0.1}$	$25.7^{\pm0.4}$	$\underline{23.7^{\pm0.2}}$	$\underline{22.4^{\pm0.4}}$	$\underline{21.0^{\pm0.4}}$	$\underline{19.3^{\pm0.2}}$	$17.8^{\pm0.3}$	$16.7^{\pm0.4}$
	20 %	$17.1^{\pm0.3}$	$27.9^{\pm0.4}$	$26.6^{\pm0.2}$	$24.9^{\pm0.4}$	$\underline{22.9^{\pm0.1}}$	$\underline{21.7^{\pm0.5}}$	$\underline{20.0^{\pm0.4}}$	$\underline{18.3^{\pm0.2}}$	$17.0^{\pm0.1}$	$15.7^{\pm0.1}$
	25 %	$15.6^{\pm0.3}$	$27.1^{\pm0.3}$	$25.7^{\pm0.2}$	$24.1^{\pm0.3}$	$22.1^{\pm0.2}$	$\underline{20.7^{\pm0.4}}$	$\underline{19.2^{\pm0.5}}$	$\underline{17.4^{\pm0.2}}$	$\underline{16.3^{\pm0.2}}$	$14.9^{\pm0.1}$
	30 %	$13.8^{\pm0.2}$	$25.9^{\pm0.3}$	$24.5^{\pm0.4}$	$22.8^{\pm0.3}$	$20.9^{\pm0.3}$	$19.6^{\pm0.2}$	$\underline{18.3^{\pm0.2}}$	$\underline{16.3^{\pm0.4}}$	$\underline{15.1^{\pm0.1}}$	$\underline{13.9^{\pm0.2}}$
	35 %	$12.5^{\pm0.2}$	$24.6^{\pm0.4}$	$22.5^{\pm0.4}$	$20.9^{\pm0.2}$	$19.2^{\pm0.1}$	$18.1^{\pm0.5}$	$16.7^{\pm0.3}$	$15.3^{\pm0.3}$	$\underline{14.1^{\pm0.2}}$	$\underline{12.9^{\pm0.1}}$
	40 %	$10.8^{\pm0.1}$	$23.4^{\pm0.4}$	$21.4^{\pm0.1}$	$20.1^{\pm0.3}$	$18.3^{\pm0.5}$	$17.3^{\pm0.2}$	$16.0^{\pm0.1}$	$14.4^{\pm0.1}$	$13.2^{\pm0.2}$	$\underline{12.1^{\pm0.1}}$

BLEU		**news11**	**En WER**								
		single-src.	0 %	5 %	10 %	15 %	20 %	25 %	30 %	35 %	40 %
s-src.			$23.4^{\pm0.0}$	$21.1^{\pm0.2}$	$19.2^{\pm0.1}$	$17.1^{\pm0.0}$	$15.3^{\pm0.1}$	$13.6^{\pm0.2}$	$12.2^{\pm0.2}$	$10.6^{\pm0.3}$	$9.6^{\pm0.0}$
De WER	0 %	$22.9^{\pm0.0}$	$\underline{24.0^{\pm0.0}}$	$22.7^{\pm0.0}$	$21.4^{\pm0.2}$	$20.2^{\pm0.1}$	$18.9^{\pm0.2}$	$17.8^{\pm0.1}$	$16.9^{\pm0.1}$	$15.5^{\pm0.1}$	$14.6^{\pm0.2}$
	5 %	$20.6^{\pm0.1}$	$23.2^{\pm0.1}$	$\underline{21.8^{\pm0.1}}$	$\underline{20.7^{\pm0.0}}$	$19.2^{\pm0.0}$	$18.2^{\pm0.1}$	$17.1^{\pm0.1}$	$16.1^{\pm0.1}$	$14.8^{\pm0.0}$	$13.9^{\pm0.1}$
	10 %	$18.8^{\pm0.1}$	$22.5^{\pm0.1}$	$\underline{21.3^{\pm0.2}}$	$\underline{20.1^{\pm0.1}}$	$18.6^{\pm0.2}$	$17.7^{\pm0.1}$	$16.4^{\pm0.1}$	$15.5^{\pm0.1}$	$14.1^{\pm0.1}$	$13.2^{\pm0.2}$
	15 %	$17.0^{\pm0.3}$	$21.6^{\pm0.2}$	$20.3^{\pm0.2}$	$19.1^{\pm0.2}$	$\underline{17.8^{\pm0.1}}$	$16.9^{\pm0.1}$	$15.6^{\pm0.0}$	$14.7^{\pm0.1}$	$13.4^{\pm0.1}$	$12.5^{\pm0.1}$
	20 %	$15.4^{\pm0.2}$	$20.8^{\pm0.0}$	$19.5^{\pm0.1}$	$18.3^{\pm0.1}$	$17.0^{\pm0.2}$	$\underline{16.0^{\pm0.1}}$	$14.9^{\pm0.1}$	$14.0^{\pm0.1}$	$12.7^{\pm0.2}$	$12.0^{\pm0.1}$
	25 %	$13.8^{\pm0.1}$	$19.9^{\pm0.2}$	$18.7^{\pm0.2}$	$17.7^{\pm0.1}$	$16.3^{\pm0.1}$	$\underline{15.4^{\pm0.0}}$	$\underline{14.0^{\pm0.2}}$	$13.2^{\pm0.1}$	$11.9^{\pm0.0}$	$11.1^{\pm0.1}$
	30 %	$12.3^{\pm0.3}$	$19.2^{\pm0.3}$	$17.9^{\pm0.2}$	$16.9^{\pm0.3}$	$15.6^{\pm0.1}$	$14.5^{\pm0.2}$	$13.5^{\pm0.3}$	$\underline{12.7^{\pm0.2}}$	$11.3^{\pm0.1}$	$10.6^{\pm0.1}$
	35 %	$11.2^{\pm0.1}$	$18.4^{\pm0.0}$	$17.1^{\pm0.1}$	$16.1^{\pm0.1}$	$15.0^{\pm0.2}$	$13.8^{\pm0.1}$	$12.7^{\pm0.2}$	$11.7^{\pm0.1}$	$10.6^{\pm0.2}$	$9.9^{\pm0.2}$
	40 %	$9.9^{\pm0.3}$	$17.1^{\pm0.0}$	$16.1^{\pm0.2}$	$14.9^{\pm0.2}$	$14.0^{\pm0.1}$	$12.9^{\pm0.1}$	$11.7^{\pm0.2}$	$10.7^{\pm0.1}$	$9.9^{\pm0.1}$	$9.1^{\pm0.2}$

Table 6.10: BLEU (avg±stddev) with transcription noise on ESIC dev set whose reference translations were English and on Newstest11 with balanced reference source language. The green-backgrounded area is where the English single-source outperforms German single-source. Black underlined numbers indicate the area where multi-sourcing achieves higher scores than both single-sourcing options. In **bold** is near the maximum gap from single-source, more than 2.1 BLEU. Red-colored numbers are where at least one single-source scores higher.

6.3.2 Results with Transcription Noise

Table 6.10 summarizes the BLEU scores of two-source MT with different levels of transcription noise in each of the sources on two sets: ESIC dev with reference translated from English, and Newstest11 with balanced reference, originating in five languages. Table 6.11 contains the corresponding chrF2 scores. Table 6.12 shows the results on the ESIC test set for the settings where multi-source models achieved the highest improvement due to noisy inputs.

In Table 6.10, on both sets, we observe that the less noisy single source achieves higher BLEU than the other single source. When the difference in noise levels between the sources is small (close to diagonal in the table), then multi-sourcing reaches slightly higher BLEU than single sources. In the case of balanced Newstest11, this

	chrF2	**ESIC dev**	**En WER**								
		single-src.	0 %	5 %	10 %	15 %	20 %	25 %	30 %	35 %	40 %
	s-src.		$60.2^{\pm 0.0}$	$57.2^{\pm 0.2}$	$54.4^{\pm 0.2}$	$51.4^{\pm 0.3}$	$49.2^{\pm 0.7}$	$46.8^{\pm 0.8}$	$44.3^{\pm 0.0}$	$42.1^{\pm 0.3}$	$40.1^{\pm 0.0}$
De WER	0 %	$54.0^{\pm 0.0}$	$58.6^{\pm 0.0}$	$56.9^{\pm 0.2}$	$55.6^{\pm 0.1}$	$53.7^{\pm 0.2}$	$52.3^{\pm 0.5}$	$50.9^{\pm 0.4}$	$49.2^{\pm 0.3}$	$47.5^{\pm 0.2}$	$46.1^{\pm 0.2}$
	5 %	$51.8^{\pm 0.1}$	$57.7^{\pm 0.1}$	$56.2^{\pm 0.2}$	$54.8^{\pm 0.1}$	$52.9^{\pm 0.2}$	$51.4^{\pm 0.6}$	$50.0^{\pm 0.4}$	$48.3^{\pm 0.3}$	$46.7^{\pm 0.3}$	$45.4^{\pm 0.2}$
	10 %	$49.9^{\pm 0.2}$	$56.8^{\pm 0.2}$	$55.1^{\pm 0.1}$	$53.7^{\pm 0.3}$	$51.8^{\pm 0.3}$	$50.4^{\pm 0.3}$	$49.0^{\pm 0.4}$	$47.3^{\pm 0.1}$	$45.6^{\pm 0.2}$	$44.3^{\pm 0.2}$
	15 %	$47.6^{\pm 0.3}$	$55.8^{\pm 0.0}$	$54.2^{\pm 0.1}$	$52.8^{\pm 0.3}$	$50.9^{\pm 0.2}$	$49.6^{\pm 0.4}$	$48.1^{\pm 0.5}$	$46.4^{\pm 0.3}$	$44.9^{\pm 0.3}$	$43.6^{\pm 0.1}$
	20 %	$45.7^{\pm 0.3}$	$54.9^{\pm 0.2}$	$53.5^{\pm 0.1}$	$51.9^{\pm 0.3}$	$50.2^{\pm 0.1}$	$48.7^{\pm 0.6}$	$47.2^{\pm 0.4}$	$45.4^{\pm 0.2}$	$43.9^{\pm 0.3}$	$42.6^{\pm 0.3}$
	25 %	$44.0^{\pm 0.4}$	$54.2^{\pm 0.4}$	$52.9^{\pm 0.1}$	$51.3^{\pm 0.2}$	$49.3^{\pm 0.2}$	$48.1^{\pm 0.4}$	$46.5^{\pm 0.4}$	$44.7^{\pm 0.2}$	$43.3^{\pm 0.2}$	$41.7^{\pm 0.0}$
	30 %	$42.1^{\pm 0.3}$	$53.1^{\pm 0.3}$	$51.7^{\pm 0.3}$	$50.2^{\pm 0.2}$	$48.3^{\pm 0.3}$	$46.8^{\pm 0.3}$	$45.4^{\pm 0.5}$	$43.5^{\pm 0.3}$	$42.2^{\pm 0.2}$	$40.6^{\pm 0.1}$
	35 %	$40.5^{\pm 0.2}$	$52.0^{\pm 0.3}$	$50.0^{\pm 0.3}$	$48.7^{\pm 0.2}$	$46.9^{\pm 0.1}$	$45.7^{\pm 0.5}$	$44.1^{\pm 0.4}$	$42.4^{\pm 0.2}$	$41.0^{\pm 0.1}$	$39.7^{\pm 0.1}$
	40 %	$38.6^{\pm 0.2}$	$51.1^{\pm 0.2}$	$49.2^{\pm 0.2}$	$47.8^{\pm 0.3}$	$46.0^{\pm 0.4}$	$44.8^{\pm 0.3}$	$43.2^{\pm 0.4}$	$41.5^{\pm 0.1}$	$39.8^{\pm 0.3}$	$38.6^{\pm 0.1}$

	chrF2	**news11**	**En WER**								
		single-src.	0 %	5 %	10 %	15 %	20 %	25 %	30 %	35 %	40 %
	s-src.		$51.0^{\pm 0.0}$	$48.8^{\pm 0.2}$	$46.9^{\pm 0.1}$	$44.9^{\pm 0.1}$	$43.0^{\pm 0.1}$	$41.0^{\pm 0.1}$	$39.4^{\pm 0.1}$	$37.3^{\pm 0.1}$	$35.8^{\pm 0.0}$
De WER	0 %	$50.3^{\pm 0.0}$	$50.8^{\pm 0.0}$	$49.5^{\pm 0.0}$	$48.0^{\pm 0.1}$	$46.7^{\pm 0.1}$	$45.3^{\pm 0.2}$	$43.8^{\pm 0.3}$	$42.6^{\pm 0.2}$	$41.0^{\pm 0.1}$	$39.7^{\pm 0.1}$
	5 %	$48.3^{\pm 0.1}$	$50.0^{\pm 0.0}$	$48.6^{\pm 0.0}$	$47.3^{\pm 0.0}$	$45.7^{\pm 0.1}$	$44.5^{\pm 0.1}$	$43.1^{\pm 0.1}$	$41.8^{\pm 0.0}$	$40.1^{\pm 0.1}$	$38.8^{\pm 0.1}$
	10 %	$46.5^{\pm 0.2}$	$49.2^{\pm 0.1}$	$47.9^{\pm 0.2}$	$46.4^{\pm 0.1}$	$44.8^{\pm 0.1}$	$43.8^{\pm 0.0}$	$42.3^{\pm 0.0}$	$40.9^{\pm 0.1}$	$39.2^{\pm 0.2}$	$38.0^{\pm 0.1}$
	15 %	$44.7^{\pm 0.2}$	$48.1^{\pm 0.1}$	$46.7^{\pm 0.1}$	$45.4^{\pm 0.1}$	$43.9^{\pm 0.1}$	$42.8^{\pm 0.0}$	$41.2^{\pm 0.0}$	$40.1^{\pm 0.1}$	$38.4^{\pm 0.0}$	$37.1^{\pm 0.0}$
	20 %	$42.9^{\pm 0.1}$	$47.1^{\pm 0.0}$	$45.8^{\pm 0.1}$	$44.3^{\pm 0.0}$	$42.9^{\pm 0.1}$	$41.7^{\pm 0.1}$	$40.4^{\pm 0.1}$	$39.1^{\pm 0.0}$	$37.4^{\pm 0.1}$	$36.3^{\pm 0.1}$
	25 %	$41.1^{\pm 0.1}$	$46.1^{\pm 0.2}$	$44.8^{\pm 0.0}$	$43.6^{\pm 0.1}$	$42.0^{\pm 0.1}$	$40.8^{\pm 0.1}$	$39.3^{\pm 0.2}$	$38.1^{\pm 0.1}$	$36.4^{\pm 0.1}$	$35.3^{\pm 0.1}$
	30 %	$39.4^{\pm 0.2}$	$45.3^{\pm 0.3}$	$43.9^{\pm 0.2}$	$42.6^{\pm 0.2}$	$41.1^{\pm 0.1}$	$39.9^{\pm 0.2}$	$38.5^{\pm 0.2}$	$37.3^{\pm 0.1}$	$35.7^{\pm 0.0}$	$34.5^{\pm 0.2}$
	35 %	$38.0^{\pm 0.2}$	$44.3^{\pm 0.2}$	$42.9^{\pm 0.3}$	$41.5^{\pm 0.1}$	$40.2^{\pm 0.2}$	$38.9^{\pm 0.2}$	$37.6^{\pm 0.1}$	$36.4^{\pm 0.1}$	$34.9^{\pm 0.0}$	$33.7^{\pm 0.2}$
	40 %	$36.2^{\pm 0.2}$	$43.2^{\pm 0.2}$	$41.9^{\pm 0.2}$	$40.5^{\pm 0.2}$	$39.1^{\pm 0.1}$	$37.9^{\pm 0.1}$	$36.4^{\pm 0.2}$	$35.2^{\pm 0.1}$	$33.8^{\pm 0.2}$	$32.8^{\pm 0.2}$

Table 6.11: chrF2 (avg±stddev) with transcription noise on ESIC dev set whose reference translations were English and on Newstest11 (news11) with balanced reference source language. The area with the green background is where the English single-source outperforms German single-source. Black underlined numbers indicate the area where multi-sourcing achieves higher scores than both single-sourcing options. Red-colored numbers are where at least one single-source scores higher.

WER	**En**	**De**	**En+De**
15% En, 10% De	23.58±0.16	23.23±0.05	**26.50±0.27**

Table 6.12: ESIC test multi-sourcing vs. single-sourcing BLEU scores on the artificial WER noise level where multi-sourcing achieved the largest improvement on the dev set.

area matches the diagonal. In the case of ESIC with an English original source and reference translated from English, the area of multi-source outperforming single-source is shifted. This tendency is reflected in the test set results in Table 6.12 as well. Only when the German source is less noisy than the English one, it does improve BLEU in multi-sourcing. We explain it by the discrepancy of source languages used for MT vs. reference that affect BLEU the same way as in offline mode in Section 6.2.5. On Newstest11, with the references translated from German, we expect the reverse. We inspect the effect of reference source language in Section 6.4.4.

We also observe the expected behavior that the more noise, the lower BLEU in all setups. Compare e.g. 33.3 BLEU with zero noise and 12.1 with 40% WER in both sources. With very large noise, it is possible that neither option would be usable. In ESIC dev, e.g. when English WER is 20%, we observe a large span, between 5 and 25% WER in German, where multi-sourcing outperforms single source at least by several hundreths of BLEU. This span in Newstest11 is much narrower, only 20 to 25% WER in German. We hypothesize that it may be caused by the domain difference. The lexical noise model is trained on Europarl. In the news domain, there may be fewer words for substitution, so the noise consists more of deletions and insertions, and it might be more harmful for MT in the combination of two sources. However, multi-sourcing appears to be robust to ASR errors regardless of whether we have one or both sources as original.

chrF2 scores There is evidence that chrF2 correlates with human judgments better than BLEU. In Table 6.11, we see that for multi-sourcing with noisy inputs on ESIC dev, chrF2 are indeed higher than single-sourcing and this correlates with the BLEU score gains in Table 6.10. On the other hand, for Newstest11, chrF2 scores do not indicate any improvements. While the corresponding BLEU scores in Table 6.10 indicated improvements of multi-sourcing with noisy inputs, the magnitudes of these gains were minor, much smaller than those observed for ESIC. This gives us sufficient reason to believe that multi-sourcing should be useful in a setting like ESIC, where the reference is created from only one source, which is more realistic than the "balanced" use-case of Newstest11, where the reference originates from 5 languages.

6.4 Simultaneous Multi-Sourcing

In the previous sections, we experimented with offline translation with artificial ASR noise and showed that multi-source models are indeed robust to noise. However, one important use case of speech translation is in a real-time setting where simultaneous MT is used. We therefore adapt our offline multi-source NMT model for simultaneous mode and evaluate its robustness to transcription noise. Again, we use a simplified setup. We simulate SST on text sources that are available one token at a time, or one token in each of the multiple sources. Instead of simultaneous interpreting, we use parallel revised translations in ESIC and Newstest11. Again, we simulate the ASR noise by artificial errors inserted to texts by lexical noise model as in Section 6.3.

Focus on streaming For this and further experiments in this thesis, we decided to use streaming, and not re-translating MT (recall Section 3.1.2), because the most recent research advancements, e.g. in IWSLT (Anastasopoulos et al., 2022; Agarwal et al., 2023; Fukuda et al., 2023; Papi et al., 2023a), in Papi et al. (2023b,c); Dugan et al. (2023); Polák et al. (2023) and many other recent works focus on streaming. We plan to follow up on their findings. Fortunately, we are able to use streaming in the ELITR infrastructure that is designed for re-translation. It is possible because streaming meets the conditions that are defined for re-translation.

6.4.1 Creating Simultaneous MT

Adapting MT to simultaneity Simultaneous MT can be created from any standard text-to-text NMT by applying a simultaneous decoding algorithm. However, it is recommendable first to adapt NMT to be inclined to translate consecutive sentence prefixes with the same target prefix. We use Local Agreement (LA-n) as a decoding streaming algorithm. It achieved good performance by the best-performing system CUNI-KIT (Polák et al., 2022) in the recent IWSLT competition (Anastasopoulos et al., 2022). Local Agreement (LA-n) means that n consecutive updates must agree on a target prefix to commit it and write it out. The last committed prefix is then forced as a prefix to decode the next units. Agreement level n is a parameter that controls the latency.

Finetuning for stable prefixes In Section 6.2, we used multi-way models trained on full sentences, but in a simultaneous setting, these models will tend to artificially finish sentences when translating partial ones using the LA-n approach. Therefore, our multi-way models should first be adapted for partial sentence translation. To this end, we used the multi-way English and German to Czech MT model as a base for simultaneous MT. We fine-tuned the last trained model checkpoint for stable translation on a 1:1 mix of incomplete sentence prefixes and full sentences as Niehues et al. (2018). For each source-target pair of the training data, we selected 5 times 1 to 90 % of source and target characters and rounded them to full words. Then, we ran training for 1 day on 1 GPU. We validated the BLEU score on ESIC dev and

Normalized Erasure (NE, Arivazhagan et al., 2020b) on all prefixes of the first 65 sentences (around 1 500 words) of ESIC dev set. We ran fine-tuning with multi-way data for English and German as source languages, and for bilingual English-Czech and German-Czech MT.

We stopped training after one day when there were no improvements in stability (NE) or quality (BLEU). Then, we selected one checkpoint for English and one for German that reached acceptable quality and stability values. The checkpoints that we selected for simultaneous multi-source decoding was the multi-way checkpoint for {English,German}→Czech and bilingual one for German→Czech. Table 6.13 summarizes the results of fine-tuning for stability. BLEU decreased marginally (by 0.2 on English and 0.9 on German), while Normalized Erasure (NE) dropped by 40% in English and 52% on German.

Based on some outputs, we explain higher NE in German-to-Czech by word order differences. Many erasures were caused by an incorrect presumption of the final verb. Regardless, our fine-tuned models exhibit significantly reduced NE and can be reliably used for simultaneous translation using the LA-n approach.

	En		De	
checkpoint	BLEU	NE	BLEU	NE
starting	33.2	1.77	25.9	3.15
selected	33.0	1.21	25.0	1.52
diff	-0.2	-40%	-0.9	-52%

Table 6.13: The results of fine-tuning for stability, on ESIC dev. NE stands for "Normalized Erasure" (Arivazhagan et al., 2020b), measure of stability of re-translating simultaneous MT.

6.4.2 Multi-Sourcing in Simultaneous MT

We use late averaging of the two selected checkpoints for multi-sourcing in simultaneous MT. The only aspects of multi-sourcing in the simultaneous mode that differ from single-source or non-simultaneous mode are 1) synchronization of sources and 2) how to measure latency with Average Lagging.

Synchronization In a realistic use case, it is necessary to synchronize the original speech and simultaneous interpreting. However, we leave it for further work, as our goal is to inspect the limits of multi-sourcing. Therefore, we simulate a case where the sources are optimally synchronized, aligned, and parallel to the sentence level.

LA update	En update	En source	De source	→ Cs target
1	1	<lang:en> The	<lang:de>	→
2			Die	→
3			Ausnahme	→
4	2	derogation		→
5			sollte	→Výjimka
6	3	should		→
7			die	→ by měla
8	4	instead		→
9			allgemeine	→ být
10	5	be		→
11	6	a		→
12	7	general		→ obecná
13			Regel	→
14			sein.	→
15	8	rule.		→ pravidla.

Figure 6.8: Example synchronization of two sources for streaming SST with LocalAgreement-2. The LocalAgreement considers all updates (left-most column), while Average Lagging, latency measure, considers only English updates (second column). The Czech multi-source MT target contains wrong inflection "obecná pravidla."

In multi-source mode, we sort all sentence prefixes by proportion of the character length to the sentence length. Each "Read" operation of the multi-source system then receives two prefixes in two languages. One of them is updated by one new token, as illustrated in Figure 6.8. Every such update is counted to the local agreement size. We note that there are other strategies, e.g. count only English source updates to LA-n, but in this experiment, we have another goal than searching for the best strategy.

Synchronizing noise For synchronizing the sources with WER noise, we implemented an option for the WER model that produces a tag indicating word deletion and its character length. The deleted word contributes to synchronization but does not appear in the noised source at all. Instead, there is no word on its position, simulating that the ASR produced silence. Without this, we would have an unrealistic simulation, the deletions could make succeeding words appear early.

Average Lagging in multi-source In a multi-source setup, we only count Read operations of the English source to Average Lagging calculations that we report, and not of the German source because the sources are simultaneous. Counting only German tokens differs negligibly, approximately by 0.1 tokens.

6.4.3 Simultaneous Multi-Source with Artificial Noise

We want to compare the multi-sourcing model to single-sourcing with artificial ASR noise model as in Section 6.3. We evaluate each system on the latency levels with local agreement sizes 2, 5, 10 and 15. Since each evaluation on 2 000 sentences takes approximately 5 hours, we report only one run, and not average and deviation on multiple randomly noised inputs.

The results on ESIC dev set are in Figure 6.9. We can observe the same trends as in the offline case. The single source that is noised less achieves higher BLEU. Multi-sourcing outperforms both single sources when both noise levels are similar and when the English is noisier, e.g. in the case with 10% WER in German and 20% WER in English. We explain it again by the fact that the Czech reference is translated from English, and not German.

Furthermore, on both ESIC and Newstest11 (Figure 6.9) we observe that multi-sourcing performs worse in the low-latency modes, i.e. in AL<5 that roughly corresponds to LA<5. We assume that the proportional synchronization of the two sources is often inaccurate and may confuse late averaging. In higher latency modes, the synchronization noise at the end of input may be lowered by local agreement. Having validated that the multi-source NMT is robust to ASR errors in both full sentence and simultaneous settings, we have paved the way for harder settings where multi-lingual interpretations of the original source available with different amounts of delay can be used for translation.

6.4.4 Effect of Reference Source Language

To explain the effect of reference source language, we run a contrastive evaluation on the subset of Newstest11 that consists only of the documents that originate in English. We compare BLEU measures with a reference translated directly from Czech, and with three additional references translated only from German (Bojar et al., 2012).

The results of the simultaneous mode are in Figure 6.10. We observe the same trends as in offline mode in Section 6.3. The BLEU score is higher for the single source with the language from which the reference was translated. When this source is noised substantially more than the other, multi-sourcing outperforms both by a small margin.

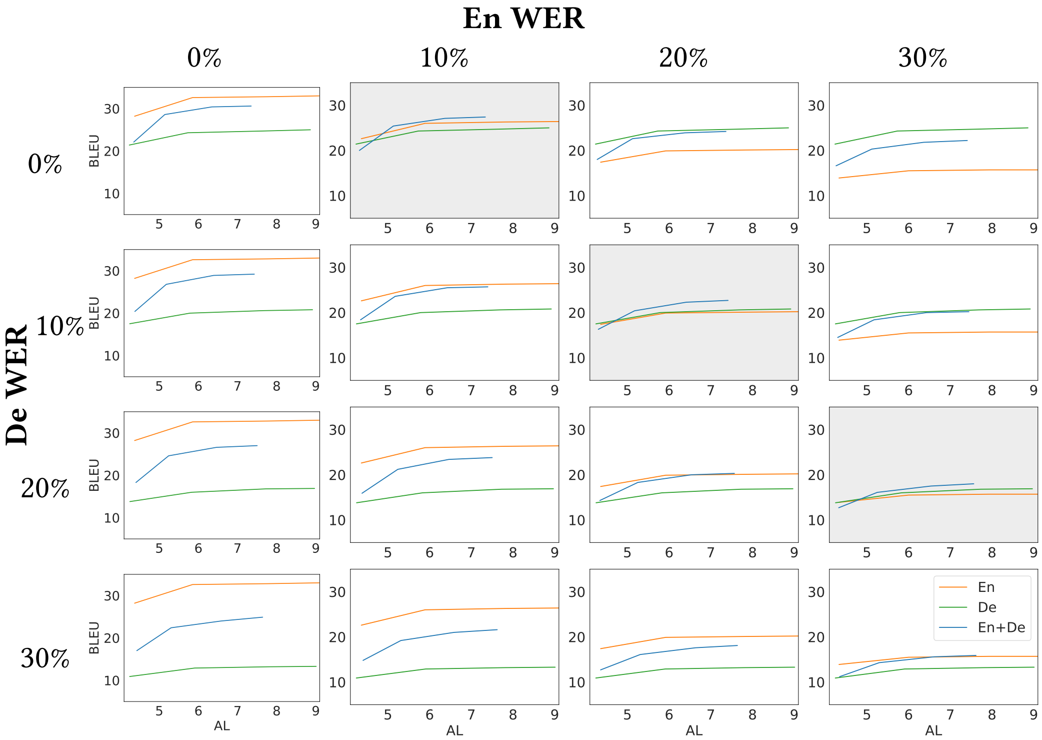


Figure 6.9: Single-sourcing vs. multi-sourcing with different levels of artificial ASR noise of the sources (% WER) in simultaneous mode on ESIC dev set. The results are depicted as quality (BLEU) and latency (AL) trade-offs of the candidate systems. The plots highlighted by gray background show noise levels where multi-sourcing (En+De, blue line) outperforms both single sources in BLEU at least for AL>5.5.

In the case of German references, the nearest margin to single-sourcing is much smaller than with the English references. We assume it is because the structural difference between the English source and German-Czech references is larger than the German source and English-Czech references. This is documented also by BLEU scores with zero noise (33 and 20 on references from English vs. 16 and 30 on references from German).

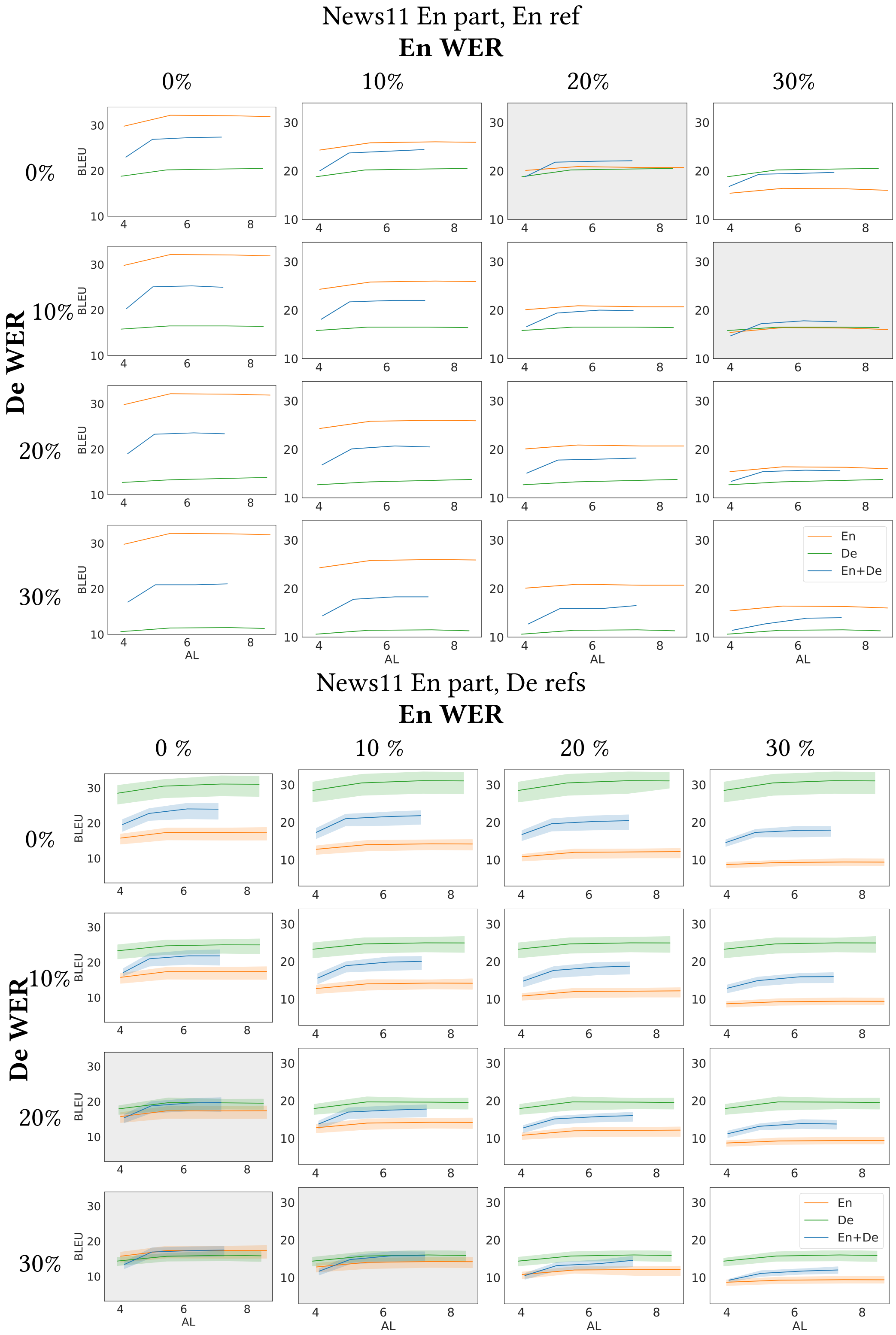


Figure 6.10: Single-sourcing vs. multi-sourcing with different levels of artificial ASR noise of the sources (% WER) in simultaneous mode on Newstest11 subset (598 sentences) originally in English. In the upper grid, the Czech reference is translated from English, while in the lower, there is an average and standard deviation of single-reference BLEU counted against each of the the 3 additional references translated from German (Bojar et al., 2012). Grey highlighting indicates noise levels where multi-sourcing (En+De, blue line) outperforms or is on par with both single sources in BLEU.

6.5 Summary

We investigated the robustness of multi-source SST to transcription errors in order to motivate its use in settings where ASRs for the original speech and parallel simultaneous interpreting are available.

For this, we first analyzed the 10-hour ESIC corpus and documented that the ASR errors in the two sources are indeed independent, indicating their complementary nature. We then simulated transcription noise for English and German when translating into Czech in single and multi-source NMT settings and observed that using multiple noisy sources is significantly better than individual noisy sources. We then repeated experiments in a simultaneous translation setting and showed that multi-source translation continues to be robust to noise. This robustness of multi-source NMT to noise motivates future research into simultaneous multi-source speech translation, where one source is available with a delay.

The limitation of the experiments in this chapter is the simplification of realistic conditions. We analyzed multi-sourcing using parallel sentence-segmented texts and artificial errors inserted by the lexical noise model. Multi-sourcing with realistic ASR applied on parallel speech, original and experimenting, is left for the next chapters and for future work, as well as the question of how to deal with interpreting delay and the fact that it is not fully parallel in meaning to original.

7 Evaluation Questions

When researching multi-lingual simultaneous speech translation (SST) from the original and parallel simultaneous interpreting (SI), we had to consider several questions regarding evaluation. The first one is very practical. We have the ESIC corpus (Section 4.5) that we use for SST into Czech. There are two parallel Czech versions: revised normalized translations, and SI. What should we use as a reference for SST; translation, or interpreting?

This leads us to a more general question. What should optimal SST be like? More like translation, or interpreting, a combination of both, or something else?

First of all, we leave the investigation of other alternatives than translation and interpreting to further work because it is not in our primary scope. We only note that imitating human performance in translating or interpreting by the machine does not need to be the best goal anyway because performance of humans is often suboptimal (Kloudová et al., 2023). However, we investigate the two options, translation and interpreting, because we plan to apply our findings in our next research for developing multi-source SST in realistic conditions.

The advantage of offline text translation over SI is the fact that the translators can afford more time, effort, and consultation with external resources and translation memories because, unlike the simultaneous interpreters, they are not limited by time constraints and by the speech output modality that does not allow post-editing and revisions. The translation, in contrast to SI, is usually more faithful, grammatically correct, and the translator has the capacity to decide on and maintain a certain level of literalness. On the other hand, as we found in Macháček et al. (2021) and in Sections 5.1.1 and 5.1.2, SI tends to use simpler vocabulary and is usually shorter. It may be better understandable and therefore more preferable by end users.

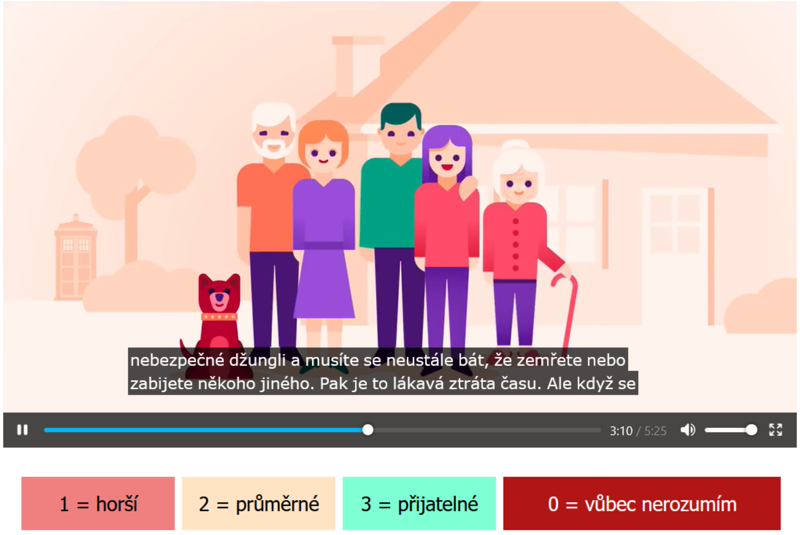

nebezpečné džungli a musíte se neustále bát, že zemřete nebo
zabijete někoho jiného. Pak je to lákavá ztráta času. Ale když se

Figure 7.1: A preview of the Continuous Rating session setup. There is a video at the top, overlaid with two lines of subtitles in Czech produced by SST, followed by buttons for Continuous Rating. The button labels are 1: Worse; 2: Average; 3: Good; 0: I do not understand at all. Figure reprinted from Javorský et al. (2022).

Our next question is how to make SST evaluation efficient and accurate, so that it can be easily, frequently, and reliably used in the SST development. The current standard is to adopt the evaluation metrics from the offline text-to-text machine translation (MT) or speech translation (ST), and assume that they are reliable, despite that the simultaneity makes SST different from MT and ST.

To answer these questions, we were involved in proposing Continuous Rating (CR, Macháček and Bojar, 2020), a method for collecting human ratings of SST on simulated live events, and in the study for assessing its reliability (Javorský et al., 2022). We describe CR in Section 7.1. Then, we assume that knowledgeable human experts will give their quality preference using CR. In Section 7.2, we analyze automatic MT metrics by comparing them to human ratings and assess their reliability in SST. In Section 7.3, we find the most reliable automatic evaluation method and answer the question of translation or interpreting reference.

Sections 7.2 and 7.3 contain text, figures and tables that we previously published in the paper "MT Metrics Correlate to Human Ratings of Simultaneous Speech Translation" (Macháček et al., 2023a).

7.1 Continuous Rating

Continuous Rating (CR) is a method where SST is evaluated in a simulated online event. Human evaluators watch SST generated subtitles placed over video (or only with audio) document and they are asked to continuously express their satisfaction by pressing rating buttons, e.g. every 20 seconds. CR was first proposed by Ondřej Bojar in our joint work (Macháček and Bojar, 2020). We implemented it and used it for human evaluation of the size of the subtitling window for re-translating SST.

Later, our colleague Dávid Javorský implemented a framework for CR evaluation as a web application (preview in Figure 7.1) and performed a study with human evaluators. One of the goals of the study was to assess the reliability of CR by contrasting it to the level of understanding the document content. The level of understanding was assessed by factual questionnaires that the evaluators filled out after each rating. We collaborated on the study mostly by consultation, result analysis, and writing the paper (Javorský et al., 2022).

The results show that the evaluators with advanced knowledge of the source language are reliable to assess SST quality with CR, and that the factual questionnaires are not necessary. This saves a considerable amount of effort because the questionnaires are not easy to prepare and evaluate.

CR is analogous to Direct Assessment (Graham et al., 2015), a method of human text-to-text MT evaluation in which a bilingual evaluator expresses the MT quality by a number on a scale. However, CR is directly designed for end-to-end simultaneous evaluation, enabling to cover all the aspects of simultaneity. Every evaluator sees a document for their first time when they are rating a system on it, to avoid expectations of the system outputs. They also see the system outputs with the same timing as in real-time. They can not make a pause during rating. They receive the corresponding source as the original speech, not as transcription that can omit the paralinguistics, and they access the document continuously, from the beginning to the end, without skipping to the future or past context, which is not possible in the real-time event.

7.2 MT Metrics in SST

CR requires human labor, and therefore it is not suitable for frequent evaluation during SST development. For that, standard MT metrics such as BLEU (Papineni et al., 2002), chrF2 (Popović, 2017), BertScore (Zhang et al., 2020b), COMET (Rei et al., 2020) and others (Freitag et al., 2022) are used to assess the quality of SST system, along the measures for latency. Figure 7.2 shows typical quality and latency comparison of SST system. However, the MT metrics are designed primarily for segment-level text-to-text MT. There are challenges of SST evaluation that are not included in MT metrics design: simultaneity, speech source modality, and document-level phenomena. We illustrate them in Figure 7.3.

The MT metrics were[1] used for quality assessment of SST, despite there was no evidence that they correlate to human ratings in simultaneous mode. Such a standard stems from the belief that the translation quality is currently the most critical issue of SST, and the MT metrics capture it sufficiently, as evaluated in WMT Metrics tasks (Freitag et al., 2022).

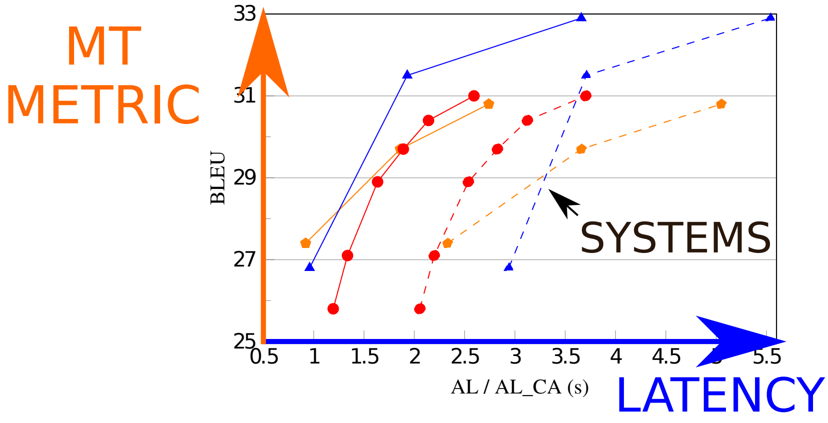


Figure 7.2: Illustration of typical comparison of SST candidate systems by their quality and latency. Lines inside the plot represent the candidate systems that are evaluated with certain configurations whose quality and latency is empirically observed and plotted as the data points. The quality is estimated by a MT metric (vertical axis), the latency measure is on the horizontal axis. The plot is reprinted from Papi et al. (2023a).

7.2.1 Continuous Rating in IWSLT22

To rigorously test the hypothesis that the MT metrics are reliable for SST quality assessment, we used the CR data that were collected in IWSLT 2022 English-to-German Simultaneous Translation Task, which is described in "IWSLT 2022 Findings" (Anastasopoulos et al., 2022). The task focused on speech-to-text translation and was reduced to the translation of individual sentences. The segmentation of the source audio to sentences was provided by organizers, and not by the systems themselves. The

[1] In late 2022, when we worked on this issue.

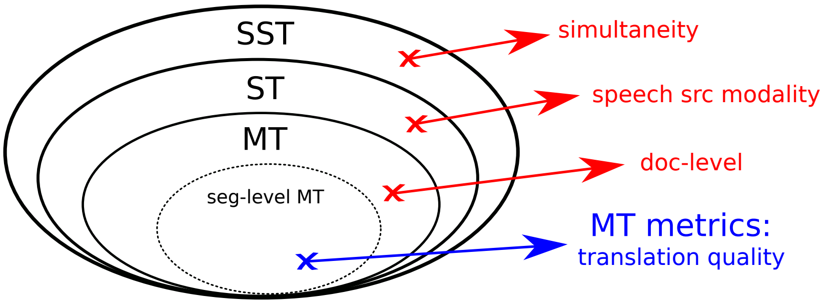


Figure 7.3: Illustration of simultaneous speech translation (SST) challenges that are not included in the design of MT metrics. We see SST as a superset of offline speech translation (ST). The simultaneity lies in the difference. Then, we see ST as a superset of offline text-to-text machine translation (MT). Speech as source modality is in the difference. MT is a superset of the segment-level MT, which is designed to translate individual sentences, not documents that require document-level consistency that lies in the difference. The MT metrics are designed primarily for the segment-level MT, where the primary concert is the translation quality.

source sentence segmentation that was used in human evaluation was gold (oracle). It only approximates a realistic setup where the segmentation would be provided by an automatic system, e.g. Tsiamas et al. (2022), and may be partially incorrect, causing more translation errors than the gold segmentation.

The simultaneous mode in the Simultaneous Translation Task means that the source is provided gradually, one audio chunk at a time. After receiving each chunk, the system decides to either wait for more source context or produce target tokens. Once the target tokens are generated, they can not be rewritten.

The participating systems are submitted and studied in three latency regimes: low, medium, and high. It means that the maximum Average Lagging (Ma et al., 2019) between the source and target on the validation set must be 1, 2, or 4 seconds, respectively, in a "computationally unaware" simulation that includes counting the time for waiting for the source context, but not the time spent by computation, which is dependent on the hardware and implementation optimization. It enables easy comparison of the SST systems as algorithms. The alternative is "computationally aware" latency.

One system in the low latency did not pass the latency constraints (see IWSLT 2022 Findings, page 44, numbered 141), but it is manually evaluated regardless.

The computationally unaware latency was one of the main criteria in IWSLT 2022. It means that the algorithmic aspect of the SST systems was the priority, so the participants did not need to focus on a low-latency implementation. However, the subtitle timing in the manual evaluation was created in a way such that waiting for the first target token was dropped, and then it continued with computationally aware latency.

Criteria of CR In IWSLT 2022, the evaluators were instructed that the primary criterion in CR should be meaning preservation (or adequacy), and other aspects such as fluency should be secondary. The instructions do not mention readability due to output segmentation frequency or verbalizing non-linguistic sounds such as "laughter," despite the system candidates differ in these aspects.

Automatic SST systems There are 5 evaluated SST systems: FBK (Gaido et al., 2022), NAIST (Fukuda et al., 2022), UPV (Iranzo-Sánchez et al., 2022), HW-TSC (Wang et al., 2022), and CUNI-KIT (Polák et al., 2022).

Human SI In order to compare the state-of-the-art SST with human reference, the organizers hired one expert human interpreter to simultaneously interpret all the test documents. Then, they employed annotators to transcribe the voice into texts. The annotators worked in the offline mode. The transcripts were then formed as subtitles including the original interpreter's timing and were used in the CR evaluation the same way as SST. However, the human interpreters use their own segmentation to translation units so they often do not translate one source sentence as one target sentence. There is no gold alignment of the translation sentences to interpreting chunks. This alignment has to be resolved before applying automatic metrics to interpreting.

Evaluation data There are two subsets of evaluation data used in IWSLT22 En-De Simultaneous Translation task. The "Common" subset consists of TED talks of the native speakers. The "Non-Native" subset consists of mock business presentations of European high school students (Macháček et al., 2019, it is "AntreCorp" discussed in Section 4.3), and of presentations by representatives of two European supreme audit institutions.

7.2.2 Correlation of CR and MT Metrics

Metrics First, we study the correlation of CR and MT metrics BLEU, chrF2, BertScore and COMET using translation reference. BLEU and chrF2 are based on lexical overlap and are available for any language. BertScore (Zhang et al., 2020b) is based on embedding similarity of a pre-trained BERT language model. COMET (Rei et al., 2020) is a neural metric trained to estimate the Direct Assessment (Graham et al., 2015) style of human evaluation.

COMET requires sentence-to-sentence aligned source, translation, and reference in the form of texts, which may be unavailable in some SST use cases. BertScore and COMET are also available only for a limited set of languages. Therefore, there may be use cases in which the most recommendable metrics for MT, the ones that correlate most to human ratings, are not available. Then the other, less correlating, but more versatile metrics can be considered as fallback options. Therefore, we analyze multiple metric types.

We use sacreBLEU (Post, 2018) for BLEU and chrF2 computation,[2] BertScore with the original implementation[3] and COMET (Rei et al., 2020) with `wmt20-comet-da` model.

Aggregation level We measure the correlation at the level of documents, and not at the test set level in order to increase the number of observations for significance tests. There are 60 evaluated documents (17 in the Common subset and 43 in Non-Native) and 15 system candidates (5 systems, each in 3 latency regimes), which yields 900 data points.

Pre-processing We discovered that CUNI-KIT system outputs are tokenized, while the others are detokenized. Therefore, we first detokenized CUNI-KIT outputs. Then, we removed the final end of sequence token (`</s>`) from the outputs of all systems.

Aggregating CR In total, there are 1584 rating sessions of the 900 candidate document translations. Each candidate document translation is rated either twice with different evaluators, once, or not at all. We aggregate the individual rating clicks in each rating session by plain average to get the CR scores for each rating session. Then, we average the CR across multple sessions of the same documents and candidate translations, and we correlate it with MT metrics.

[2] Metric signatures: BLEU|nrefs:1|case:mixed|eff:no|tok:13a|smooth:exp|version:2.3.1, chrF2|nrefs:1|case:mixed|eff:no|tok:13a|smooth:exp|version:2.3.1

[3] F1 score, signature bert-base-multilingual-cased_L9_no-idf_version=0.3.12(hug_trans=4.23.1)_fast-tokenizer

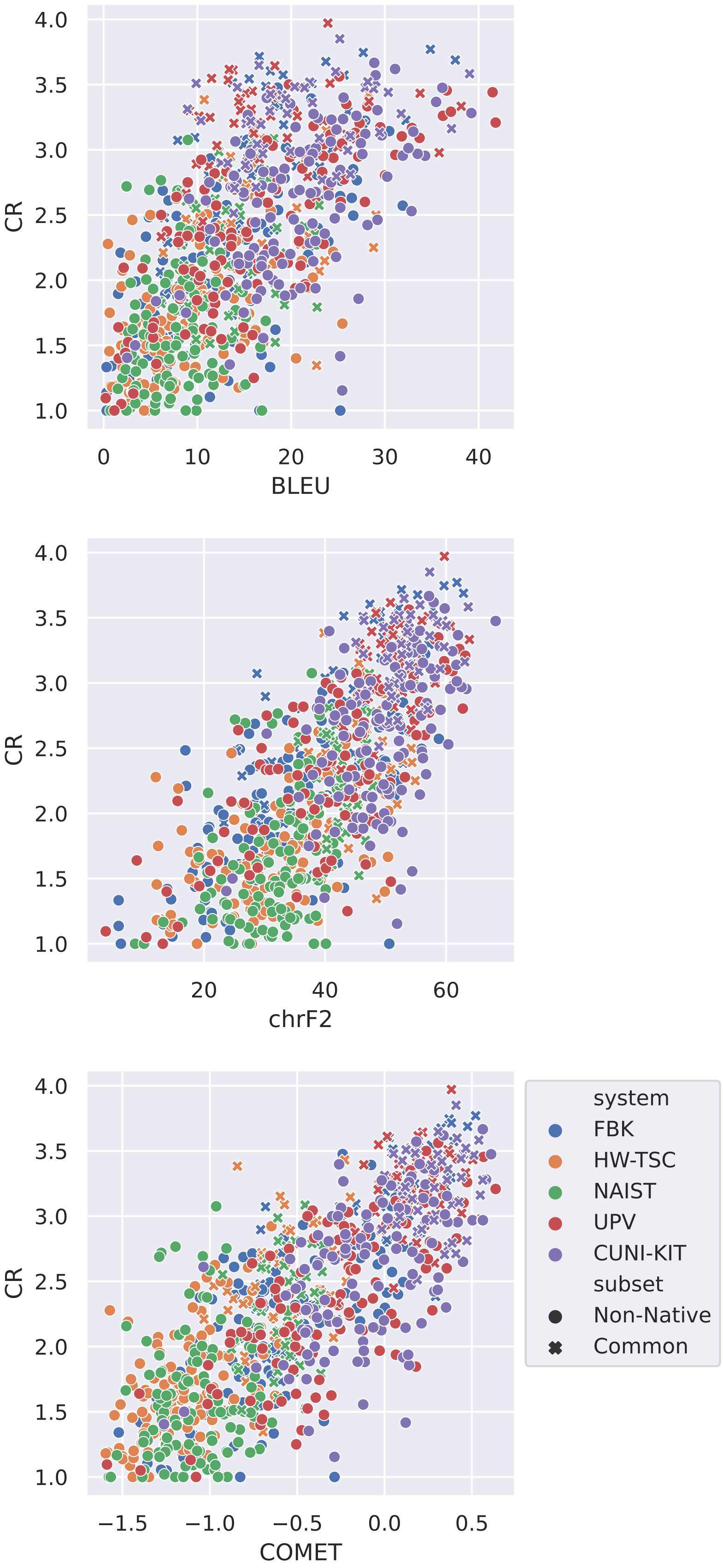


Figure 7.4: Averaged document CR vs. MT metrics BLEU, chrF2 and COMET on both subsets.

Averaged document ratings

subsets	num.	BLEU	chrF2	BertS.	COMET
both	823	0.65	0.73	0.77	0.80
Common	228	*0.42*	0.63	0.68	0.76
Non-Native	595	0.70	0.70	0.73	0.75

All document ratings

subsets	num.	BLEU	chrF2	BertS.	COMET
both	1584	0.61	0.68	0.71	0.73
Common	441	*0.37*	*0.57*	0.60	0.68
Non-Native	1143	0.64	0.64	0.66	0.67

Table 7.1: Pearson correlation coefficients for CR vs. MT metrics BLEU, chrF2, BertScore and COMET for averaged document ratings by all 5 SST systems and 3 latency regimes (upper), and all ratings (lower). When the coefficient is less than 0.6 (in *italics*), the correlation is not considered as strong. Significance values are $p < 0.01$ in all cases, meaning strong confidence.

Correlation Results In Table 7.1, we report correlation coefficients with and without averaging, together with the number of observations. Figure 7.4 displays the relation between CR and COMET.

Pearson correlation is considered as strong if the coefficient is larger than 0.6 (Evans, 1996). The results show a strong correlation (above 0.65) of CR with BLEU, chrF2, BertScore, and COMET at the document level on both test subsets. When we consider only one subset, the correlation is lower, but still strong for chrF2, BertScore, and COMET (0.63, 0.68, and 0.76, resp.). It is because the Common subset is generally translated better than Non-Native, so with only one subset, the points span a smaller part of the axes and contain a larger proportion of outliers.

The strong correlation is not the case of BLEU on the Common subset where the Pearson coefficient is 0.42. We assume it is because BLEU is designed for use on a larger test set, but we use it on short single documents. However, BLEU correlates with chrF2 and COMET (0.81 and 0.62 on the Common subset). BLEU also correlates with CR on the level of test sets, as reported in the Findings in the caption of Table 18 (page 48, numbered 145).

Conclusion We conclude that with the current overall levels of speech translation quality, BLEU, chrF2, BertScore, and COMET can be used for reliable assessment of human judgment of SST quality at least at the level of test sets. chrF2, BertScore and COMET are reliable also at the document level.

7.3 Translation or Interpreting Reference

In the works that assess the reliability of MT metrics in offline text-to-text MT, it is assumed that knowledgeable human experts reliably assess and express the MT quality. In Javorský et al. (2022), we showed that the CR of the bilingual evaluators correlates with the SST users' understanding. Therefore, we assume that CR can serve as golden truth of the SST quality, and we use CR to find the automatic evaluation method that is the nearest approximation of CR. It is the one that has the highest correlation to CR, similarly as in the text-to-text MT (Papineni et al., 2002; Freitag et al., 2022).

Moreover, for every metric we can find the optimal reference. We consider human translation (transl) and transcript of simultaneous interpreting (intp) as two possible references, and also multi-reference metrics with both.

Since interpreting is not sentence-aligned to SST candidate translations, we consider two alignment methods: single sequence (SingleSeq), and mWERSegmenter (Matusov et al., 2005, mWER). SingleSeq method means that we concatenate all the sentences in the document to one single sequence, and then apply the metric to it as if it was one sentence. mWERSegmenter is a tool for aligning translation candidates to reference if their sentence segmentation differs. It finds the alignment with the minimum WER when comparing tokens in aligned segments. For translation, we also apply the default sentence alignment (Sent).

7.3.1 Results

In Table 7.2, we report the correlations for different metric, reference and alignment variants.

For every pair of metric setups, we test the statistical significance of the difference in their correlation to CR. The test can help determine whether one method is significantly, or only slightly better than another one. If the difference is not significant, then another criterion can be used for selection, e.g. a simpler method or the method that is more often used in other works may be preferred.

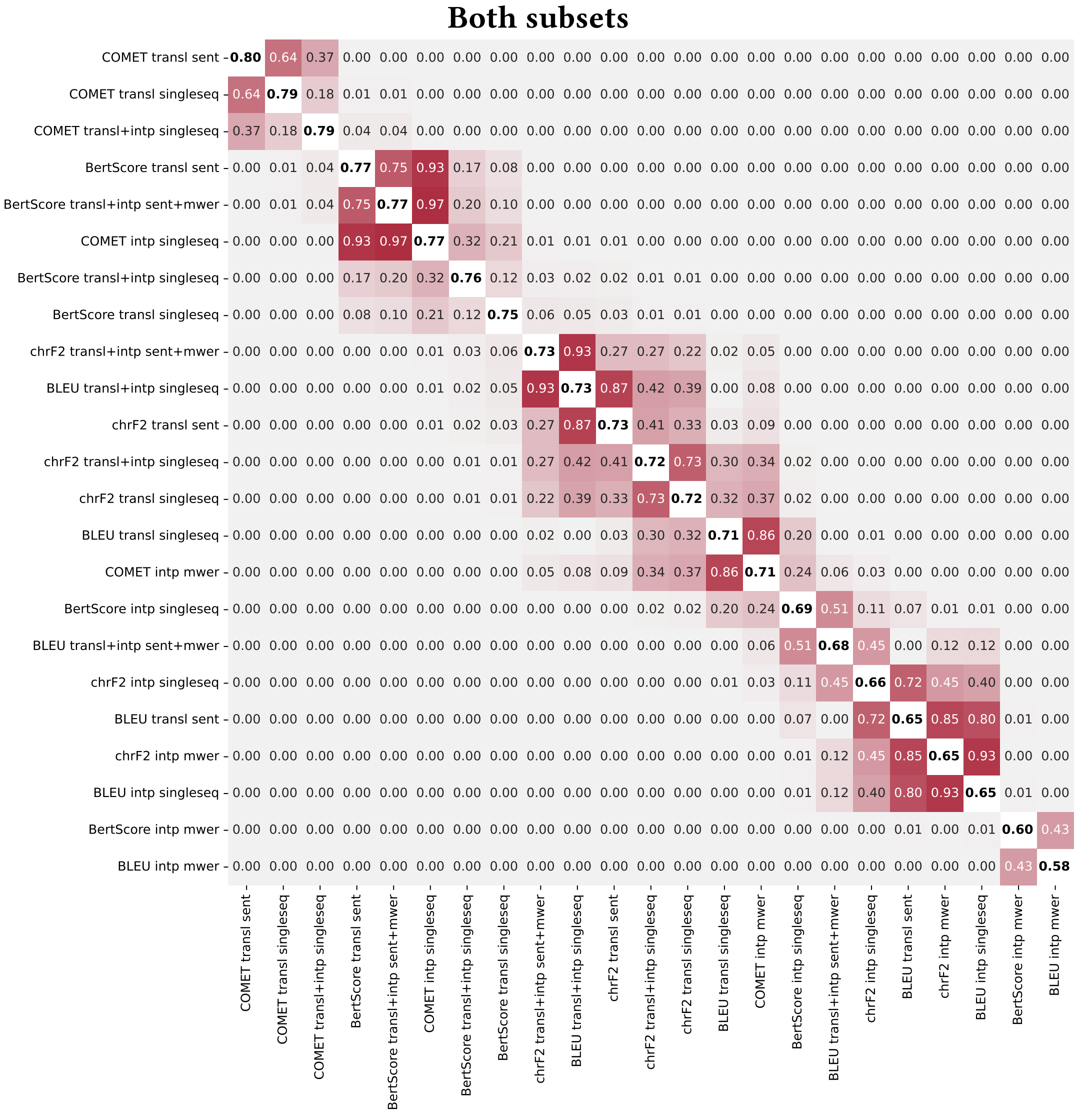
To test the significance of correlation differences, we use Steiger's method.[4] The method takes into account the number of data points and the fact that all three compared variables correlate, which is the case of the MT metrics that are applied to the same texts. We use a two-tailed test.

[4] https://github.com/psinger/CorrelationStats/

metric	reference	alignment	corr.
COMET	transl	Sent	0.80
COMET	transl	SingleSeq	0.79
COMET	transl+intp	SingleSeq	0.79
BertScore	transl	Sent	0.77
BertScore	transl+intp	Sent+mWER	0.77
COMET	intp	SingleSeq	0.77
BertScore	transl+intp	SingleSeq	0.76
BertScore	transl	SingleSeq	0.75
chrF2	transl+intp	Sent+mWER	0.73
BLEU	transl+intp	SingleSeq	0.73
chrF2	transl	Sent	0.73
chrF2	transl+intp	SingleSeq	0.72
chrF2	transl	SingleSeq	0.72
BLEU	transl	SingleSeq	0.71
COMET	intp	mWER	0.71
BertScore	intp	SingleSeq	0.69
BLEU	transl+intp	Sent+mWER	0.68
chrF2	intp	SingleSeq	0.66
BLEU	transl	Sent	0.65
chrF2	intp	mWER	0.65
BLEU	intp	SingleSeq	0.65
BertScore	intp	mWER	0.60
BLEU	intp	mWER	0.58

Table 7.2: Pearson correlation of metric variants to averaged CR on both subsets, ordered from the most to the least correlating ones. Lines indicate "clusters of significance," i.e. boundaries between groups where all metric variants significantly differ from all in the other groups, with $p < 0.05$ for dashed line and $p < 0.1$ for dotted line. See the complete pair-wise comparison in Figures 7.5 to 7.7.

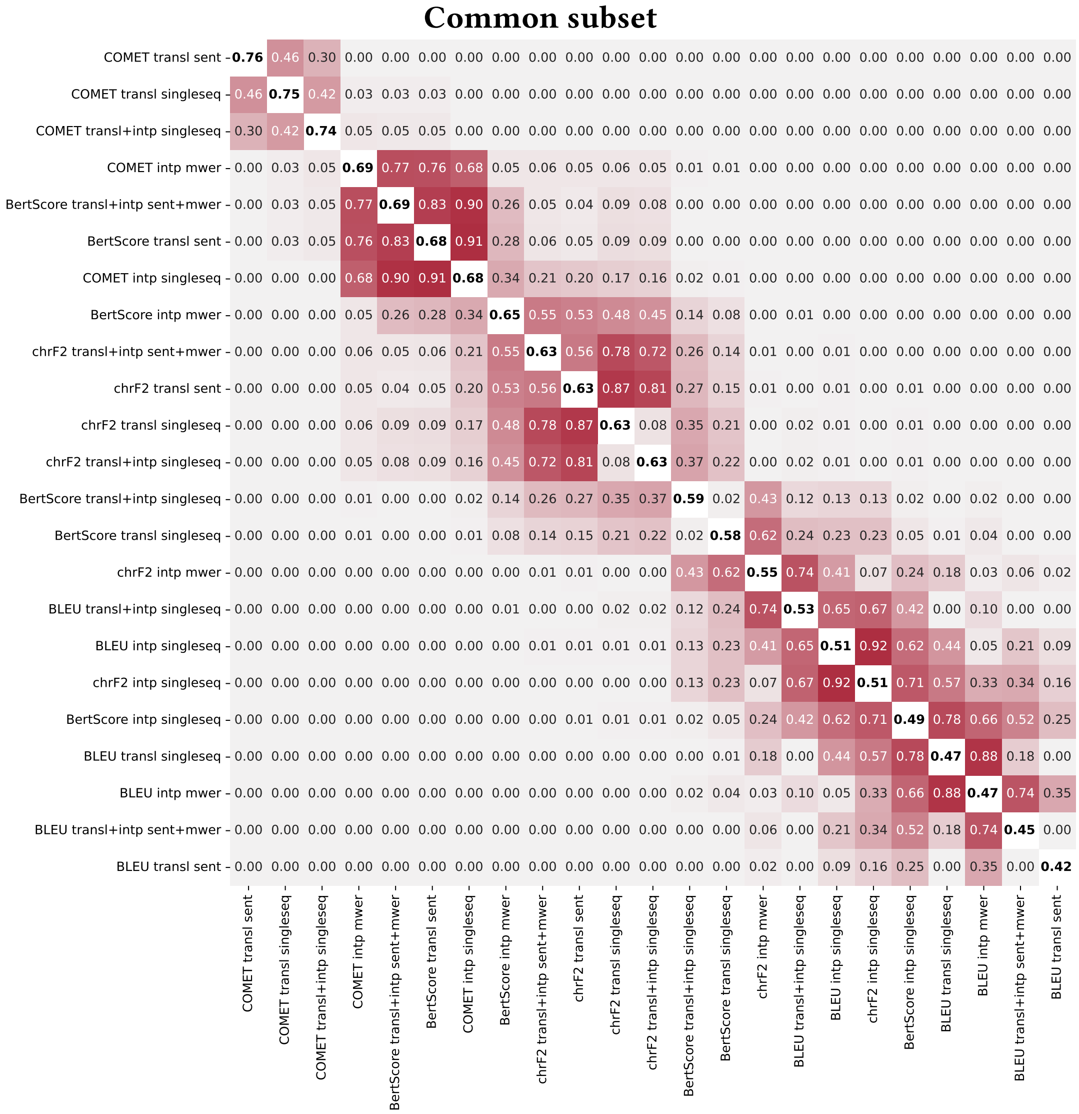
The results of significance tests for both subsets are in Figure 7.5. Figure 7.6 displays results on the Common subset, and Figure 7.7 for the Non-Native subset. These results are analogous to those in Table 7.1 in Section 7.2.2. The correlation scores for the two subsets treated separately are lower and the differences along the diagonal are less significant. We explain it by the fact that in the smaller dataset, there is a larger impact of noise.



Both subsets

	COMET transl sent	COMET transl singleseq	COMET transl+intp singleseq	BertScore transl sent	BertScore transl+intp sent+mwer	COMET intp singleseq	BertScore transl+intp singleseq	BertScore transl singleseq	chrF2 transl+intp sent+mwer	BLEU transl+intp singleseq	chrF2 transl sent	chrF2 transl+intp singleseq	chrF2 transl singleseq	BLEU transl singleseq	COMET intp mwer	BertScore intp singleseq	BLEU transl+intp sent+mwer	chrF2 intp singleseq	BLEU transl sent	chrF2 intp mwer	BLEU intp singleseq	BertScore intp mwer	BLEU intp mwer
COMET transl sent	**0.80**	0.64	0.37	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
COMET transl singleseq	0.64	**0.79**	0.18	0.01	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
COMET transl+intp singleseq	0.37	0.18	**0.79**	0.04	0.04	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
BertScore transl sent	0.00	0.01	0.04	**0.77**	0.75	0.93	0.17	0.08	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
BertScore transl+intp sent+mwer	0.00	0.01	0.04	0.75	**0.77**	0.97	0.20	0.10	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
COMET intp singleseq	0.00	0.00	0.00	0.93	0.97	**0.77**	0.32	0.21	0.01	0.01	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
BertScore transl+intp singleseq	0.00	0.00	0.00	0.17	0.20	0.32	**0.76**	0.12	0.03	0.02	0.02	0.01	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
BertScore transl singleseq	0.00	0.00	0.00	0.08	0.10	0.21	0.12	**0.75**	0.06	0.05	0.03	0.01	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
chrF2 transl+intp sent+mwer	0.00	0.00	0.00	0.00	0.00	0.01	0.03	0.06	**0.73**	0.93	0.27	0.27	0.22	0.02	0.05	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
BLEU transl+intp singleseq	0.00	0.00	0.00	0.00	0.00	0.01	0.02	0.05	0.93	**0.73**	0.87	0.42	0.39	0.00	0.08	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
chrF2 transl sent	0.00	0.00	0.00	0.00	0.00	0.01	0.02	0.03	0.27	0.87	**0.73**	0.41	0.33	0.03	0.09	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
chrF2 transl+intp singleseq	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.01	0.27	0.42	0.41	**0.72**	0.73	0.30	0.34	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00
chrF2 transl singleseq	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.01	0.22	0.39	0.33	0.73	**0.72**	0.32	0.37	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00
BLEU transl singleseq	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.00	0.03	0.30	0.32	**0.71**	0.86	0.20	0.00	0.01	0.00	0.00	0.00	0.00	0.00
COMET intp mwer	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.05	0.08	0.09	0.34	0.37	0.86	**0.71**	0.24	0.06	0.03	0.00	0.00	0.00	0.00	0.00
BertScore intp singleseq	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.02	0.20	0.24	**0.69**	0.51	0.11	0.07	0.01	0.01	0.00	0.00
BLEU transl+intp sent+mwer	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.06	0.51	**0.68**	0.45	0.00	0.12	0.12	0.00	0.00
chrF2 intp singleseq	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.03	0.11	0.45	**0.66**	0.72	0.45	0.40	0.00	0.00
BLEU transl sent	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.07	0.00	0.72	**0.65**	0.85	0.80	0.01	0.00
chrF2 intp mwer	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.12	0.45	0.85	**0.65**	0.93	0.00	0.00
BLEU intp singleseq	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.12	0.40	0.80	0.93	**0.65**	0.01	0.00
BertScore intp mwer	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.01	**0.60**	0.43
BLEU intp mwer	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.43	**0.58**

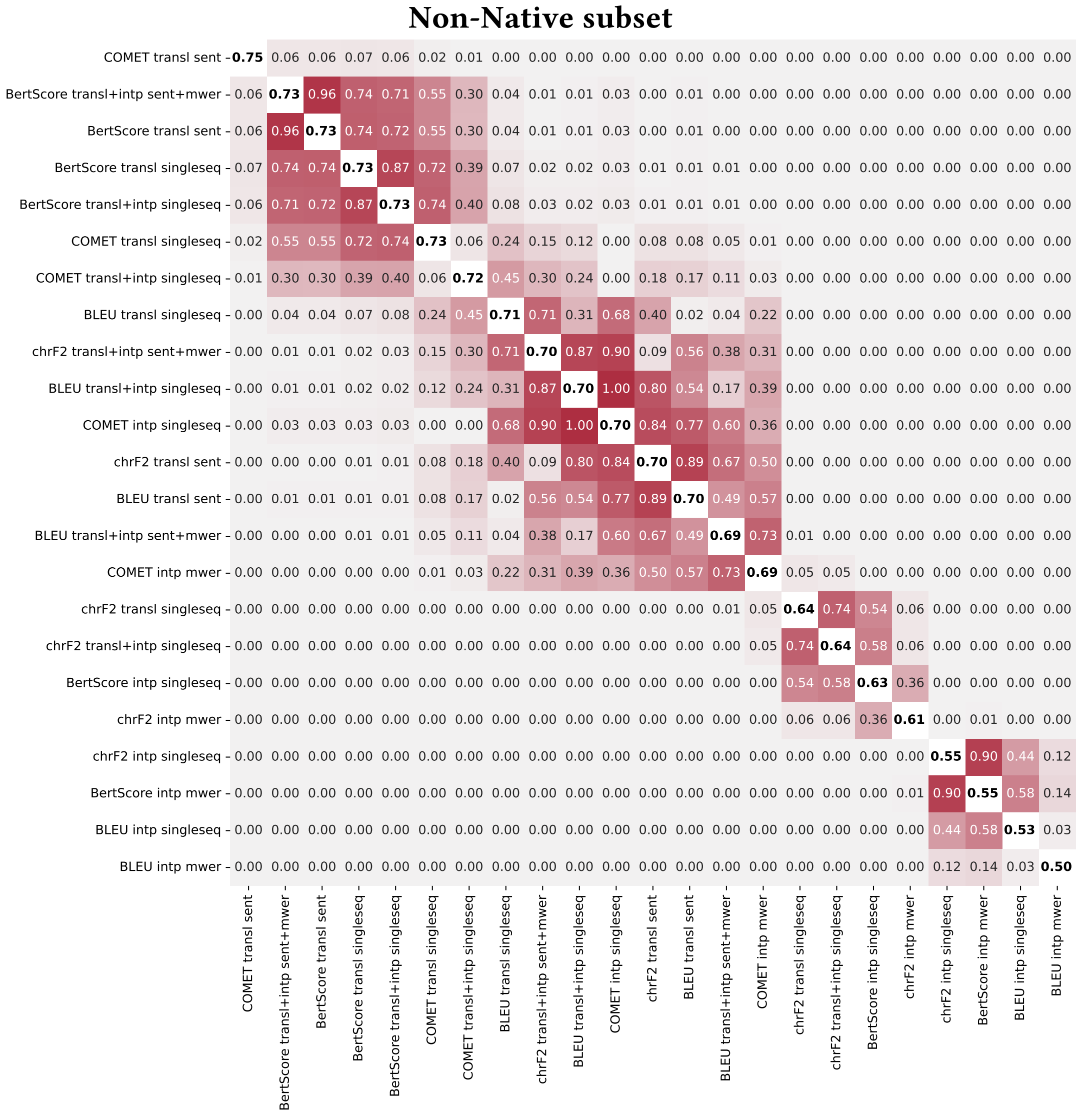
Figure 7.5: Results of significance test (p-values rounded to two decimal digits) for the difference of correlations of the metrics variants to CR. The metrics variants are ordered by Pearson correlation to CR on both subsets from the most correlating (top left) to the least (bottom right). The bold numbers on the diagonal are the correlation coefficients to CR.



Common subset

	COMET transl sent	COMET transl singleseq	COMET transl+intp singleseq	COMET intp mwer	BertScore transl+intp sent+mwer	BertScore transl sent	COMET intp singleseq	BertScore intp mwer	chrF2 transl+intp sent+mwer	chrF2 transl sent	chrF2 transl singleseq	chrF2 transl+intp singleseq	BertScore transl+intp singleseq	BertScore transl singleseq	chrF2 intp mwer	BLEU transl+intp singleseq	BLEU intp singleseq	chrF2 intp singleseq	BertScore intp singleseq	BLEU transl singleseq	BLEU intp mwer	BLEU transl+intp sent+mwer	BLEU transl sent
COMET transl sent	**0.76**	0.46	0.30	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
COMET transl singleseq	0.46	**0.75**	0.42	0.03	0.03	0.03	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
COMET transl+intp singleseq	0.30	0.42	**0.74**	0.05	0.05	0.05	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
COMET intp mwer	0.00	0.03	0.05	**0.69**	0.77	0.76	0.68	0.05	0.06	0.05	0.06	0.05	0.01	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
BertScore transl+intp sent+mwer	0.00	0.03	0.05	0.77	**0.69**	0.83	0.90	0.26	0.05	0.04	0.09	0.08	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
BertScore transl sent	0.00	0.03	0.05	0.76	0.83	**0.68**	0.91	0.28	0.06	0.05	0.09	0.09	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
COMET intp singleseq	0.00	0.00	0.00	0.68	0.90	0.91	**0.68**	0.34	0.21	0.20	0.17	0.16	0.02	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
BertScore intp mwer	0.00	0.00	0.00	0.05	0.26	0.28	0.34	**0.65**	0.55	0.53	0.48	0.45	0.14	0.08	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00
chrF2 transl+intp sent+mwer	0.00	0.00	0.00	0.06	0.05	0.06	0.21	0.55	**0.63**	0.56	0.78	0.72	0.26	0.14	0.01	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00
chrF2 transl sent	0.00	0.00	0.00	0.05	0.04	0.05	0.20	0.53	0.56	**0.63**	0.87	0.81	0.27	0.15	0.01	0.00	0.01	0.00	0.01	0.00	0.00	0.00	0.00
chrF2 transl singleseq	0.00	0.00	0.00	0.06	0.09	0.09	0.17	0.48	0.78	0.87	**0.63**	0.08	0.35	0.21	0.00	0.02	0.01	0.00	0.01	0.00	0.00	0.00	0.00
chrF2 transl+intp singleseq	0.00	0.00	0.00	0.05	0.08	0.09	0.16	0.45	0.72	0.81	0.08	**0.63**	0.37	0.22	0.00	0.02	0.01	0.00	0.01	0.00	0.00	0.00	0.00
BertScore transl+intp singleseq	0.00	0.00	0.00	0.01	0.00	0.00	0.02	0.14	0.26	0.27	0.35	0.37	**0.59**	0.02	0.43	0.12	0.13	0.13	0.02	0.00	0.02	0.00	0.00
BertScore transl singleseq	0.00	0.00	0.00	0.01	0.00	0.00	0.01	0.08	0.14	0.15	0.21	0.22	0.02	**0.58**	0.62	0.24	0.23	0.23	0.05	0.01	0.04	0.00	0.00
chrF2 intp mwer	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.01	0.00	0.00	0.43	0.62	**0.55**	0.74	0.41	0.07	0.24	0.18	0.03	0.06	0.02
BLEU transl+intp singleseq	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.02	0.02	0.12	0.24	0.74	**0.53**	0.65	0.67	0.42	0.00	0.10	0.00	0.00
BLEU intp singleseq	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.01	0.01	0.01	0.13	0.23	0.41	0.65	**0.51**	0.92	0.62	0.44	0.05	0.21	0.09
chrF2 intp singleseq	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.13	0.23	0.07	0.67	0.92	**0.51**	0.71	0.57	0.33	0.34	0.16
BertScore intp singleseq	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.01	0.01	0.02	0.05	0.24	0.42	0.62	0.71	**0.49**	0.78	0.66	0.52	0.25
BLEU transl singleseq	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.18	0.00	0.44	0.57	0.78	**0.47**	0.88	0.18	0.00
BLEU intp mwer	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.04	0.03	0.10	0.05	0.33	0.66	0.88	**0.47**	0.74	0.35
BLEU transl+intp sent+mwer	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.06	0.00	0.21	0.34	0.52	0.18	0.74	**0.45**	0.00
BLEU transl sent	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.00	0.09	0.16	0.25	0.00	0.35	0.00	**0.42**

Figure 7.6: Results of significance test (p-values rounded to two decimal digits) for the difference of correlations of the metrics variants to CR. The metrics variants are ordered by Pearson correlation to CR on the Common subset from the most correlating (top left) to the least (bottom right). The bold numbers on the diagonal are the correlation coefficients to CR.



Non-Native subset

	COMET transl sent	BertScore transl+intp sent+mwer	BertScore transl sent	BertScore transl singleseq	BertScore transl+intp singleseq	COMET transl singleseq	COMET transl+intp singleseq	BLEU transl singleseq	chrF2 transl+intp sent+mwer	BLEU transl+intp singleseq	COMET intp singleseq	chrF2 transl sent	BLEU transl sent	BLEU transl+intp sent+mwer	COMET intp mwer	chrF2 transl singleseq	chrF2 transl+intp singleseq	BertScore intp singleseq	chrF2 intp mwer	chrF2 intp singleseq	BertScore intp mwer	BLEU intp singleseq	BLEU intp mwer
COMET transl sent	**0.75**	0.06	0.06	0.07	0.06	0.02	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
BertScore transl+intp sent+mwer	0.06	**0.73**	0.96	0.74	0.71	0.55	0.30	0.04	0.01	0.01	0.03	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
BertScore transl sent	0.06	0.96	**0.73**	0.74	0.72	0.55	0.30	0.04	0.01	0.01	0.03	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
BertScore transl singleseq	0.07	0.74	0.74	**0.73**	0.87	0.72	0.39	0.07	0.02	0.02	0.03	0.01	0.01	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
BertScore transl+intp singleseq	0.06	0.71	0.72	0.87	**0.73**	0.74	0.40	0.08	0.03	0.02	0.03	0.01	0.01	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
COMET transl singleseq	0.02	0.55	0.55	0.72	0.74	**0.73**	0.06	0.24	0.15	0.12	0.00	0.08	0.08	0.05	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
COMET transl+intp singleseq	0.01	0.30	0.30	0.39	0.40	0.06	**0.72**	0.45	0.30	0.24	0.00	0.18	0.17	0.11	0.03	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
BLEU transl singleseq	0.00	0.04	0.04	0.07	0.08	0.24	0.45	**0.71**	0.71	0.31	0.68	0.40	0.02	0.04	0.22	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
chrF2 transl+intp sent+mwer	0.00	0.01	0.01	0.02	0.03	0.15	0.30	0.71	**0.70**	0.87	0.90	0.09	0.56	0.38	0.31	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
BLEU transl+intp singleseq	0.00	0.01	0.01	0.02	0.02	0.12	0.24	0.31	0.87	**0.70**	1.00	0.80	0.54	0.17	0.39	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
COMET intp singleseq	0.00	0.03	0.03	0.03	0.03	0.00	0.00	0.68	0.90	1.00	**0.70**	0.84	0.77	0.60	0.36	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
chrF2 transl sent	0.00	0.00	0.00	0.01	0.01	0.08	0.18	0.40	0.09	0.80	0.84	**0.70**	0.89	0.67	0.50	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
BLEU transl sent	0.00	0.01	0.01	0.01	0.01	0.08	0.17	0.02	0.56	0.54	0.77	0.89	**0.70**	0.49	0.57	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
BLEU transl+intp sent+mwer	0.00	0.00	0.00	0.01	0.01	0.05	0.11	0.04	0.38	0.17	0.60	0.67	0.49	**0.69**	0.73	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00
COMET intp mwer	0.00	0.00	0.00	0.00	0.00	0.01	0.03	0.22	0.31	0.39	0.36	0.50	0.57	0.73	**0.69**	0.05	0.05	0.00	0.00	0.00	0.00	0.00	0.00
chrF2 transl singleseq	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.05	**0.64**	0.74	0.54	0.06	0.00	0.00	0.00	0.00
chrF2 transl+intp singleseq	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.05	0.74	**0.64**	0.58	0.06	0.00	0.00	0.00	0.00
BertScore intp singleseq	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.54	0.58	**0.63**	0.36	0.00	0.00	0.00	0.00
chrF2 intp mwer	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.06	0.06	0.36	**0.61**	0.00	0.01	0.00	0.00
chrF2 intp singleseq	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	**0.55**	0.90	0.44	0.12
BertScore intp mwer	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.90	**0.55**	0.58	0.14
BLEU intp singleseq	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.44	0.58	**0.53**	0.03
BLEU intp mwer	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.12	0.14	0.03	**0.50**

Figure 7.7: Results of significance test (p-values rounded to two decimal digits) for the difference of correlations of the metrics variants to CR. The metrics variants are ordered by Pearson correlation to CR on the Non-Native subset from the most correlating (top left) to the least (bottom right). The bold numbers on the diagonal are the correlation coefficients to CR.

7.3.2 Recommendations

Based on the results, we make the following recommendations for the most correlating metric, reference, and sentence alignment method for SST evaluation.

Which metric? COMET because it correlates significantly better with CR than BertScore does. From the fallback options, chrF2 should be slightly preferred over BLEU.

Which reference? The metrics give significantly higher correlations with CR with translations than with transcribed interpreting as the reference. The difference between translation reference and two references (transl+intp) is insignificant. Therefore, we recommend translation as the reference for SST evaluation.

Which alignment method? If the sentence alignment of the candidate and reference is not available, COMET and BertScore correlate significantly more with SingleSeq than with mWER, probably because the neural metrics are trained on full, complete sentences, which are often split into multiple segments by mWERSegmenter. chrF2 correlates insignificantly better with mWER than with SingleSeq.

7.4 Summary

We analyzed human ratings of SST in English-to-German SST shared a task in IWSLT 2022, to answer practical questions for our next research where we aim to propose methods for developing multi-source SST applicable in the realistic use case.

First, the results show that we can safely use the MT metrics with translation reference. We will prefer BertScore because it correlates to human ratings the most and is available for multi-source translation into Czech, in contrast to COMET which requires one source.

Second, regarding using translation or interpreting references, we conclude that the current SST systems should be more like translation than like interpreting. When both are available, the difference between BertScore in multi-reference setup using both and using only translation was insignificantly more in favor to the multi-reference one.

Last but not least, we found interesting questions that a further analysis of these data could answer. The first one is that an inter-annotator agreement or other relevant way can be measured to figure out to what extent the CR is reliable. The second one is the question of machine achieving super-human performance (Tedeschi et al., 2023) in interpreting. Simultaneous interpreting is rated with CR the same way as the

SST system, and the SST systems were rated better than interpreting on the Native test set, but worse on the Non-Native set. This fact suggests to study whether this result from English-to-German IWSLT 2022 task is reliable and significant. If yes, it may be recommendable to apply speech translation instead of human interpreters on domains like English TED talks. Furthermore, it could enlarge the practical impact of multi-sourcing methods that we research because the second auxiliary source could be an automatic system instead of an interpreter, which would be more affordable.

8 Multi-Sourcing in Reality

In the previous chapters, we figured out that multi-source simultaneous speech translation (SST) from the original and simultaneous interpreting (SI) could be useful in the realistic use case on long-form monologue speech. In this chapter, we describe how we aimed to propose methods for developing it. First, in Section 8.1, we outline the general strategy for researching new technology, the loop of model, evaluation, and improvements.

Then, in Section 8.2, we describe our initial attempt, the late averaging multi-source model applied on automatic speech recognition (ASR) transcripts of parallel audios, and evaluation on real ASR noise levels, similarly as in Chapter 6. The results show that this baseline does not outperform the single-source model, so in Section 8.3 we propose improvement options that can be focused on in further work.

Since we first used offline ASR, and not simultaneous one, we improve this aspect by implementing Whisper-Streaming, a tool in which we adapt the state-of-the-art Whisper ASR model for real-time mode. It turned out that this tool is very useful, effective, robust, and innovative. We have published the implementation in a system demonstration paper "Turning Whisper into Real-Time Transcription System" (Macháček et al., 2023b) and received lots of appreciative feedback. We present Whisper-Streaming in Section 8.4 which includes the text that we previously published in this paper.

8.1 Strategy

We adopt the standard way of technology progress:

1. Baseline – we start with any model that solves the problem in any, even insufficient way. The model is denoted as a baseline.
2. Evaluation – we evaluate the model, analyze the results, and observe the weak spots of the model.
3. Improvement – we propose and implement improvement of the model, or of the evaluation process, e.g. if it is flawed, or if it is an approximation of realistic conditions that can be better.

The last two steps, evaluation and improvement, are iterated. In theory, the iteration does not stop until the performance is optimal. In reality, the researchers continuously evaluate their priorities, and may temporarily abort progress in the research task if another task gets a priority.

The baseline that we start with is the late averaging multi-source neural machine translation (NMT) from Chapter 6. We evaluate it on transcripts of original and interpreting, although we are aware that averaging requires totally parallel sources, which is not the case of SI on ESIC. If an evaluation reveals that it might be indeed a reason for low performance, we improve it.

8.2 Multi-Sourcing on Parallel ASR Transcripts

Previously, in Chapter 6, we evaluated multi-sourcing in simplified conditions – on parallel sentence aligned translations instead of ASR transcripts of original and interpreting. Now, we switch to conditions that are more realistic, although not entirely. We use the transcribed audio in ESIC (Macháček et al., 2021, Chapter 4). For that, we manually aligned the English original and English-German SI on the level of parallel sentence chunks (recall Section 4.6.3). We also do not use the artificial ASR noise model anymore, but real state-of-the-art ASR models.

There is still one unrealistic simplification left because we decomposed the main task into two subtasks that are easier to focus on separately than together. The first is the translation of synchronized parallel audio streams. The other task, synchronization of original and SI, is left for further work.

8.2.1 Real ASR Systems

For realistic evaluation of the ASR errors, we made a selection of the top performing ASR models for English and German on the domain of the European Parliament that is included in ESIC. We considered the models that were publicly available at Huggingface model repository[1] in early 2023 when we worked on this task. We compared the following models:

- **Whisper** by OpenAI (Radford et al., 2022). It is a Transformer model for speech-to-text transcription and translation trained on a massive amount of multi-lingual data. It supports 96 languages for ASR, including English, German, and Czech, and translation from these languages into English. Whisper also produces punctuation and supports long-form speech, not only individual sentences. It is available in multiple model sizes: large, medium, small, base, and tiny. The smaller models require less computational time and memory for processing, but they achieve lower quality because of lower capacity, especially for languages less represented in the training data. There are also finetuned Whisper models for English.

- **Wav2Vec 2.0** (Conneau et al., 2020) by Facebook is speech-to-text multilingual ASR model. We consider Wav2Vec 2.0 model finetuned on VoxPopuli for German.[2] VoxPopuli (Wang et al., 2021) is a large speech and text corpus from the European Parliament, the same domain as ESIC. Wav2Vec does not produce punctuation.

We compared the models on ESIC dev set in offline mode, on individually segmented sentences in the original English or German interpreting audio. Note that this is an approximation of a realistic setup, the simultaneous mode is more challenging.

Results The results are in Table 8.1. We present two types of scores, WER (word error rate), and CER (character error rate) with casing and punctuation on tokenized outputs, and in verbatim mode, which means without casing and punctuation. The highest difference 6% CER in punctuated versus verbatim modes is in the case of Facebook's Wav2Vec model on German, because it does not produce any punctuation or casing, while all Whisper models do.

[1] `https://huggingface.co/`

[2] `https://huggingface.co/facebook/wav2vec2-base-10k-voxpopuli-ft-de`

English original:

cased, punct.		verbatim			
CER	WER	CER	WER	model	quality level
4.1	12.1	3.1	5.8	**whisper-large**	①
4.7	13.5	3.6	6.9	whisper-small	
4.4	12.6	3.9	7.0	whisper-medium	
5.7	14.8	4.3	7.3	**whisper-medium-hi**	②
6.1	16.1	4.5	7.9	whisper-medium.en	
6.6	16.4	5.2	8.6	**whisper-small.en**	③
7.4	17.9	6.1	11.0	**whisper-base**	④
9.4	21.8	7.9	14.6	**whisper-tiny**	⑤
19.0	30.6	17.6	22.8	**whisper-tiny.en**	⑥
73.9	94.8	73.9	94.8	whisper-base-bn-trans	

German SI:

cased, punct.		verbatim			
CER	WER	CER	WER	model	quality level
7.2	16.3	6.4	11.6	**whisper-medium**	①
7.8	16.6	7.0	11.8	whisper-large	
8.2	19.7	6.9	13.2	**whisper-medium-hi**	②
10.4	22.8	9.4	17.6	**whisper-small**	③
19.2	48.8	13.0	20.8	**fb/wav2vec2**	④
15.5	33.1	14.2	27.5	**whisper-base**	⑤
24.3	48.7	22.7	42.4	whisper-tiny	

Table 8.1: ASR quality scores of segmented offline evaluation on ESIC dev of various models for English original speech (top) and German simultaneous interpreting (bottom). There is CER (character error rate) and WER (word error rate) calculated twice, with punctuation and casing, and without ("verbatim"). Both CER and WER are in 0-100% range, a low error rate means high quality. The models are ordered by the fourth column. Bold-highlighted models are the ones that we selected for multi-sourcing evaluation. The Whisper models that end with ".en" are finetuned for English, the other ones are multi-lingual. All the models can be found in HuggingFace.co repository under the respective label, except for fb/wav2vec2 which stands for `facebook/wav2vec2-base-10k-voxpopuli-ft-de`.

The worst quality model, whisper-base-bn-trans for English with nearly 95% WER is caused by the fact that this model is a preliminary checkpoint from the first stage of model training. It is not intended for ASR evaluation but as a base for further training.

Selection Based on the verbatim WER scores in Table 8.1, we selected models that represent ASR quality levels. We aimed to select verbatim WER levels with uniform distance from each other, therefore we disregard some of the similarly performing models. We highlight the selected models in Table 8.1, and we assign numbers to them so that we can refer them easily.

Examples Figures 8.1 and 8.2 show example outputs of English and German ASRs where we highlight the errors. The problem in these examples is the surname "Brejc" because it is pronounced incorrectly as "brake" in the original English speech. It is a Slovenian surname,[3] but the German ASRs propose German spelling. The other problem is the acronym "VIS" because it sounds similar to "visa," and the models tend to assume that similarly sounding words within one sentence are identical, but in this case, they are different. Last but not least, we observe words that were transcribed by German ASRs with different orthographical forms than in gold. It may not be a problem in reality, but the automatic metrics WER and CER will detect it as an error.

Other observations We thoroughly read and compared the ASR transcripts with gold, and made the following observations.

- The ASR quality is often optimal (0% WER) in all quality levels if there is no challenging case in the output, such as a rare word or a proper name.
- On many segments, all ASRs performed very low quality near 100% WER in all ASR quality levels. They either did not produce any output or produced some "hallucination," a long non-sense token repetition. We assume it could be because of non-standard audio, such as noise or a wrong sentence segmentation. The models are not trained for sequences in which a word or sentence starts in the middle.
- We noticed some errors in the segmentation of the evaluation data into individual sentences. In some cases, a word is included at the end of a sentence, but in audio, this word is at the beginning of the succeeding sentence. An example is in Figure 8.3. It may be caused by the fact that the audio segmentation was

[3] https://en.wikipedia.org/wiki/Mihael_Brejc

gold	On behalf of the Greens **Verts** Group I would like to thank Mr **Brejc** for his great report.
①	On behalf of the Green **Zephyr** Group, I would like to thank Mr. **Brick** for his great report.
②	On behalf of the Green **Zephyr** Group, I would like to thank Mr. **Brick** for his great report.
③	On behalf of the Greens, **I have a group** I would like to thank Mr. **Breck** for his great report.
④	On behalf of the Green **Zefa** Group, I would like to thank Mr. **Breck** for his great report.
⑤	On behalf of the **Greens** ∅, I would like to thank Mr. **Break** for his great report.
⑥	on behalf of the **Greens, if a group** I would like to thank Mr. **Break** for his great report.

gold	And we appreciate his point that consulting the **VIS** using the number of the **visa** sticker in combination with verification of fingerprints will create a lot of problems.
①	And we appreciate his point that consulting the **Viz** using the number of the **Viz** sticker in combination with the verification of fingerprints will create a lot of problems.
②	And we appreciate his point that consulting the **visa** using the number of the **visa** sticker in combination with verification of fingerprints will create a lot of problems.
③	And we appreciate his point that consulting the **Veeze** using the number of the **Veeze** sticker in combination with the verification of fingerprints will create a lot of problems.
④	And we appreciate his point that consulting **Zeev is** using the number of **Zeev's Tika** in combination with the verification of **finger prints** will create a lot of problems.
⑤	And we appreciate his point that consulting **ZV is** using the number of **ZVs TK** in combination with the verification of **finger prints. We'll** create a lot of problems.
⑥	**a**nd we appreciate his points that consulting the **V is** using the number of the **V's** sticker in combination with the verification of fingerprints**. We'll** create a lot of problems.

Figure 8.1: Example of English ASR transcripts on the first two sentences of ESIC dev 20080901.018_006_EN_Ždanoka speech. The numbers in the first column correspond to the English ASR models in Table 8.1. We highlight serious errors in red (neglecting small morphological changes near the serious errors), orange are errors that probably not influence understanding of English ASR transcript, but will be serious in translation. Green is a correct transcript where all other systems were incorrect.

gold	Im Namen der **Grünen-Fraktion** möchte ich Herrn **Brejc** für seinen Bericht danken.
①	Im Namen der **Grünenfraktion** möchte ich Herrn **Breitz** für seinen Bericht danken.
②	Im Namen der **Grünenfraktion** möchte ich Herrn **Braetz** für seinen Bericht danken.
③	Im Namen der **Grünen Fraktion** möchte ich Herrn **Breitz** für seinen Bericht danken.
④	im namen der **grünen fraktion** möchte ich herrn **breit** für seinen bericht danken
⑤	Im Namen der **Grünfraktion** möchte ich **an Breiz** für seinen Bericht danken.

gold	Wir schätzen seinen Standpunkt, dass die **Visa-Inhaber** überprüft werden sollen durch eine Kontrolle der **Visa-Marke** und durch eine Abnahme von Fingerabdrücken, aber das wird zu sehr vielen Komplikationen führen.
①	Wir schätzen seinen Standpunkt, dass die **Visainhaber** überprüft werden sollen durch eine Kontrolle der **Visa Marke** und durch eine Abnahme von Fingerabdrücken**. Aber** das wird zu sehr vielen Komplikationen führen.
②	Wir schätzen seinen Standpunkt, dass die **Visa-Inhaber** überprüft werden sollen durch eine Kontrolle der **Visa-Marke** und durch eine Abnahme von Fingerabdrücken**. Aber** das wird zu sehr vielen Komplikationen führen.
③	Wir schätzen seinen Standpunkt, dass die **Visorinhaber prüft** werden sollen durch eine Kontrolle der **Visormarke** und durch eine Abnahme von Fingerabdrücken**. Aber** das wird zu sehr vielen Komplikationen führen.
④	wir schätzen seinen standpunkt dass die **visa inhaber** überprüft werden sollen durch eine kontrolle der **visa marke** und durch eine abnahme von fingerabdrücken aber das wird zu sehr vielen komplikationen führen
⑤	Wir schätzen seinen Standpunkt, dass sie **Wieserinhaber probt** werden sollen, durch eine Kontrolle der **Wiesermarke** und durch eine Abnahme von Fingerabdrücken**. Aber** das wird zu sehr vielen Komplikationen führen.

Figure 8.2: Example of ASR transcripts of German SI on the first two sentences of ESIC dev 20080901.018_006_EN_Ždanoka speech. The numbers in the first column correspond to the German ASR models in Table 8.1. Note that model 4 does not insert punctuation and casing, so we do not evaluate it. In blue, we mark the words that use spellings or punctuations that are different from gold, but they are grammatical and acceptable. The other color highlighting is as in Figure 8.1.

gold	man sollte das jetzt nicht **einfrieren**
ASR	man sollte das jetzt nicht ∅

The next sentence:

gold	ist ja auch nicht schuld der kommission
ASR	**einfrieren** ist ja auch nicht auf schulterkontakt

Figure 8.3: Example of wrong segmentation of evaluation data to sentences. The words "einfrieren" is at the end of the sentence in the gold transcript and at the beginning of the next sentence in audio and in ASR transcript. Automatic evaluation scores (e.g. WER) may count two errors, one deletion and one insertion.

based on the automatic forced alignment of the transcript to audio. In challenging cases such as noise, hesitations, non-transcribed false start, etc., the timestamps may be inaccurate. Wrong segmentation makes automatic evaluation wrong.

The issue of segmentation that causes wrong evaluation and may cause hallucination can be resolved by long-form mode evaluation. It means that we will not process ASR on individual sentences but on the whole, unsegmented documents.

8.2.2 Averaging Multi-Source Results

We evaluate late averaging multi-source model from Chapter 6 in offline mode, with the text inputs produced by selected ASR models.

Evaluation metric We use the evaluation method based on our findings from Chapter 7. The most correlating metric to human judgments is COMET (Rei et al., 2020), but COMET is not available for multi-sourcing because it requires one source on the input. We could count COMET twice, with German and English sources, but it would give us two scores, not one for a simple comparison. We would also need to investigate whether it is better to use the inaccurate ASR transcripts as a source in COMET, or the perfect gold transcripts.

The second most correlating metric to human judgments that we can consider is BERTScore (Zhang et al., 2020b). We use it with two references, with normalized text translation into Czech, and with transcripts of simultaneous interpreting because in Chapter 7 we found out that it correlates slightly more to human judgments than translation as a single reference. However, using only translation is also reasonable, the difference is not statistically significant. Our case also differs from the one that was investigated in Chapter 7 by normalization of the translations.

Results The results are in Figure 8.4. We can observe that with the gold transcripts (topmost group of bars) and with English ASR quality levels ①-⑤ which is 5.8% to 14.6% verbatim WER (Table 8.1), single sourcing (blue bar) achieves higher BertScore than multi-sourcing with German gold (orange) or ASR of any level (green, red, violet, brown, pink; also see the legend). With English ASR level ⑥, 22.8% WER, multi-sourcing with the German ASR source achieves higher BertScore, except with the ASR model ④, the one without punctuation and casing. However, the difference is very small, less than 0.1 BertScore.

In Figure 8.4, we do not show German single-source results. This is because translation from German achieves significantly lower BertScores. Probable reasons are the traces of the Czech reference source language (English), the same as we analyzed in Chapter 6. Furthermore, there may be traces of simultaneous interpreting style in German-Czech translations that BertScore penalizes because the references are normalized English-Czech translations averaged with English-Czech simultaneous interpreting, which may use a slightly different style because the language pair is structurally different than English-German.

8.2.3 Human Evaluation

We performed human evaluation to reliably compare late-averaging multi-sourcing versus single sourcing machine translation (MT) that use the top performing ASRs for English and German.

The single source translation achieves BertScore 0.8301, and multi-source 0.8246; see Figure 8.4, English ASR ① (second group from the top) and blue vs. green bar, which means no German source and German ASR ①, respectively. The underlying ASR models are multi-lingual Whisper large for English and Whisper medium for German. They achieve 5.8% verbatim WER for English and 11.6% for German on ESIC dev (Table 8.1).

We used ESIC dev subset for this evaluation because we consider it as one step within the development of the SST system. We keep ESIC test set undisclosed to the evaluator because he is also a developer and the main author, and disclosing the test set to him could lead to overfitting.

We randomly shuffled documents in the subset and presented blind translations of the whole document together with the English normalized source and Czech normalized translation reference. We aligned the sentences of the candidates to the reference with mWERSegmenter (Matusov et al., 2005). This automatic tool may be partially inaccurate, however, we instructed the annotator that this alignment is only for easier orientation, he should not mind misalignments and consider the document-level context.

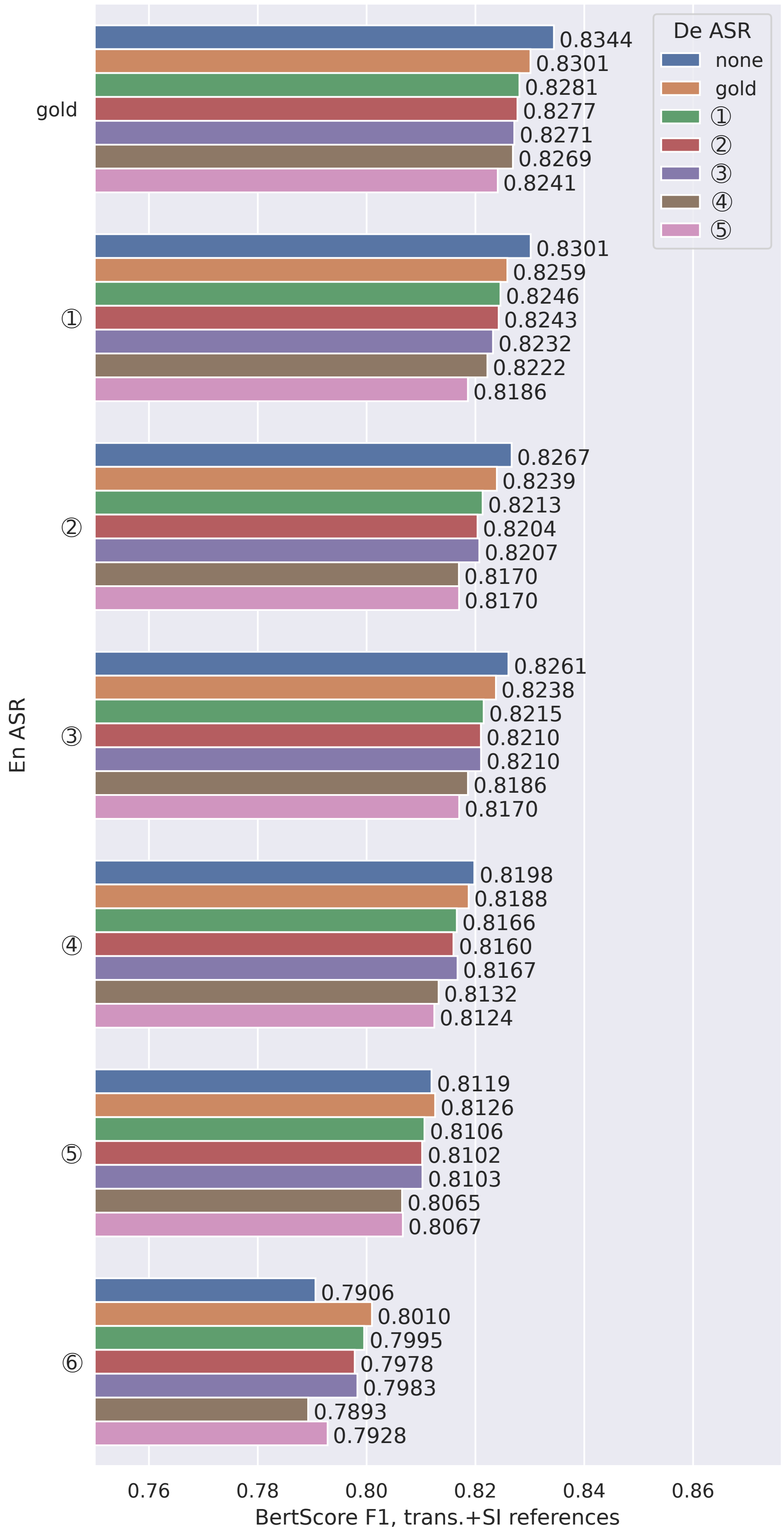


Figure 8.4: The results of late averaging multi-sourcing in offline mode on ESIC dev with ASR transcripts of different quality levels (①-⑥ on English, ①-⑤ on German, the numbers correspond to systems in Table 8.1). The evaluation metric is BertScore F1 on a 0-1 scale, the higher, the better quality.

	En+De multi-src	**En single-src**	description
total +	28	**34**	better translated expressions
+	19	28	1 better in segment
++	1	3	2 better in segment
+++	3	0	3 better in segment
—	2	2	worse translated expr. in segment
?	2	3	not graded segment
0	50	43	not +/-/?, or comparable
total	79 segments, 6 documents 1 evaluator, 2 hours		total rated segments

Table 8.2: Results of human evaluation. English single-source system is evaluated as better because it achieved more total + grades (34 vs. 28) assigned to better-translated words or phrases in the 79 rated segments.

Then, we instructed the annotator to grade the translation candidates by putting zero or more "+" grades for every expression, word or phrase, that is translated better in one candidate than in the other. He could also use "—" (minus) grades for very bad translation, e.g. "hallucination," or skip grading. Comparable or identical translations were graded by "0." The evaluator was instructed to focus only on adequacy, and mind grammar only if it affects adequacy.

In total, the evaluator rated 6 documents, 79 sentence-like segments in translation, and spent 2 hours.

The results are in Table 8.2. They show that single-source translation performs better. It achieved 34 better-rated translations, which is 6 more than single-sourcing. This result is consistent with BertScore.

However, the fact that 28 expressions in 79 sentences of the multi-sourcing candidate were translated better than in the single source leads us to the conclusion that multi-sourcing is able to benefit the translation quality, but multi-sourcing probably brings more issues than benefits. However, we can investigate the issues and propose a multi-sourcing model that performs better than the late averaging.

In Figure 8.5, we show a cherry-picked example where the additional German source helped to improve the transcription error in the English source.

ESIC dev 20110215/005_017_EN_Tarand:

SRC	- Madam President, in my opinion, Mr Werner Schulz has drafted a resolution which is very well **founded** with arguments and draws correct conclusions.
REF	- Paní předsedající, pan Werner Schulz navrhl podle mého názoru usnesení, které je dobře **odůvodněno** argumenty, a jeho výsledkem jsou správné závěry.
En ASR 1	Thank you. In my opinion Mr. Schulz has drafted a resolution which is very well **funded** with arguments and draws correct conclusions.
De SI ASR 1	Herzlichen Dank! Ich denke, dass Herr Schulz eine Entschließung verpasst hat, die wirklich sehr gute Argumente **beinhaltet** und auch die richtigen Schlüsse zieht.
En→Cs	Děkuji vám. Podle mého názoru pan Schulz vypracoval usnesení, které je velmi dobře **financováno** argumenty a vyvozuje správné závěry.
En+De→Cs	Díky. Myslím, že pan Schulz přišel s usnesením, které je velmi dobře **podloženo** argumenty a vyvozuje správné závěry.

Figure 8.5: Cherry-picked example of multi-source translation outperforming single source in the translation of the word “founded.” English ASR 1 (Whisper large) following the original speech incorrectly transcribed “funded” instead of “founded.” English→Czech single source system translated it wrongly as “financed” (“financováno”), while the multi-source English+German→Czech translated it correctly as “grounded” (“podloženo”). It is very likely thanks to the German ASR 1 (Whisper medium) following German SI that correctly transcribed the corresponding word “beinhaltet” (“contains arguments“).

8.3 Improvement Options

In the previous section, we did not observe the benefits of multi-sourcing in offline mode with the averaging model on ESIC with real ASRs. In this section, we propose improvement options for further research of multi-source SST for realistic use case, but for time and capacity reasons we elaborate them only very superficially or not at all. Therefore, we describe the options only very briefly, without all the details for reproduction because we are aware that our investigation in these areas is far from completed.

We decided to first focus thoroughly on investigating multi-sourcing quality with real simultaneous ASR. We describe it below, in Section 8.4.

8.3.1 More Challenging Domain

It is possible that we do not observe the benefits of multi-sourcing on ESIC because it may not be a very challenging test set for ASR and SST. We can consider another test set that represents a domain where multi-sourcing may be beneficial, e.g. due to difficult acoustic conditions and non-native accent. There is the non-native IWSLT 2022 test set that consists of English original speech, parallel German simultaneous interpreting, and there exist corresponding Czech reference translations, although not in IWSLT 2022. The IWSLT 2022 non-native test set consists of mock student business presentations (Macháček et al., 2019) with very challenging background noise and strong non-native accents. The second part consists of auditing presentations of non-native English speakers. Another advantage of this test set is that the reference translations may be direct, and not revised and normalized as in ESIC.

8.3.2 Error Detection

The next set of improvement options aims to resolve the possible issue that the averaging multi-sourcing does not have the ability to detect and avoid errors in the sources. We suggest the following options to improve it:

Confidence scores The multi-sourcing model may use the *confidence scores* that estimate how much correct is the transcription of each word in the sources (Laptev and Ginsburg, 2023; Afshan et al., 2021). They may be useful e.g. in early averaging multi-source model (Firat et al., 2016b, recall Chapter 6).

More sources for voting We experimented with only two sources, English and German. If there were three or more sources, there would be an opportunity to vote on what is correct and what is not. However, the limiting factor for experimenting with more sources is the test set. ESIC contains only three languages that we use as two sources and one target. We would need to add a fourth language, e.g. French.

Multi-sequence and noise-robust training We made an initial experiment where we trained a multi-sequence model with concatenation (Dabre et al., 2017). We used multi-parallel data for English and German into Czech, an 8 million sentence triple subset of the 30 million that we used in Chapter 6. The results were analogical to those in Section 8.2, multi-sourcing was not outperforming baseline. Then, we estimated the distribution of counts of ASR errors of the top performing offline Whisper models on ESIC dev original and simultaneous interpreting, and we inserted errors in the same counts to the multi-parallel training data using the lexical noise model from Chapter 6. The results showed that multi-sourcing again did not outperform the baseline; however, a more detailed analysis and finishing of this experiment are pending.

8.3.3 Quality Estimation

Next direction of research that we considered is applying MT quality estimation (QE, Zerva et al., 2022; Rei et al., 2022; Rubino et al., 2021) for SST (Stewart et al., 2018) and for multi-sourcing. QE is a task that either assigns a quality score to the candidate translation, given the source sequence, or ranks one or more candidate translations by quality. QE usually works at the sentence level, but it can be adapted to score the tokens in a sequence (Zhao et al., 2021b).

Adapting QE for multi-source SST from original and SI is a challenging, but not an impossible task. First, it requires only parallel corpora that can be machine-translated and scored by automatic MT metric. Then, a neural network predicting the quality scores can be trained on it. Adaptation for multi-sourcing is straightforward, e.g. multi-sequence NMT (Dabre et al., 2017) can be used. Adaptation for speech source modality and simultaneity is more challenging, but we assume that training on incomplete sentence prefixes as in SST (Niehues et al., 2018) can be applied.

Then, QE can be e.g. applied to select the better candidate from the set of single-source and multi-source translations. Token-level QE can be applied as a confidence score.

8.3.4 Interpreting Style Training

Another option why multi-sourcing may not be beneficial on ESIC original and interpreting, is the specific interpreting style. We experimented with multi-sequence training on multi-parallel translations, and not on original and interpreting. Therefore, we can consider using either authentic interpreting data from VoxPopuli (Wang et al., 2021), corpus of speeches and interpreting from the European Parliament, or we can synthesize interpreting from parallel translation corpus using style transfer model as Zhao et al. (2021a).

In Zhao et al. (2021a), the authors create a training corpus of German original and parallel English simultaneous interpreting sentence pairs. Then, they trained a statistical MT system to change the style of English translation to English simultaneous interpreting. It e.g. changed words and phrases to shorter variants, as in interpreting style. Then they finetuned NMT on synthetic interpreting, and observed higher quality (i.e. better match with references) when evaluating against interpreting.

In our future work, we can use the same approach to create a German translation-to-interpreting style-transfer model, and synthesize German simultaneous interpreting training data for multi-sequence NMT.

Moreover, we can consider ChatGPT for generating the interpreting style data.

8.3.5 ChatGPT

ChatGPT (Chat Generative Pre-trained Transformer, OpenAI, 2022) is a large language model adapted by reinforcement learning from human feedback to serve as a chatbot assistant. It was developed by the company OpenAI and made accessible to the public through API (application programming interface) and public website.[4] It was first released for research and the public in late 2022, and it soon received lots of public attention because it appeared to be very robust, useful and effective tool for generating texts based on natural language prompts, and at the same time performing very poorly or questionably on many other tasks, especially when working with factual knowledge.

[4] https://chat.openai.com/

ChatGPT is able to chat in many languages. Moreover, it is able to translate,[5] change style, summarize, answer questions, estimate translation quality (Kocmi and Federmann, 2023), and work with long background context. We see an opportunity to utilize ChatGPT for our primary research task, multi-source speech translation from original and simultaneous interpreting. Therefore, we briefly investigated ChatGPT on this task, to get an initial overview of its performance, limitations, and possible future work.

We used `gpt-3.5-turbo` model in March 2023. We experimented with prompting ASR error correction, translation of the whole document from English and German into Czech given a context of ESIC speeches – European Parliament, name of the speaker and date, because we supposed that the relevant background information (e.g. European Parliament, Wikipedia, and news from ESIC period 2008-2011) are in ChatGPT training data and it can retrieve them and use them for answering. We also prompted translation from one language source and from two parallel language sources, both from gold text transcripts and from the ASR transcripts with errors. We did not complete a rigorous analysis, but we made the following initial observations that could be tested rigorously:

1. ChatGPT is able to correct simple ASR errors only from the textual context. For example, in the salutation in the European Parliament, "I'm Commissioner" corrects to "Mr. Commissioner." It also corrects "Green Zephyr Group" into "Green-EFA Group," which is a suitable and existing alternative acronym of the political fraction that is mentioned,[6] although the speaker said "Green Verts Group."

2. When prompted directly, ChatGPT knows the speaker, MEP Tatjana Ždanoka, and the topics she was focusing on back in the period 2008-2011 that is included in ESIC. We therefore assume that ChatGPT (or another generative large language model) could be used to generate background context data or finetuning data for ASR of the speaker.

3. ChatGPT is able to translate whole ESIC documents at once. We assume it can use the document-level context, but we did not analyze it in detail. The segmentation of translation output alternates, sometimes there is one sentence per line, sometimes a paragraph per line.

4. ChatGPT is able to translate from two language sources, but we did not analyze whether it combines them for higher quality, or uses only the first one.

[5] https://www.makeuseof.com/how-to-translate-with-chatgpt/

[6] https://en.wikipedia.org/wiki/Greens%E2%80%93European_Free_Alliance

5. It is possible to give ChatGPT a complex prompt such as "Here are possibly incorrect ASR transcripts of the English original and simultaneous interpreting into German. Please, correct them and translate them into Czech." However, we observed that the answers are rather problematic, ChatGPT usually performs only a part of the complex prompted task. Sometimes it catches a keyword "correct," and makes stylistic changes, sometimes it answers in German, in the second language, and not in Czech.

Moreover, we are aware of ChatGPT limitations stemming from the fact that the model checkpoint is not available for inference outside OpenAI. It is therefore not advisable for confidential use cases. The model is also very large and inference is expensive. ChatGPT training data are not accessible, and it is possible that the ESIC development and evaluation set is included in the training data, so that evaluation on ESIC is unreliable. On the other hand, there may be other large language models with similar capabilities without these limitations (Radford et al., 2019; Scao et al., 2022). We assume that using large language models for improving SST from multiple languages could be possible and reasonable research direction.

8.4 Whisper-Streaming

So far, in Section 8.2 we used offline ASR systems, but the more realistic is to apply ASR in simultaneous mode, which is more challenging. There may be more errors and more opportunities for multi-sourcing to correct them.

In this section, we describe our first step in investigating multi-sourcing with state-of-the-art simultaneous ASR. We implement Whisper-Streaming, a tool that we plan to apply in the realistic multi-sourcing simulation. We use it to process the source speech signals, each of them separately, into the text streams. Then we apply the multi-sourcing text-to-text MT that we propose.

In Whisper-Streaming, we use Whisper (Radford et al., 2022), a state-of-the-art speech-to-text model that works very well in offline mode and on long-form speech, but it does not support simultaneous mode. However, adapting it for online mode is very easy with the LocalAgreement streaming policy. Our colleague Peter Polák implemented a demonstration of LocalAgreement with any Huggingface speech-to-text model.[7] We realized that his implementation assumes only the segmented speech, not the long-form unsegmented speech. However, Whisper produces punctuation and word-level timestamps, and it is therefore possible to use them to segment the incoming audio buffer to contain only the last single sentence.

[7] https://github.com/pe-trik/transformers/blob/online_decode/examples/pytorch/online-decoding/whisper-online-demo.py

We created Whisper-Streaming, an implementation of real-time speech transcription and translation of Whisper-like models. Since it is a very robust, well-performing and innovative tool, achieving 3.3 seconds latency on ESIC English ASR, we wrote and published a system demonstration paper "Turning Whisper into Real-Time Transcription System" (Macháček et al., 2023b). It was accepted for presentation at IJCNLP-AACL 2023 conference. In this section, we use text that we published in this paper.

8.4.1 Background

Whisper (Radford et al., 2022) is a Transformer model for speech-to-text transcription and translation trained on a massive amount of multi-lingual data. We use "large-v2"[8] model because it achieves the highest quality of all Whisper model size options. Since the original release of the Whisper backend is rather slow, we use the `faster-whisper`[9] reimplementation of Whisper inference using CTranslate2, a fast inference engine for Transformer models. It is approximately four times faster than the standard implementation (as reported by the authors). We use it on NVIDIA A40 GPU with 16-bit float precision.

Although we primarily use Whisper, the underlying model in our implementation can be easily replaced by any other speech-to-text transcription or translation model (e.g. MMS, Pratap et al., 2023; SeamlessM4T, Barrault et al., 2023) if it produces word-level timestamps and punctuation.

Streaming Let us assume a model M that processes a source sequence $c_1, \cdots, c_n$ into a target sequence $t_1, \cdots, t_m$, given a target of the previous sequence s that can be used as a "prompt" in Whisper for inter-sentence coherence. Streaming involves receiving the source sequence consecutively in discretized units, one chunk at a time, and producing the target simultaneously. A *streaming policy* P predicts a target segment t_T at time T as $t_T := P_M(c_{i<T}|s, t_{j<T})$. It operates the model M on available source chunks $c_{i<T}$, previous sequence target s, and previous target segments $t_{j<T}$. The policy is triggered every time a new source segment is available. An empty target segment can be emitted, e.g. when waiting for context. The policy aims to minimize latency and maximize target quality.

Streaming was originally proposed for simultaneous translation (Ma et al., 2019), but it is applicable for any sequence-to-sequence task including ASR. Dong et al. (2022) give a summary of streaming speech translation.

[8] `https://huggingface.co/openai/whisper-large-v2`

[9] `https://github.com/guillaumekln/faster-whisper`

LocalAgreement (Liu et al., 2020) is a streaming policy that outputs the longest common prefix of the model on n consecutive source chunks, or an empty segment when less than n chunks are available. Based on the IWSLT 2022 shared task on simultaneous translation (Anastasopoulos et al., 2022), the CUNI-KIT system (Polák et al., 2022) compared LocalAgreement to other policies (hold-n and wait-k) with different chunk sizes. They found that LocalAgreement with $n = 2$ was the best effective policy. Therefore, we use LocalAgreement-2 for identifying stabilized target segments.

8.4.2 Implementation

We describe the core components and inner workings of Whisper-Streaming. It consists of the update loop, audio buffer, skipping the confirmed output in audio buffer, trimming the buffer, joining for inter-sentence context, and optional voice activity detection.

Update loop The main part of Whisper-Streaming is a program that utilizes a loop to receive source audio chunks and trigger streaming policy updates. The parameter MinChunkSize determines the minimal duration processed per iteration. If the update computation exceeds MinChunkSize, the next update is performed immediately on the accumulated audio input. This parameter impacts both latency and quality.

Audio buffer Whisper is trained to handle sequences that are up to 30 seconds long and contain one full sentence. It provides punctuation and word-level timestamps.[10] The process is illustrated in Figure 8.6. Each update involves storing incoming audio at the end of the audio buffer and processing the entire buffer with Whisper. We keep an invariant that the buffer always starts at the sentence boundary, to maintain the high quality of Whisper. LocalAgreement-2 is applied to the current and previous Whisper output. The timestamp of the last word in the "confirmed output" is saved. In subsequent updates, we always reprocess Whisper from the beginning of the buffer, including the portion preceding the last "confirmed output" timestamp (indicated by the gray background in Figure 8.6). Changes to the transcription in the confirmed portion are disregarded, as they are often insignificant in terms of meaning alteration.

[10] When using "faster-whisper" or another implementation that supports it.

Skipping the confirmed part When determining the position of transcribed words relative to the last confirmed word from the previous update, we account for the potential inaccuracies and updates in Whisper timestamps due to new audio chunks. If a word's timestamp falls within a 1-second interval from the last confirmed word, we compare its preceding n-grams (where n ranges from 1 to 5) with the suffix in the last confirmed output. If they match, we skip those words. However, this rule can be further enhanced in future work by incorporating measures such as setting and fine-tuning a character edit distance threshold, trimming punctuation and casing from the n-grams, etc.

Trimming the audio buffer To avoid unacceptably long spikes in latency, the audio buffer is limited to around 30 seconds. When the confirmed output includes a sentence-ending punctuation mark followed by a word starting a new sentence, the buffer is trimmed at the punctuation mark's timestamp. A language-specific sentence segmentation tool (e.g. Koehn et al., 2007) is used for this purpose, ensuring that the buffer always contains a single sentence. Despite this, if the buffer length exceeds 30 seconds, we retain the last confirmed segment marked by Whisper.

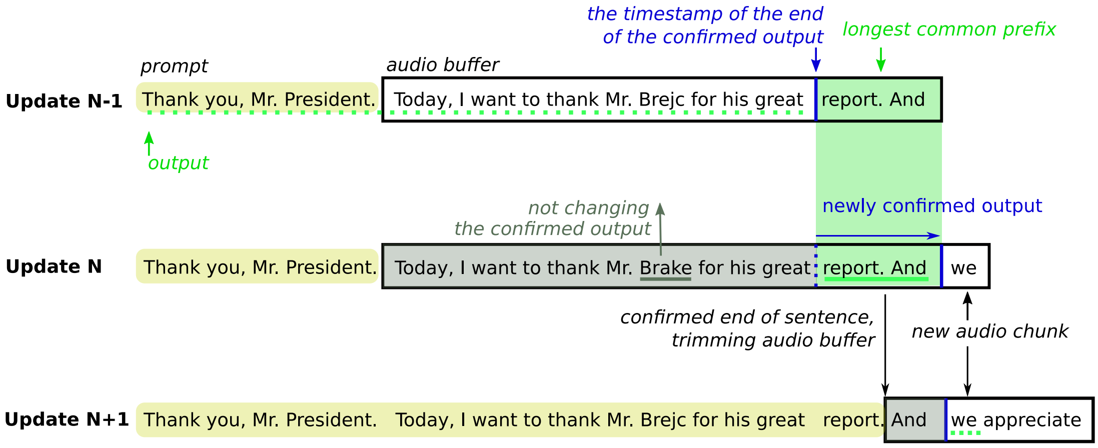
Joining for inter-sentence context The Whisper transcribe function utilizes a "prompt" parameter to maintain consistency within a document (consistent style, terminology, and inter-sentence references). We extract the last 200 words from the confirmed output of previous audio buffers as the "prompt" parameter, as shown in Figure 8.6 (yellow backgrounded text).

Voice activity detection There is a parameter to activate or deactivate Whisper's default voice activity detection (VAD) filter, impacting both quality and latency.

8.4.3 Evaluation Setting

We describe the dataset for evaluation, metrics, settings, and hardware we used to evaluate our model.

Evaluation Data For latency and quality analysis, we utilize the dev set of the manually transcribed ESIC corpus (Macháček et al., 2021) for English, German, and Czech ASR containing 179 documents. This corpus contains 5 hours of original English speeches from the European Parliament, including simultaneous interpreting into German and Czech. It provides audio tracks with manual transcripts and word-level timestamps.



the timestamp of the end of the confirmed output
longest common prefix
prompt
audio buffer
Update N-1
Thank you, Mr. President.
Today, I want to thank Mr. Brejc for his great
report. And
output
not changing the confirmed output
newly confirmed output
Update N
Thank you, Mr. President.
Today, I want to thank Mr. Brake for his great
report. And
we
confirmed end of sentence, trimming audio buffer
new audio chunk
Update N+1
Thank you, Mr. President. Today, I want to thank Mr. Brejc for his great report.
And
we appreciate

Figure 8.6: Illustration of processing three consecutive updates. The yellow highlighted text is a "prompt," the previous context to follow. The black-bordered rectangle is an audio buffer, and the text inside is Whisper's transcript is generated from that sound segment. The blue vertical line is a timestamp that splits the buffer into two parts, the left being previously confirmed, and the right being unconfirmed. The LocalAgreement-2 policy, or searching the longest common prefix, is applied on the unconfirmed (right) part in two subsequent updates. The longest common prefix is highlighted in green and the green underline highlights the *newly* confirmed output, whereas the green dashed underline indicates previously and subsequently confirmed output. The gray underline demonstrates an update in the confirmed part that is disregarded.

WER We use word error rate (WER) after removing punctuation and casing as the standard measure of ASR quality.

Latency In our latency analysis, we implement our own method wherein we use the timestamps provided in the ESIC corpus to align the gold transcripts to the ASR output using edit distance.[11] This allows us to determine the edit operations for each gold word. We calculate the ASR latency by measuring the time difference between when the ASR emitted a word and when the corresponding gold word was spoken, excluding words deleted by the ASR. We compute the average latency within each document and, when comparing different setups across multiple documents, we report the average latency along with the standard deviation.

We do not use the tool SimulEval (Ma et al., 2020) because it does not support our use case, the long-form ASR evaluation. We also do not use SLTev (Ansari et al., 2021) because it is not as simple and transparent as our code.

[11] https://pypi.org/project/edlib/

GPU	VAD	% WER	latency [s]
A40	**off**	**5.8±0.9**	**2.85±0.45**
A40	on	5.2±0.9	3.12±0.36
L40	off	5.1±1.0	3.58±0.62
L40	on	5.0±0.6	3.96±0.81

Table 8.3: Average (±stddev) WER and latency of English ASR of 10 repeated runs of ESIC dev.20080925.013_007 document, with MinChunkSize 0.1 seconds, using or not using the VAD filter, on two GPU types. Bold is the setup that we later use.

Hardware For benchmarking, we use NVIDIA A40 GPUs. We run Whisper on a computer in a cluster that is used by other processes at the same time, which may allocate the same resources (except the GPU itself) and influence the latency. Since it is not always possible to have a dedicated server for a given service, this makes our evaluation very realistic. Since there will be variations in the latency metrics, we report mean and standard deviations.

Ensuring Reproducibility We simulate real-time processing of long-form transcription and record the times when Whisper emitted the outputs. We run the simulation on computers in a cluster that is not entirely under our control. For our simulation process, we block one GPU and a sufficient number of CPUs and RAM capacity. However, it can happen that other processes run at the same time, making a CPU and RAM load that is unpredictably slowing down our simulation. If the MinChunkSize is smaller than the time for processing an update, then two runs of the same simulation have different segmentation to chunks, leading to different WER and latency.

Therefore, we run a simulation of the same setup of one document 10 times, to measure the standard deviation of the latency and quality. The setup is the English transcription of the ESIC dev.20080925.013_007 document that is 3 minutes 36 seconds long, on NVIDIA A40 or L40 GPU with 48GB GPU RAM, 8 blocked CPU cores, and 200GB of CPU RAM, with or without VAD filter, with MinChunkSize 0.1 seconds.

The results are in Table 8.3. We observe a small, negligible standard deviation in WER, below or near 1%. The standard deviation in the average latency is much larger, from 0.36 to 0.81 seconds depending on the setup. We conclude that we must be aware of the standard deviation of latency due to uncontrollable computation conditions.

8.4.4 Results

We evaluated Whisper-Streaming with various setups for English, German, and Czech ASR. We first show the impact of outliers and voice activity detection (VAD) to determine optimal settings, and then present our main results with these settings.

Outliers After processing many setups, we observed extraordinarily high WER on the English ASR of the document titled dev2.20101213.015_018_EN_Gallagher. We realized it was due to noise in the ESIC dataset. The first half of the mentioned document is in Irish, and not English as intended. Only the English part is transcribed in gold, but Whisper transcribed both, leading to a more precise transcription than the reference. Except for the Gallagher document, all the reported setups achieved WER between 0 and 52%, and average latency between 0 and 16.1 seconds.

Voice activity detection We studied the effect of the VAD filter that is integrated within the Whisper backend. The results are in Table 8.4 and Figure 8.7. We realized that in the ESIC corpus, it is advisable to deactivate the VAD filter for the English original speech because it is very fluent, not interleaved with silence, and has no non-voice sounds. Without VAD, the quality remains nearly the same (difference within 0.2% WER), and the average latency was substantially lower, between 0.23 to 0.41 seconds.

For the processing of simultaneous interpreting, we recommend activating the VAD filter. The speech of a simultaneous interpreter contains many pauses, especially when waiting for context. With VAD, the latency was only 0.1 seconds larger, because VAD often filters out silence, which reduces the processing load. The quality with VAD was substantially higher, by 2 to 3 % WER with shorter MinChunkSize on German. With large chunk sizes, the quality is nearly the same (0.3 % WER difference with 2 seconds MinChunkSize) because a large chunk size causes the model to have a large context and thus a low chance of risking uncertain output. Therefore, we activated VAD for German and Czech simultaneous interpreting, and we deactivated it for English original speech.

For a real-life setup, we recommend starting Whisper-Streaming shortly before the speech actually starts, so that the first words are not missed, along with turning the VAD filter on so that the silence and non-voice sounds do not cause Whisper to make mistakes. If reducing the latency is important, an adaptive protocol for setting VAD on and off can be implemented.

		avg. % WER			avg. latency [s]		
	m.ch.	off	on	diff	off	on	diff
en	0.1s	8.4	8.3	-0.1	**3.30**	3.72	+0.41
	0.5s	8.5	8.3	-0.2	**3.27**	3.54	+0.27
	1.0s	8.1	8.1	+0.1	**3.62**	3.88	+0.26
	2.0s	8.0	7.9	-0.0	**5.45**	5.68	+0.23
de	0.1s	12.8	**9.7**	-3.1	3.83	3.93	+0.10
	0.5s	12.3	**9.5**	-2.8	3.97	4.11	+0.14
	1.0s	11.4	**9.4**	-2.0	4.19	4.37	+0.18
	2.0s	9.6	**9.3**	-0.3	5.79	5.94	+0.15

Table 8.4: Impact of the VAD filter on the WER and latency on ESIC dev on the streaming ASR with different minimum chunk size (m.ch., in seconds) of the English original speech (en) and German simultaneous interpreting (de). We highlight the remarkable benefit in bold: the original speech without pauses is processed with lower latency (by 0.23 seconds or more) and comparable quality with VAD off. On the contrary, the VAD on achieves higher quality for interpreting with frequent pauses, with a small difference in latency.

Performance Table 8.5 and Figure 8.8 summarize the WER and average latency of Whisper-Streaming on ESIC validation set for the three language tracks. Overall, with 1 second MinChunkSize, the average computationally aware latency is 3.6 seconds for English, 4.4 seconds for German, and 4.8 seconds for Czech, while the WER is by 0.2% higher than in the offline mode for English and German, and by 0.6% higher for Czech. Both WER and latency are the lowest in English, followed by German and Czech. This is related to the amount of language-specific data used for training Whisper, as well as the morphological complexity of these languages. The latency increases with larger uncertainty because it requires more updates for an agreement. Moreover, the larger MinChunkSize, the larger the latency, but the higher the quality because the system has sufficient context.

Offline mode WER We contrast the results with setups that serve as maximum performance estimates. One of them is offline mode in which processing of the whole audio document is done after recording, without any limitations on processing time. It is the default and most optimized setup for Whisper. The WER in offline mode and with VAD is lower than in streaming mode because the context size is not restricted. The model can use even the right (future) context that is unavailable or limited in streaming mode. Moreover, the internal segmentation of the long-form speech into processing chunks is optimized in the offline mode.

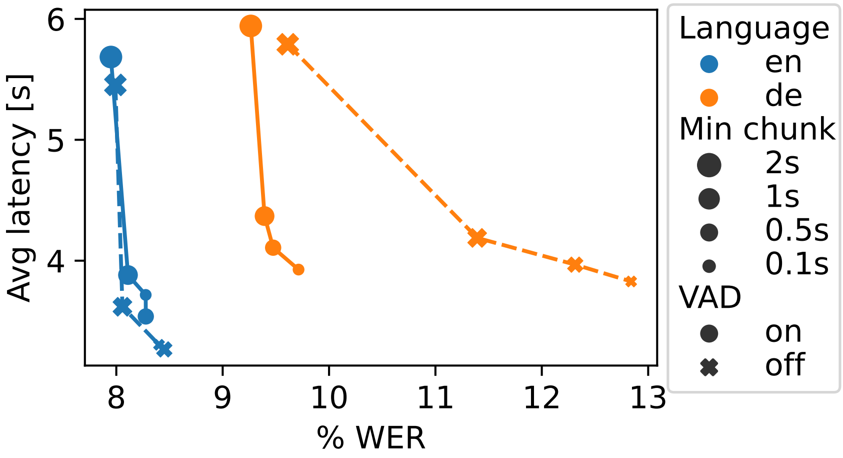


Figure 8.7: Impact of the VAD filter on latency and quality. The striking difference in VAD activated or deactivated for English vs. German is due to German being the speech of an interpreter.

Computationally unaware latency Another contrastive setup is computationally unaware simulation. It uses an unrealistic assumption that computation for Whisper processing any audio segment is instant so that the latency caused by computation is not included in the latency measurement. The measurement includes latency caused by uncertainty in the language. The gap between latency in computationally unaware and aware evaluation can be reduced by optimizing the hardware or inference algorithm. Computationally unaware latency can be reduced by improving the model or streaming policy.

We observe that the average computationally unaware latency is approximately twice the chunk size. This is expected because we use a local agreement of two consecutive updates. However, the processing of English is actually faster, a little less than twice the chunk size. We hypothesize that this could be caused by the anticipation ability of the Whisper model. The second possible reason is the inaccuracy of the gold timestamps in ESIC. The timestamps were computed by automatic forced alignment, and thus they may be less accurate in non-standard situations such as overlapping and non-transcribed speech, e.g. hesitations and foreign language insertions.

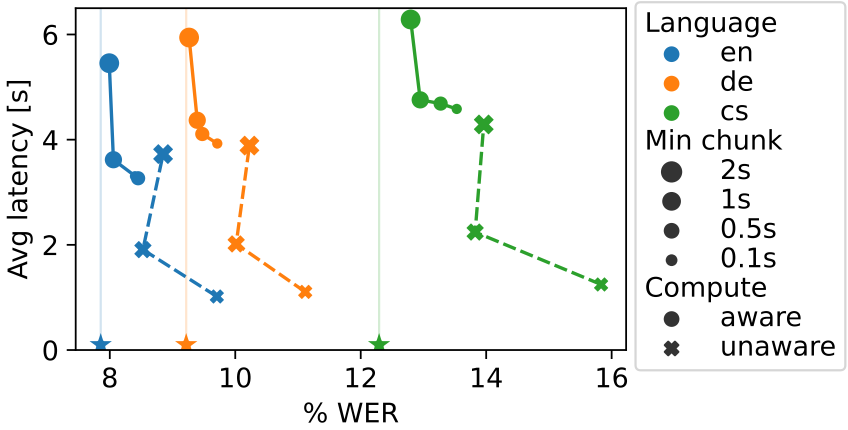


Figure 8.8: Latency and quality in computationally aware and unaware simulations (solid lines and dots vs. dashed lines and crosses), together with offline WER (stars and light vertical lines). VAD is deactivated for English, and activated for the other two.

8.4.5 Demonstration

Demonstration video is available at `https://vimeo.com/840442741`. It is a screencast video of Whisper-Streaming real-time outputs that processes live ASR on one ESIC document in three parallel instances for English, German, and Czech speech, the original and simultaneous interpreting. The video shows a contrast to gold transcripts with original timing so that the latency can be observed. The video also contains color highlighting for ASR errors.

Integration with ELITR To demonstrate practical usability, we integrate Whisper-Streaming with the ELITR (European Live Translator, Bojar et al., 2020) framework for complex distributed systems for multi-source and multi-target live speech transcription and translation (Bojar et al., 2021a). Within Whisper-Streaming, we implement and release a server that is connected as a worker to the Mediator server (Franceschini et al., 2020). A mediator allows a client to request a service from a worker. The client is then allowed to further process the text outputs received by the worker, e.g. translate them with another worker and present them at the web view server that delivers real-time captions to event participants during a live multi-lingual event.

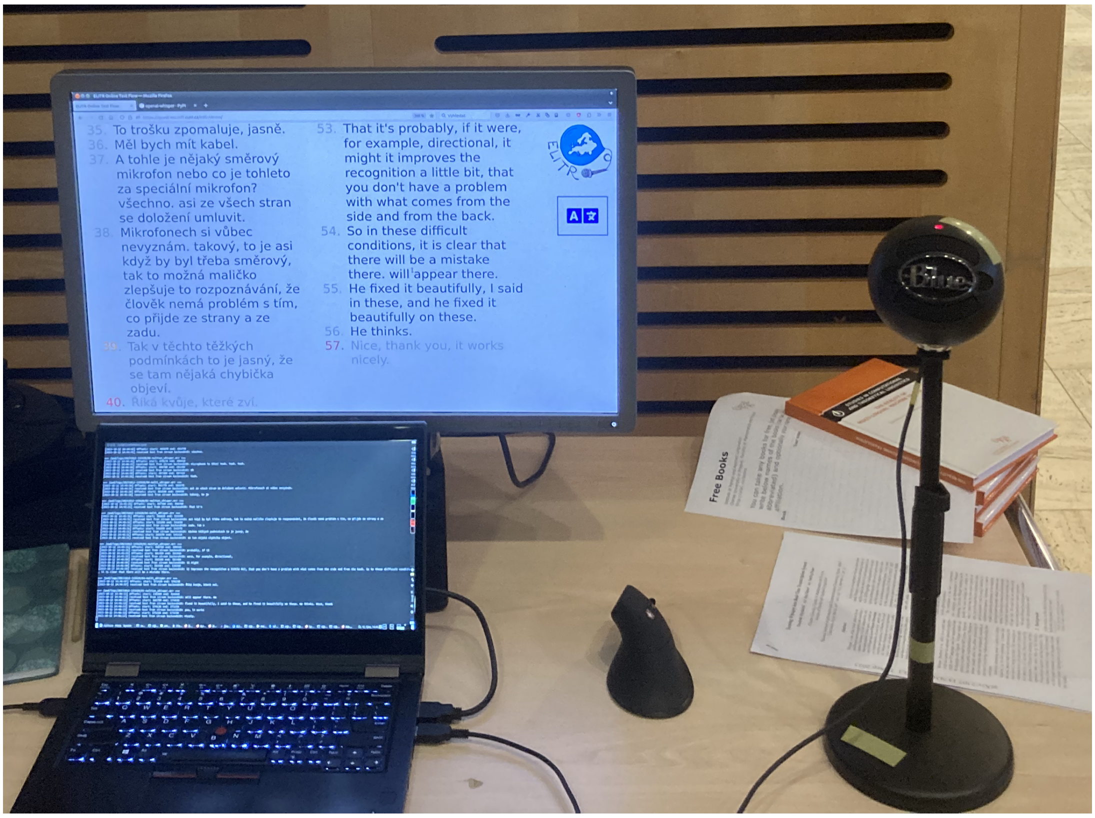
	% WER		% WER		latency [s]		
lang.	offline	m.ch.	un.	aw.	un.	aw.	diff
en	7.9	0.5s	9.7	8.5	1.02	3.27	+2.25
		1.0s	8.5	8.1	1.91	3.62	+1.71
		2.0s	8.8	8.0	3.73	5.45	+1.73
de	9.2	0.5s	11.1	9.5	1.11	4.11	+3.00
		1.0s	10.0	9.4	2.02	4.37	+2.35
		2.0s	10.2	9.3	3.89	5.94	+2.05
cs	12.3	0.5s	15.8	13.3	1.25	4.69	+3.44
		1.0s	13.8	12.9	2.24	4.76	+2.51
		2.0s	14.0	12.8	4.29	6.29	+2.00

Table 8.5: WER and an average latency of Whisper-Streaming on ESIC dev set in three language tracks using different MinChunkSize ("m.ch."). The realistic setup is computationally aware ("aw."), put into contrast with offline WER ("offline") and with the computationally unaware simulation ("un."). The data are the same as in Figure 8.8.

Evaluation event We evaluated Whisper-Streaming as a component in an experimental live speech translation service at a multi-lingual conference. For this, we built a pipeline that used five parallel Whisper-Streaming workers, three of them for ASR only (English, Czech and Ukrainian), and two for speech translation (Czech-to-English and Ukrainian-to-English). There were three parallel language streams at the conference, Czech, English, and Ukrainian. One of the languages was spoken on the main floor, and the others were provided by human simultaneous interpreting.

A human operator (as in Bojar et al., 2021b) was controlling the technical setup and the outputs using the language knowledge and had the option to redirect the streams, if necessary. The qualitative evaluation at the event showed that Whisper-Streaming is a robust and reliable part of the service, reaching acceptable latency and unexpectedly high quality in English, Czech, and Ukrainian long-form speech.

Interactive demonstration We have prepared an interactive demonstration of Whisper-Streaming using the ELITR framework. We can simulate speech sources from audio recordings, or allow demo participants to speak into a microphone in any of the 97 languages supported by Whisper, and observe the real-time outputs. Figure 8.9 is a photograph from the interactive demonstration.



35. To trošku zpomaluje, jasně.
36. Měl bych mít kabel.
37. A tohle je nějaký směrový mikrofon nebo co je tohleto za speciální mikrofon? všechno. asi ze všech stran se doložení umluvit.
38. Mikrofonech si vůbec nevyznám. takový, to je asi když by byl třeba směrový, tak to možná maličko zlepšuje to rozpoznávání, že člověk nemá problém s tím, co přijde ze strany a ze zadu.
39. Tak v těchto těžkých podmínkách to je jasný, že se tam nějaká chybička objeví.
40. Říká kvůje, které zví.

53. That it's probably, if it were, for example, directional, it might it improves the recognition a little bit, that you don't have a problem with what comes from the side and from the back.
54. So in these difficult conditions, it is clear that there will be a mistake there. will appear there.
55. He fixed it beautifully, I said in these, and he fixed it beautifully on these.
56. He thinks.
57. Nice, thank you, it works nicely.

Figure 8.9: A photograph of the interactive demonstration of Whisper-Streaming. A participant was speaking Czech into the microphone on the right. The speech was sent by the laptop computer to the Whisper-Streaming server running in a remote cluster. Then, the transcripts and translations were posted on the ELITR presentation server and displayed on the screen. There are Czech transcripts in the left column and Czech-to-English translations in the right.

8.5 Summary

In this chapter, we described the general strategy that the future research may take to propose methods for multi-source SST for realistic use case. We evaluated the late averaging multi-sourcing model with real audio and state-of-the-art ASR systems, and did not observe significant benefits of multi-sourcing using this model, both using automatic MT metric BertScore and human evaluation.

Then, we proposed and superficially surveyed and analyzed directions for future research. They include improving the multi-source combination ability of the model, using a more challenging domain, adapting for interpreting style, etc.

The first research option that we decided to focus on was evaluation with state-of-the-art streaming ASR. For that, we implemented and published Whisper-Streaming, an innovative tool that combines simultaneous streaming mode with the state-of-the-art Whisper ASR and speech translation (ST) model that is by default available only for offline mode.

Whisper-Streaming received lots of public interest and appreciating feedback, documenting e.g. by 6 external contributors who delivered pull requests with bug fixes or useful new features, 103 repository forks and 726 stars on GitHub (on 29th February 2024). There are many developers who integrate it into their applications, including e.g. videoconferencing, mobile applications, live speech lecture translation,[12] automatic minuting (Kmječ, 2023), etc.

[12] `https://github.com/ufal/correctable-lecture-translator`

9
Conclusion

In this chapter, we conclude our thesis. We remind the most important points, but not all of them. We summarized the whole thesis in brief points in Section 1.1, and our publications in Section 1.2.

Conclusion In this thesis, we investigated methods for multi-source simultaneous speech translation from the original and parallel simultaneous interpreting. The expected benefits of this method are improvements in quality and expected cost is acceptably higher latency. The simultaneous speech translation could then help to overcome language barrier in multi-lingual conferences and meetings where human simultaneous interpreting is not used for capacity reasons, but the simultaneity is necessary to enable real-time interaction between the speaker and audience.

Main finding Our results show that multi-sourcing may bring quality gains. We set the foundations for this task and advanced the state of the art, however, more research is necessary to design a multi-sourcing model that would be applicable to a real-life use case. We experimented with a late averaging model that showed the robustness of multi-sourcing to transcription errors in simplified simulated conditions – using parallel and aligned text sentences of the original and interpreting in the ESIC corpus, however, this model did not show improvements with real automatic transcripts on parallel audio. However, it can be improved by further research in the directions that we propose in Section 8.3.

Main contributions During our work, we advanced the state of the art with the following main contributions (we briefly repeat Section 1.1):

1. We created ESIC evaluation corpus (Chapter 4, Macháček et al., 2021).

2. We analyzed simultaneous interpreting in simultaneous speech translation (SST) (Chapter 5, Macháček et al., 2021).
3. We found robustness of multi-sourcing to automatic speech recognition (ASR) noise (Chapter 6, Macháček et al., 2023c).
4. We confirmed the previously untested assumption that the machine translation (MT) metrics can be used in SST (Chapter 7, Macháček et al., 2023a).
5. We implemented Whisper-Streaming, a very practical and innovative tool for long-form simultaneous speech-to-text transcription and translation demonstrating the state-of-the-art (Chapter 8, Macháček et al., 2023b).
6. We thoroughly describe all the relevant and potentially useful information in context (this whole thesis, Chapters 1-9).

Future work Our main task, multi-source SST, may be advanced in the areas that we propose in Section 8.3, e.g. designing multi-source model architecture, interpreting style training, using more than two sources for voting, confidence scores, etc.

Future related work While working on various chapters in this thesis, we found opportunities for research and innovations in the areas that are related to multi-source SST. For example, the hundreds of hours of multi-parallel interpreting data from the European Parliament that we downloaded in Chapter 4 can be processed into the training corpus, or be used for multi-lingual interpreting analysis. During work on Chapter 5 we found that the simultaneous speech translation could be inspired by simultaneous interpreting, e.g. summarization, removing redundancies, shortening, and segmentation to translation units. In Chapter 7 we suggested that the inter-annotator agreement of Continuous Rating and the question of human parity in SST should be studied. During work on Whisper-Streaming in Chapter 8, we found an opportunity to enhance the quality with out-of-vocabulary words included in the prompts. They could be inserted by a human post-editor, or found in slide texts that are relevant to the speech and known in advance.

Bibliography

Afshan, A. – Kumar, K. – Wu, J. Sequence-Level Confidence Classifier for ASR Utterance Accuracy and Application to Acoustic Models. In *Proc. Interspeech 2021*, p. 4084–4088, 2021. doi: 10.21437/Interspeech.2021-1666.

Agarwal, M. et al. FINDINGS OF THE IWSLT 2023 EVALUATION CAMPAIGN. In Salesky, E. – Federico, M. – Carpuat, M. (Ed.) *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023)*, p. 1–61, Toronto, Canada (in-person and online), July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.iwslt-1.1. Available at: `https://aclanthology.org/2023.iwslt-1.1`.

Aharoni, R. – Johnson, M. – Firat, O. Massively Multilingual Neural Machine Translation. In Burstein, J. – Doran, C. – Solorio, T. (Ed.) *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, p. 3874–3884, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1388. Available at: `https://aclanthology.org/N19-1388`.

Alam, T. – Khan, A. – Alam, F. Punctuation Restoration using Transformer Models for High-and Low-Resource Languages. In Xu, W. – Ritter, A. – Baldwin, T. – Rahimi, A. (Ed.) *Proceedings of the Sixth Workshop on Noisy User-generated Text (W-NUT 2020)*, p. 132–142, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.wnut-1.18. Available at: `https://aclanthology.org/2020.wnut-1.18`.

Ali, A. – Renals, S. Word Error Rate Estimation for Speech Recognition: e-WER. In Gurevych, I. – Miyao, Y. (Ed.) *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, p. 20–24, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-2004. Available at: `https://aclanthology.org/P18-2004`.

Anastasopoulos, A. et al. FINDINGS OF THE IWSLT 2021 EVALUATION CAMPAIGN. In Federico, M. – Waibel, A. – Costa-jussà, M. R. – Niehues, J. – Stuker, S. – Salesky, E. (Ed.) *Proceedings of the 18th International Conference on Spoken Language Translation (IWSLT 2021)*, p. 1–29, Bangkok, Thailand (online), August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.iwslt-1.1. Available at: `https://aclanthology.org/2021.iwslt-1.1`.

ANASTASOPOULOS, A. et al. Findings of the IWSLT 2022 Evaluation Campaign. In SALESKY, E. – FEDERICO, M. – COSTA-JUSSÀ, M. (Ed.) *Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT 2022)*, p. 98–157, Dublin, Ireland (in-person and online), May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.iwslt-1.10. Available at: `https://aclanthology.org/2022.iwslt-1.10`.

ANSARI, E. et al. FINDINGS OF THE IWSLT 2020 EVALUATION CAMPAIGN. In FEDERICO, M. – WAIBEL, A. – KNIGHT, K. – NAKAMURA, S. – NEY, H. – NIEHUES, J. – STÜKER, S. – WU, D. – MARIANI, J. – YVON, F. (Ed.) *Proceedings of the 17th International Conference on Spoken Language Translation*, p. 1–34, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.iwslt-1.1. Available at: `https://aclanthology.org/2020.iwslt-1.1`.

ANSARI, E. – BOJAR, O. – HADDOW, B. – MAHMOUDI, M. SLTEV: Comprehensive Evaluation of Spoken Language Translation. In GKATZIA, D. – SEDDAH, D. (Ed.) *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, p. 71–79, Online, April 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.eacl-demos.9. Available at: `https://aclanthology.org/2021.eacl-demos.9`.

ARDILA, R. – BRANSON, M. – DAVIS, K. – HENRETTY, M. – KOHLER, M. – MEYER, J. – MORAIS, R. – SAUNDERS, L. – TYERS, F. M. – WEBER, G. Common Voice: A Massively-Multilingual Speech Corpus. In *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, p. 4211–4215, 2020.

ARIVAZHAGAN, N. – CHERRY, C. – TE, I. – MACHEREY, W. – BALJEKAR, P. – FOSTER, G. F. Re-Translation Strategies for Long Form, Simultaneous, Spoken Language Translation. *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2020a, p. 7919–7923.

ARIVAZHAGAN, N. – BAPNA, A. – FIRAT, O. – LEPIKHIN, D. – JOHNSON, M. – KRIKUN, M. – CHEN, M. X. – CAO, Y. – FOSTER, G. F. – CHERRY, C. – MACHEREY, W. – CHEN, Z. – WU, Y. Massively Multilingual Neural Machine Translation in the Wild: Findings and Challenges. *CoRR*. 2019, abs/1907.05019. Available at: `http://arxiv.org/abs/1907.05019`.

ARIVAZHAGAN, N. – CHERRY, C. – MACHEREY, W. – FOSTER, G. Re-translation versus Streaming for Simultaneous Translation. In FEDERICO, M. – WAIBEL, A. – KNIGHT, K. – NAKAMURA, S. – NEY, H. – NIEHUES, J. – STÜKER, S. – WU, D. – MARIANI, J. – YVON, F. (Ed.) *Proceedings of the 17th International Conference on Spoken Language Translation*, p. 220–227, Online, July 2020b. Association for Computational Linguistics. doi: 10.18653/v1/2020.iwslt-1.27. Available at: `https://aclanthology.org/2020.iwslt-1.27`.

Bañón, M. et al. ParaCrawl: Web-Scale Acquisition of Parallel Corpora. In Jurafsky, D. – Chai, J. – Schluter, N. – Tetreault, J. (Ed.) *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, p. 4555–4567, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.417. Available at: `https://aclanthology.org/2020.acl-main.417`.

Barrault, L. et al. SeamlessM4T—Massively Multilingual & Multimodal Machine Translation. *ArXiv*. 2023.

Bartz, C. – Herold, T. – Yang, H. – Meinel, C. Language Identification Using Deep Convolutional Recurrent Neural Networks. In Liu, D. – Xie, S. – Li, Y. – Zhao, D. – El-Alfy, E.-S. M. (Ed.) *Neural Information Processing*, p. 880–889, Cham, 2017. Springer International Publishing. ISBN 978-3-319-70136-3.

Bentivogli, L. – Cettolo, M. – Gaido, M. – Karakanta, A. – Martinelli, A. – Negri, M. – Turchi, M. Cascade versus Direct Speech Translation: Do the Differences Still Make a Difference? In Zong, C. – Xia, F. – Li, W. – Navigli, R. (Ed.) *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, p. 2873–2887, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.224. Available at: `https://aclanthology.org/2021.acl-long.224`.

Bernardini, S. – Ferraresi, A. – Milicevic, M. From EPIC to EPTIC — Exploring simplification in interpreting and translation from an intermodal perspective. *Target*. 05 2016, 28, p. 61–86. doi: 10.1075/target.28.1.03ber.

Bojar, O. – Zeman, D. – Dušek, O. – Břečková, J. – Farkačová, H. – Grošpic, P. – Kačenová, K. – Knechtová, E. – Koubová, A. – Lukavská, J. – Nováková, P. – Petrdlíková, J. Additional German-Czech reference translations of the WMT'11 test set, 2012. Available at: `http://hdl.handle.net/11858/00-097C-0000-0008-D259-7`. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Bojar, O. – Dušek, O. – Kocmi, T. – Libovický, J. – Novák, M. – Popel, M. – Sudarikov, R. – Variš, D. CzEng 1.6: Enlarged Czech-English Parallel Corpus with Processing Tools Dockered. In *Text, Speech, and Dialogue: 19th International Conference, TSD 2016*, no. 9924, p. 231–238. Masaryk University, Springer International Publishing, 2016. ISBN 978-3-319-45509-9.

Bojar, O. et al. ELITR: European Live Translator. In Martins, A. – Moniz, H. – Fumega, S. – Martins, B. – Batista, F. – Coheur, L. – Parra, C. – Trancoso, I. – Turchi, M. – Bisazza, A. – Moorkens, J. – Guerberof, A. – Nurminen, M. – Marg, L. – Forcada, M. L. (Ed.) *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, p. 463–464, Lisboa, Portugal, November 2020. European Association for Machine Translation. Available at: `https://aclanthology.org/2020.eamt-1.53`.

BOJAR, O. et al. ELITR Multilingual Live Subtitling: Demo and Strategy. In GKATZIA, D. – SEDDAH, D. (Ed.) *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, p. 271–277, Online, April 2021a. Association for Computational Linguistics. doi: 10.18653/v1/2021.eacl-demos.32. Available at: `https://aclanthology.org/2021.eacl-demos.32`.

BOJAR, O. – SRDEČNÝ, V. – KUMAR, R. – SMRŽ, O. – SCHNEIDER, F. – HADDOW, B. – WILLIAMS, P. – CANTON, C. Operating a Complex SLT System with Speakers and Human Interpreters. In TURCHI, M. – FANTINUOLI, C. (Ed.) *Proceedings of the 1st Workshop on Automatic Spoken Language Translation in Real-World Settings (ASLTRW)*, p. 23–34, Virtual, August 2021b. Association for Machine Translation in the Americas. Available at: `https://aclanthology.org/2021.mtsummit-asltrw.3`.

CALLISON-BURCH, C. – KOEHN, P. – MONZ, C. – ZAIDAN, O. Findings of the 2011 Workshop on Statistical Machine Translation. In CALLISON-BURCH, C. – KOEHN, P. – MONZ, C. – ZAIDAN, O. F. (Ed.) *Proceedings of the Sixth Workshop on Statistical Machine Translation*, p. 22–64, Edinburgh, Scotland, July 2011. Association for Computational Linguistics. Available at: `https://aclanthology.org/W11-2103`.

ČEŇKOVÁ, I. *Úvod do teorie tlumočení.* Česká komora tlumočníků znakového jazyka, o.s., 2008. 2. opravené vydání. ISBN 978-80-87218-09-9.

ČEŇKOVÁ, I. *Teoretické aspekty simultánního tlumočení: na materiálu rusko-českém a česko-ruském.* 99. Univerzita karlova, 1988.

ČEŠKA, P. Speech Reconstruction - Overview of State-of-the-art Systems. In *WDS'09 Proceedings of Contributed Papers*, p. 11–15, Praha, Czechia, 2009. Matfyzpress, Charles University. ISBN 978-80-7378-101-9.

CHANG, C.-C. – CHUANG, S.-P. – LEE, H.-y. Anticipation-Free Training for Simultaneous Machine Translation. In SALESKY, E. – FEDERICO, M. – COSTA-JUSSÀ, M. (Ed.) *Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT 2022)*, p. 43–61, Dublin, Ireland (in-person and online), May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.iwslt-1.5. Available at: `https://aclanthology.org/2022.iwslt-1.5`.

CHEN, P. – HELCL, J. – GERMANN, U. – BURCHELL, L. – BOGOYCHEV, N. – MICELI BARONE, A. V. – WALDENDORF, J. – BIRCH, A. – HEAFIELD, K. The University of Edinburgh's English-German and English-Hausa Submissions to the WMT21 News Translation Task. In BARRAULT, L. et al. (Ed.) *Proceedings of the Sixth Conference on Machine Translation*, p. 104–109, Online, November 2021. Association for Computational Linguistics. Available at: `https://aclanthology.org/2021.wmt-1.4`.

Chen, Q. – Chen, M. – Li, B. – Wang, W. Controllable Time-Delay Transformer for Real-Time Punctuation Prediction and Disfluency Detection. In *2020 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2020, Barcelona, Spain, May 4-8, 2020*, p. 8069–8073. IEEE, 2020. doi: 10.1109/ICASSP40776.2020.9053159. Available at: `https://doi.org/10.1109/ICASSP40776.2020.9053159`.

Chen, Y. – Yu, S. – Leng, M. Parallel Sequence Alignment Algorithm for Clustering System. In Wang, K. – Kovacs, G. L. – Wozny, M. – Fang, M. (Ed.) *Knowledge Enterprise: Intelligent Strategies in Product Design, Manufacturing, and Management*, p. 311–321, Boston, MA, 2006. Springer US. ISBN 978-0-387-34403-4.

Cho, E. – Niehues, J. – Waibel, A. H. Segmentation and punctuation prediction in speech language translation using a monolingual translation system. In *IWSLT*, 2012.

Cho, E. – Fügen, C. – Hermann, T. – Kilgour, K. – Mediani, M. – Mohr, C. – Niehues, J. – Rottmann, K. – Saam, C. – Stüker, S. – Waibel, A. A real-world system for simultaneous translation of German lectures. 01 2013, p. 3473–3477.

Chordia, V. PunKtuator: A Multilingual Punctuation Restoration System for Spoken and Written Text. In Gkatzia, D. – Seddah, D. (Ed.) *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, p. 312–320, Online, April 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.eacl-demos.37. Available at: `https://aclanthology.org/2021.eacl-demos.37`.

Christoffels, I. K. – de Groot, A. M. – Kroll, J. F. Memory and language skills in simultaneous interpreters: The role of expertise and language proficiency. *Journal of Memory and Language*. 2006, 54, 3, p. 324–345. ISSN 0749-596X. doi: https://doi.org/10.1016/j.jml.2005.12.004. Available at: `https://www.sciencedirect.com/science/article/pii/S0749596X05001476`.

Conneau, A. – Baevski, A. – Collobert, R. – Mohamed, A. – Auli, M. Unsupervised Cross-lingual Representation Learning for Speech Recognition, 2020.

Conneau, A. – Ma, M. – Khanuja, S. – Zhang, Y. – Axelrod, V. – Dalmia, S. – Riesa, J. – Rivera, C. – Bapna, A. Fleurs: Few-shot learning evaluation of universal representations of speech. In *2022 IEEE Spoken Language Technology Workshop (SLT)*, p. 798–805. IEEE, 2023.

Dabre, R. – Cromieres, F. – Kurohashi, S. Enabling Multi-Source Neural Machine Translation By Concatenating Source Sentences In Multiple Languages. In Kurohashi, S. – Fung, P. (Ed.) *Proceedings of Machine Translation Summit XVI: Research Track*, p. 96–107, Nagoya Japan, September 18 – September 22 2017. Available at: `https://aclanthology.org/2017.mtsummit-papers.8`.

Dabre, R. – Chu, C. – Kunchukuttan, A. A Survey of Multilingual Neural Machine Translation. *ACM Comput. Surv.* September 2020, 53, 5. ISSN 0360-0300. doi: 10.1145/3406095. Available at: `https://doi.org/10.1145/3406095`.

Dabre, R. – Imankulova, A. – Kaneko, M. – Chakrabarty, A. Simultaneous Multi-Pivot Neural Machine Translation. *CoRR*. 2021, abs/2104.07410. Available at: `https://arxiv.org/abs/2104.07410`.

Daró, V. Experimental Studies on Memory in Conference Interpretation. *Meta*. 1997, 42, 4, p. 622–628. doi: https://doi.org/10.7202/002484ar.

Defrancq, B. Corpus-based research into the presumed effects of short EVS. *Interpreting*. 04 2015, 17. doi: 10.1075/intp.17.1.02def.

Dessloch, F. – Ha, T.-L. – Müller, M. – Niehues, J. – Nguyen, T.-S. – Pham, N.-Q. – Salesky, E. – Sperber, M. – Stüker, S. – Zenkel, T. – Waibel, A. KIT Lecture Translator: Multilingual Speech Translation with One-Shot Learning. In Zhao, D. (Ed.) *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, p. 89–93, Santa Fe, New Mexico, August 2018. Association for Computational Linguistics. Available at: `https://aclanthology.org/C18-2020`.

Dewaele, J.-M. Why the Dichotomy 'L1 Versus LX User' is Better than 'Native Versus Nonnative Speaker'. *Applied Linguistics*. 01 2017, 39, 2, p. 236–240. ISSN 0142-6001. doi: 10.1093/applin/amw055. Available at: `https://doi.org/10.1093/applin/amw055`.

Di Gangi, M. A. – Cattoni, R. – Bentivogli, L. – Negri, M. – Turchi, M. MuST-C: a Multilingual Speech Translation Corpus. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, Minneapolis, MN, USA, June 2019.

Doi, K. – Sudoh, K. – Nakamura, S. Large-Scale English-Japanese Simultaneous Interpretation Corpus: Construction and Analyses with Sentence-Aligned Data. In Federico, M. – Waibel, A. – Costa-jussà, M. R. – Niehues, J. – Stuker, S. – Salesky, E. (Ed.) *Proceedings of the 18th International Conference on Spoken Language Translation (IWSLT 2021)*, p. 226–235, Bangkok, Thailand (online), August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.iwslt-1.27. Available at: `https://aclanthology.org/2021.iwslt-1.27`.

Dong, Q. – Zhu, Y. – Wang, M. – Li, L. Learning When to Translate for Streaming Speech. In Muresan, S. – Nakov, P. – Villavicencio, A. (Ed.) *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, p. 680–694, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.50. Available at: `https://aclanthology.org/2022.acl-long.50`.

ĎOUBALOVÁ, J. Kvalita v simultánním tlumočení – otázka definice kvality tlumočení a kognitivní přístup ke kvalitě SI jako strategickému rozhodovacímu procesu. *AUC PHILOL.* May 2020, 2019, 4, p. 45–57.

DUGAN, L. – WADHAWAN, A. – SPENCE, K. – CALLISON-BURCH, C. – MCGUIRE, M. – ZORDAN, V. Learning When to Speak: Latency and Quality Trade-offs for Simultaneous Speech-to-Speech Translation with Offline Models. In *Proc. INTERSPEECH 2023*, p. 5265–5266, 2023.

DYER, C. – CHAHUNEAU, V. – SMITH, N. A. A Simple, Fast, and Effective Reparameterization of IBM Model 2. In VANDERWENDE, L. – DAUMÉ III, H. – KIRCHHOFF, K. (Ed.) *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, p. 644–648, Atlanta, Georgia, June 2013. Association for Computational Linguistics. Available at: https://aclanthology.org/N13-1073.

EVANS, J. D. *Straightforward Statistics for the Behavioral Sciences.* Pacific Grove: Brooks/Cole Pub., 1996.

EŠNEROVÁ, K. Hledáme Dream Job: Tlumočnice. Skautský institut, https://youtu.be/f8z464rTC0Y, 2019.

FANTINUOLI, C. The technological turn in interpreting: the challenges that lie ahead. In *Proceedings of the conference Übersetzen und Dolmetschen 4.0. - Neue Wege im digitalen Zeitalter*, p. 334–354, 2019.

FIRAT, O. – CHO, K. – BENGIO, Y. Multi-Way, Multilingual Neural Machine Translation with a Shared Attention Mechanism. In KNIGHT, K. – NENKOVA, A. – RAMBOW, O. (Ed.) *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, p. 866–875, San Diego, California, June 2016a. Association for Computational Linguistics. doi: 10.18653/v1/N16-1101. Available at: https://aclanthology.org/N16-1101.

FIRAT, O. – SANKARAN, B. – AL-ONAIZAN, Y. – YARMAN VURAL, F. T. – CHO, K. Zero-Resource Translation with Multi-Lingual Neural Machine Translation. In SU, J. – DUH, K. – CARRERAS, X. (Ed.) *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, p. 268–277, Austin, Texas, November 2016b. Association for Computational Linguistics. doi: 10.18653/v1/D16-1026. Available at: https://aclanthology.org/D16-1026.

FRANCESCHINI, D. et al. Removing European Language Barriers with Innovative Machine Translation Technology. In REHM, G. – BONTCHEVA, K. – CHOUKRI, K. – HAJIČ, J. – PIPERIDIS, S. – VASIĻJEVS, A. (Ed.) *Proceedings of the 1st International Workshop on Language Technology Platforms*, p. 44–49, Marseille, France, May 2020. European Language Resources Association. Available at: https://aclanthology.org/2020.iwltp-1.7. ISBN 979-10-95546-64-1.

FREITAG, M. – REI, R. – MATHUR, N. – LO, C.-k. – STEWART, C. – AVRAMIDIS, E. – KOCMI, T. – FOSTER, G. – LAVIE, A. – MARTINS, A. F. T. Results of WMT22 Metrics Shared Task: Stop Using BLEU – Neural Metrics Are Better and More Robust. In KOEHN, P. et al. (Ed.) *Proceedings of the Seventh Conference on Machine Translation (WMT)*, p. 46–68, Abu Dhabi, United Arab Emirates (Hybrid), December 2022. Association for Computational Linguistics. Available at: `https://aclanthology.org/2022.wmt-1.2`.

FUKUDA, R. – KO, Y. – KANO, Y. – DOI, K. – TOKUYAMA, H. – SAKTI, S. – SUDOH, K. – NAKAMURA, S. NAIST Simultaneous Speech-to-Text Translation System for IWSLT 2022. In SALESKY, E. – FEDERICO, M. – COSTA-JUSSÀ, M. (Ed.) *Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT 2022)*, p. 286–292, Dublin, Ireland (in-person and online), May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.iwslt-1.25. Available at: `https://aclanthology.org/2022.iwslt-1.25`.

FUKUDA, R. – NISHIKAWA, Y. – KANO, Y. – KO, Y. – YANAGITA, T. – DOI, K. – MAKINAE, M. – SAKTI, S. – SUDOH, K. – NAKAMURA, S. NAIST Simultaneous Speech-to-speech Translation System for IWSLT 2023. In SALESKY, E. – FEDERICO, M. – CARPUAT, M. (Ed.) *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023)*, p. 330–340, Toronto, Canada (in-person and online), July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.iwslt-1.31. Available at: `https://aclanthology.org/2023.iwslt-1.31`.

GABZDILOVÁ, M. Pracovní paměť v simultánním tlumočení a její kapacita (Working Memory in Simultaneous Interpreting and Its Capacity). Magisterská diplomová práce (Master thesis), Ústav translatologie, Filozofická fakulta, Univerzita Karlova v Praze, 2008. Supervised by doc. PhDr. Ivana Čeňková, CSc., consultant Mgr. Šárka Timarová.

GAIDO, M. – PAPI, S. – FUCCI, D. – FIAMENI, G. – NEGRI, M. – TURCHI, M. Efficient yet Competitive Speech Translation: FBK@IWSLT2022. In SALESKY, E. – FEDERICO, M. – COSTA-JUSSÀ, M. (Ed.) *Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT 2022)*, p. 177–189, Dublin, Ireland (in-person and online), May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.iwslt-1.13. Available at: `https://aclanthology.org/2022.iwslt-1.13`.

GALE, W. A. – CHURCH, K. W. A Program for Aligning Sentences in Bilingual Corpora. *Computational Linguistics*. 1993, 19, 1, p. 75–102. Available at: `https://aclanthology.org/J93-1004`.

GIESHOFF, A. C. Interpreting quality and effort in expert and novice interpreters. In *YLMP 2021 : Book of Abstracts*, p. 47–48, Poznań, 2021. Adam Mickiewicz University. Available at: `http://ylmp2021.amu.edu.pl/wp-content/uploads/2021/04/BoA_YLMP2021.pdf`. 7th Young Linguists' Meeting in Poznań : Rethinking language and identity in the multilingual world, Poznań, Poland, 23-25 April 2021.

Gile, D. *Basic Concepts and Models for Interpreter and Translator Training.* John Benjamins, 1995.

Graham, Y. – Baldwin, T. – Mathur, N. Accurate Evaluation of Segment-level Machine Translation Metrics. In Mihalcea, R. – Chai, J. – Sarkar, A. (Ed.) *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, p. 1183–1191, Denver, Colorado, May–June 2015. Association for Computational Linguistics. doi: 10.3115/v1/N15-1124. Available at: https://aclanthology.org/N15-1124.

Gulati, A. – Qin, J. – Chiu, C.-C. – Parmar, N. – Zhang, Y. – Yu, J. – Han, W. – Wang, S. – Zhang, Z. – Wu, Y. – Pang, R. Conformer: Convolution-augmented Transformer for Speech Recognition. In *Proc. Interspeech 2020*, p. 5036–5040, 2020. doi: 10.21437/Interspeech.2020-3015.

Ha, T.-L. – Niehues, J. – Waibel, A. Toward Multilingual Neural Machine Translation with Universal Encoder and Decoder. In Cettolo, M. – Niehues, J. – Stüker, S. – Bentivogli, L. – Cattoni, R. – Federico, M. (Ed.) *Proceedings of the 13th International Conference on Spoken Language Translation*, Seattle, Washington D.C, December 8-9 2016. International Workshop on Spoken Language Translation. Available at: https://aclanthology.org/2016.iwslt-1.6.

He, H. – Boyd-Graber, J. – Daumé III, H. Interpretese vs. Translationese: The Uniqueness of Human Strategies in Simultaneous Interpretation. In Knight, K. – Nenkova, A. – Rambow, O. (Ed.) *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, p. 971–976, San Diego, California, June 2016. Association for Computational Linguistics. doi: 10.18653/v1/N16-1111. Available at: https://aclanthology.org/N16-1111.

Huang, Y. – Feng, X. – Geng, X. – Li, B. – Qin, B. Towards Higher Pareto Frontier in Multilingual Machine Translation. In Rogers, A. – Boyd-Graber, J. – Okazaki, N. (Ed.) *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, p. 3802–3818, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.211. Available at: https://aclanthology.org/2023.acl-long.211.

Iranzo-Sánchez, J. – Silvestre-Cerdà, J. A. – Jorge, J. – Roselló, N. – Giménez, A. – Sanchis, A. – Civera, J. – Juan, A. Europarl-ST: A Multilingual Corpus For Speech Translation Of Parliamentary Debates. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, p. 8229–8233, 2020.

Iranzo-Sánchez, J. – Jorge Cano, J. – Martos, A. – Giménez Pastor, A. – Garcés Díaz-Munío, G. – Baquero-Arnal, P. – Silvestre-Cerdà, J. A. – Civera Saiz, J. – Sanchis, A. – Juan, A. MLLP-VRAIN UPV systems for the IWSLT 2022 Simultaneous Speech Translation and Speech-to-Speech Translation tasks. In Salesky, E. – Federico, M. – Costa-jussà, M. (Ed.) *Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT 2022)*, p. 255–264, Dublin, Ireland (in-person and online), May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.iwslt-1.22. Available at: https://aclanthology.org/2022.iwslt-1.22.

Javorský, D. – Macháček, D. – Bojar, O. Continuous Rating as Reliable Human Evaluation of Simultaneous Speech Translation. In Koehn, P. et al. (Ed.) *Proceedings of the Seventh Conference on Machine Translation (WMT)*, p. 154–164, Abu Dhabi, United Arab Emirates (Hybrid), December 2022. Association for Computational Linguistics. Available at: https://aclanthology.org/2022.wmt-1.9.

Johnson, M. – Schuster, M. – Le, Q. V. – Krikun, M. – Wu, Y. – Chen, Z. – Thorat, N. – Viégas, F. – Wattenberg, M. – Corrado, G. – Hughes, M. – Dean, J. Google's Multilingual Neural Machine Translation System: Enabling Zero-Shot Translation. *Transactions of the Association for Computational Linguistics*. 2017, 5, p. 339–351. doi: 10.1162/tacl_a_00065. Available at: https://aclanthology.org/Q17-1024.

Jones, R. *Conference Interpreting Explained*. St. Jerome, 2002.

Joulin, A. – Grave, E. – Bojanowski, P. – Douze, M. – Jégou, H. – Mikolov, T. FastText.zip: Compressing text classification models. *arXiv preprint arXiv:1612.03651*. 2016a.

Joulin, A. – Grave, E. – Bojanowski, P. – Mikolov, T. Bag of Tricks for Efficient Text Classification. *arXiv preprint arXiv:1607.01759*. 2016b.

Junczys-Dowmunt, M. Dual Conditional Cross-Entropy Filtering of Noisy Parallel Corpora. In Bojar, O. et al. (Ed.) *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, p. 888–895, Belgium, Brussels, October 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-6478. Available at: https://aclanthology.org/W18-6478.

Junczys-Dowmunt, M. – Grundkiewicz, R. – Dwojak, T. – Hoang, H. – Heafield, K. – Neckermann, T. – Seide, F. – Germann, U. – Aji, A. F. – Bogoychev, N. – Martins, A. F. T. – Birch, A. Marian: Fast Neural Machine Translation in C++. In Liu, F. – Solorio, T. (Ed.) *Proceedings of ACL 2018, System Demonstrations*, p. 116–121, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-4020. Available at: https://aclanthology.org/P18-4020.

Khadivi, S. – Ney, H. Integration of Speech Recognition and Machine Translation in Computer-Assisted Translation. *IEEE Transactions on Audio, Speech, and Language Processing*. 2008, 16, 8, p. 1551–1564. doi: 10.1109/TASL.2008.2004301.

Kisler, T. – Reichel, U. – Schiel, F. Multilingual processing of speech via web services. *Computer Speech & Language*. 2017, 45, p. 326 – 347. doi: http://dx.doi.org/10.1016/j.csl.2017.01.005.

KLOUDOVÁ, V. – MRAČEK, D. – BOJAR, O. – POPEL, M. Možnosti a meze tvorby tzv. optimálních referenčních překladů: po stopách „překladatelštiny“ v profesionálních překladech zpravodajských textů. *Slovo a slovesnost*. 2023, 84, 2, p. 122–156. ISSN 0037-7031.

KMJEČ, F. Methods of User-Assisted Summarization of Meetings. Bachelor thesis, Institute of Formal and Applied Linguistics, Faculty of Mathematics and Physics, Charles University, 2023. Supervised by doc. RNDr. Ondřej Bojar, Ph.D.

KOCMI, T. – FEDERMANN, C. Large Language Models Are State-of-the-Art Evaluators of Translation Quality. In NURMINEN, M. et al. (Ed.) *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, p. 193–203, Tampere, Finland, June 2023. European Association for Machine Translation. Available at: `https://aclanthology.org/2023.eamt-1.19`.

KOCMI, T. – POPEL, M. – BOJAR, O. Announcing CzEng 2.0 Parallel Corpus with over 2 Gigawords. *arXiv preprint arXiv:2007.03006*. 2020.

KOCMI, T. – MACHÁČEK, D. – BOJAR, O. *The Reality of Multi-Lingual Machine Translation*. ÚFAL, 2021.

KOCMI, T. et al. Findings of the 2022 Conference on Machine Translation (WMT22). In KOEHN, P. et al. (Ed.) *Proceedings of the Seventh Conference on Machine Translation (WMT)*, p. 1–45, Abu Dhabi, United Arab Emirates (Hybrid), December 2022. Association for Computational Linguistics. Available at: `https://aclanthology.org/2022.wmt-1.1`.

KOCMI, T. et al. Findings of the 2023 Conference on Machine Translation (WMT23): LLMs Are Here but Not Quite There Yet. In KOEHN, P. – HADDOW, B. – KOCMI, T. – MONZ, C. (Ed.) *Proceedings of the Eighth Conference on Machine Translation*, p. 1–42, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.wmt-1.1. Available at: `https://aclanthology.org/2023.wmt-1.1`.

KOEHN, P. Europarl: A Parallel Corpus for Statistical Machine Translation. In *Proceedings of Machine Translation Summit X: Papers*, p. 79–86, Phuket, Thailand, September 13-15 2005. Available at: `https://aclanthology.org/2005.mtsummit-papers.11`.

KOEHN, P. – OCH, F. J. – MARCU, D. Statistical Phrase-Based Translation. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, p. 127–133, 2003. Available at: `https://aclanthology.org/N03-1017`.

Koehn, P. – Hoang, H. – Birch, A. – Callison-Burch, C. – Federico, M. – Bertoldi, N. – Cowan, B. – Shen, W. – Moran, C. – Zens, R. – Dyer, C. – Bojar, O. – Constantin, A. – Herbst, E. Moses: Open Source Toolkit for Statistical Machine Translation. In Ananiadou, S. (Ed.) *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, p. 177–180, Prague, Czech Republic, June 2007. Association for Computational Linguistics. Available at: `https://aclanthology.org/P07-2045`.

Kratochvíl, J. – Polák, P. – Bojar, O. Large Corpus of Czech Parliament Plenary Hearings. In *Proceedings of the 12th Language Resources and Evaluation Conference*, p. 6363–6367, Marseille, France, May 2020. European Language Resources Association. Available at: `https://aclanthology.org/2020.lrec-1.781`. ISBN 979-10-95546-34-4.

Kuchaiev, O. – Li, J. – Nguyen, H. – Hrinchuk, O. – Leary, R. – Ginsburg, B. – Kriman, S. – Beliaev, S. – Lavrukhin, V. – Cook, J. – others. Nemo: a toolkit for building ai applications using neural modules. *arXiv preprint arXiv:1909.09577*. 2019.

Kudo, T. – Richardson, J. SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing. In Blanco, E. – Lu, W. (Ed.) *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, p. 66–71, Brussels, Belgium, November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-2012. Available at: `https://aclanthology.org/D18-2012`.

Laptev, A. – Ginsburg, B. Fast Entropy-Based Methods of Word-Level Confidence Estimation for End-to-End Automatic Speech Recognition. In *2022 IEEE Spoken Language Technology Workshop (SLT)*, p. 152–159, 2023. doi: 10.1109/SLT54892.2023.10022960.

Lederer, M. *Simultaneous Interpretation — Units of Meaning and other Features*, p. 323–332. 01 1978. doi: 10.1007/978-1-4615-9077-4_28. ISBN 978-1-4615-9079-8.

Li, B. Word Alignment in the Era of Deep Learning: A Tutorial, 11 2022.

Lison, P. – Tiedemann, J. OpenSubtitles2016: Extracting Large Parallel Corpora from Movie and TV Subtitles. In Calzolari, N. – Choukri, K. – Declerck, T. – Goggi, S. – Grobelnik, M. – Maegaard, B. – Mariani, J. – Mazo, H. – Moreno, A. – Odijk, J. – Piperidis, S. (Ed.) *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, p. 923–929, Portorož, Slovenia, May 2016. European Language Resources Association (ELRA). Available at: `https://aclanthology.org/L16-1147`.

Lita, L. V. – Ittycheriah, A. – Roukos, S. – Kambhatla, N. tRuEcasIng. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, p. 152–159, Sapporo, Japan, July 2003. Association for Computational Linguistics. doi: 10.3115/1075096.1075116. Available at: `https://aclanthology.org/P03-1020`.

Liu, D. – Spanakis, G. – Niehues, J. Low-Latency Sequence-to-Sequence Speech Recognition and Translation by Partial Hypothesis Selection. In *Proc. Interspeech 2020*, p. 3620–3624, 2020. doi: 10.21437/Interspeech.2020-2897.

Liu, L. – Zhu, M. Bertalign: Improved word embedding-based sentence alignment for Chinese–English parallel corpora of literary texts. *Digital Scholarship in the Humanities*. 12 2022, 38, 2, p. 621–634. ISSN 2055-7671. doi: 10.1093/llc/fqac089. Available at: `https://doi.org/10.1093/llc/fqac089`.

Liu, M. – Zhang, W. – Li, X. – Luan, J. – Wang, B. – Guo, Y. – Chen, S. Rethinking the Reasonability of the Test Set for Simultaneous Machine Translation. In *Proceedings of the 48th IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2023)*, 2023.

Lui, M. – Baldwin, T. langid.py: An Off-the-shelf Language Identification Tool. In Zhang, M. (Ed.) *Proceedings of the ACL 2012 System Demonstrations*, p. 25–30, Jeju Island, Korea, July 2012. Association for Computational Linguistics. Available at: `https://aclanthology.org/P12-3005`.

Ma, M. – Huang, L. – Xiong, H. – Zheng, R. – Liu, K. – Zheng, B. – Zhang, C. – He, Z. – Liu, H. – Li, X. – Wu, H. – Wang, H. STACL: Simultaneous Translation with Implicit Anticipation and Controllable Latency using Prefix-to-Prefix Framework. In Korhonen, A. – Traum, D. – Màrquez, L. (Ed.) *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, p. 3025–3036, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1289. Available at: `https://aclanthology.org/P19-1289`.

Ma, X. – Dousti, M. J. – Wang, C. – Gu, J. – Pino, J. SIMULEVAL: An Evaluation Toolkit for Simultaneous Translation. In Liu, Q. – Schlangen, D. (Ed.) *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, p. 144–150, Online, October 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-demos.19. Available at: `https://aclanthology.org/2020.emnlp-demos.19`.

Macháček, D. – Kratochvíl, J. – Vojtěchová, T. – Bojar, O. A Speech Test Set of Practice Business Presentations with Additional Relevant Texts. In Martín-Vide, C. – Purver, M. – Pollak, S. (Ed.) *Statistical Language and Speech Processing*, p. 151–161, Cham, 2019. Springer International Publishing. ISBN 978-3-030-31372-2.

Macháček, D. – Kratochvíl, J. – Sagar, S. – Žilinec, M. – Bojar, O. – Nguyen, T.-S. – Schneider, F. – Williams, P. – Yao, Y. ELITR Non-Native Speech Translation at IWSLT 2020. In Federico, M. – Waibel, A. – Knight, K. – Nakamura, S. – Ney, H. – Niehues, J. – Stüker, S. – Wu, D. – Mariani, J. – Yvon, F. (Ed.) *Proceedings of the 17th International Conference on Spoken Language Translation*, p. 200–208, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.iwslt-1.25. Available at: `https://aclanthology.org/2020.iwslt-1.25`.

Macháček, D. – Bojar, O. – Dabre, R. MT Metrics Correlate with Human Ratings of Simultaneous Speech Translation. In Salesky, E. – Federico, M. – Carpuat, M. (Ed.) *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023)*, p. 169–179, Toronto, Canada (in-person and online), July 2023a. Association for Computational Linguistics. doi: 10.18653/v1/2023.iwslt-1.12. Available at: `https://aclanthology.org/2023.iwslt-1.12`.

Macháček, D. – Dabre, R. – Bojar, O. Turning Whisper into Real-Time Transcription System. In Saha, S. – Sujaini, H. (Ed.) *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics: System Demonstrations*, p. 17–24, Bali, Indonesia, November 2023b. Association for Computational Linguistics. Available at: `https://aclanthology.org/2023.ijcnlp-demo.3`.

Macháček, D. – Polák, P. – Bojar, O. – Dabre, R. Robustness of Multi-Source MT to Transcription Errors. In Rogers, A. – Boyd-Graber, J. – Okazaki, N. (Ed.) *Findings of the Association for Computational Linguistics: ACL 2023*, p. 3707–3723, Toronto, Canada, July 2023c. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.228. Available at: `https://aclanthology.org/2023.findings-acl.228`.

Macháček, D. – Bojar, O. Presenting Simultaneous Translation in Limited Space. In *Proceedings of the 20th Conference Information Technologies - Applications and Theory (ITAT 2020)*, p. 32–37, Košice, Slovakia, 2020. Tomáš Horváth.

Macháček, D. – Žilinec, M. – Bojar, O. Lost in Interpreting: Speech Translation from Source or Interpreter? In *Interspeech 2021, 22nd Annual Conference of the International Speech Communication Association.* ISCA, 2021.

Martucci, G. – Cettolo, M. – Negri, M. – Turchi, M. Lexical Modeling of ASR Errors for Robust Speech Translation. In *Proc. Interspeech 2021*, p. 2282–2286, 2021. doi: 10.21437/Interspeech.2021-265.

Mathur, N. – Wei, J. – Freitag, M. – Ma, Q. – Bojar, O. Results of the WMT20 Metrics Shared Task. In Barrault, L. et al. (Ed.) *Proceedings of the Fifth Conference on Machine Translation*, p. 688–725, Online, November 2020. Association for Computational Linguistics. Available at: `https://aclanthology.org/2020.wmt-1.77`.

Matsuda, S. – Hu, X. – Shiga, Y. – Kashioka, H. – Hori, C. – Yasuda, K. – Okuma, H. – Uchiyama, M. – Sumita, E. – Kawai, H. – Nakamura, S. Multilingual Speech-to-Speech Translation System: VoiceTra. In *2013 IEEE 14th International Conference on Mobile Data Management*, 2, p. 229–233, 2013. doi: 10.1109/MDM.2013.99.

Matusov, E. – Leusch, G. – Bender, O. – Ney, H. Evaluating Machine Translation Output with Automatic Sentence Segmentation. In *Proceedings of the Second International Workshop on Spoken Language Translation*, Pittsburgh, Pennsylvania, USA, October 24-25 2005. Available at: `https://aclanthology.org/2005.iwslt-1.19`.

Michel, J.-B. – Shen, Y. K. – Aiden, A. P. – Veres, A. – Gray, M. K. – Team, T. G. B. – Pickett, J. P. – Hoiberg, D. – Clancy, D. – Norvig, P. – Orwant, J. – Pinker, S. – Nowak, M. A. – Aiden, E. L. Quantitative Analysis of Culture Using Millions of Digitized Books. *Science*. 2011, 331, 6014, p. 176–182. doi: 10.1126/science.1199644. Available at: `https://www.science.org/doi/abs/10.1126/science.1199644`.

Minixhofer, B. – Pfeiffer, J. – Vulić, I. Where's the Point? Self-Supervised Multilingual Punctuation-Agnostic Sentence Segmentation. In Rogers, A. – Boyd-Graber, J. – Okazaki, N. (Ed.) *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, p. 7215–7235, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.398. Available at: `https://aclanthology.org/2023.acl-long.398`.

Miranda, J. – Neto, J. P. – Black, A. W. Parallel combination of multilingual speech streams for improved ASR. In *INTERSPEECH 2012, 13th Annual Conference of the International Speech Communication Association, Portland, Oregon, USA, September 9-13, 2012*, p. 1027–1030. ISCA, 2012. Available at: `http://www.isca-speech.org/archive/interspeech_2012/i12_1027.html`.

Miranda, J. – Neto, J. P. – Black, A. W. Improved punctuation recovery through combination of multiple speech streams. In *2013 IEEE Workshop on Automatic Speech Recognition and Understanding*, p. 132–137, 2013. doi: 10.1109/ASRU.2013.6707718.

Misu, T. Network-based multi-lingual speech translation system: VoiceTra. In *2010 4th International Universal Communication Symposium*, p. 405–405, 2010. doi: 10.1109/IUCS.2010.5666754.

Müller, M. – Nguyen, T. S. – Niehues, J. – Cho, E. – Krüger, B. – Ha, T.-L. – Kilgour, K. – Sperber, M. – Mediani, M. – Stüker, S. – Waibel, A. Lecture Translator - Speech translation framework for simultaneous lecture translation. In DeNero, J. – Finlayson, M. – Reddy, S. (Ed.) *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, p. 82–86, San Diego, California, June 2016. Association for Computational Linguistics. doi: 10.18653/v1/N16-3017. Available at: `https://aclanthology.org/N16-3017`.

Newman, E. J. – Schwarz, N. Good Sound, Good Research: How Audio Quality Influences Perceptions of the Research and Researcher. *Science Communication*. 2018, 40, 2, p. 246–257. doi: 10.1177/1075547018759345. Available at: `https://doi.org/10.1177/1075547018759345`.

Nguyen, T.-S. – Stueker, S. – Waibel, A. Super-Human Performance in Online Low-latency Recognition of Conversational Speech, 2021a.

Nguyen, T.-S. – Stüker, S. – Waibel, A. Super-Human Performance in Online Low-Latency Recognition of Conversational Speech. In *Proc. Interspeech 2021*, p. 1762–1766, 2021b. doi: 10.21437/Interspeech.2021-1114.

Niehues, J. – Pham, N.-Q. – Ha, T.-L. – Sperber, M. – Waibel, A. Low-Latency Neural Speech Translation. In *Proc. Interspeech 2018*, p. 1293–1297, 2018. doi: 10.21437/Interspeech.2018-1055.

Nishimura, Y. – Sudoh, K. – Neubig, G. – Nakamura, S. Multi-Source Neural Machine Translation with Data Augmentation. In Turchi, M. – Niehues, J. – Frederico, M. (Ed.) *Proceedings of the 15th International Conference on Spoken Language Translation*, p. 48–53, Brussels, October 29-30 2018. International Conference on Spoken Language Translation. Available at: `https://aclanthology.org/2018.iwslt-1.7`.

NLLB Team et al. No Language Left Behind: Scaling Human-Centered Machine Translation, 2022.

Och, F. J. – Ney, H. Statistical multi-source translation. In Maegaard, B. (Ed.) *Proceedings of Machine Translation Summit VIII*, Santiago de Compostela, Spain, September 18-22 2001. Available at: `https://aclanthology.org/2001.mtsummit-papers.46`.

Olsen, B. S. Human Interpreter Training and Practice: Insights for Simultaneous Machine Translation Research. Invited talk at workshop AutoSimTrans 2020 at ACL, 2020.

OpenAI. Introducing ChatGPT, 11 2022. Available at: `https://openai.com/blog/chatgpt`.

Pan, J. The Chinese/English Political Interpreting Corpus (CEPIC): A New Electronic Resource for Translators and Interpreters. In *Proceedings of the Human-Informed Translation and Interpreting Technology Workshop (HiT-IT 2019)*, p. 82–88, Varna, Bulgaria, September 2019. Incoma Ltd., Shoumen, Bulgaria. doi: 10.26615/issn.2683-0078.2019_010. Available at: `https://aclanthology.org/W19-8710`.

Papastratis, I. Speech Recognition: a review of the different deep learning approaches. *https://theaisummer.com/*. 2021.

Papi, S. – Gaido, M. – Negri, M. Direct Models for Simultaneous Translation and Automatic Subtitling: FBK@IWSLT2023. In Salesky, E. – Federico, M. – Carpuat, M. (Ed.) *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023)*, p. 159–168, Toronto, Canada (in-person and online), July 2023a. Association for Computational Linguistics. doi: 10.18653/v1/2023.iwslt-1.11. Available at: `https://aclanthology.org/2023.iwslt-1.11`.

Papi, S. – Negri, M. – Turchi, M. Attention as a Guide for Simultaneous Speech Translation. In Rogers, A. – Boyd-Graber, J. – Okazaki, N. (Ed.) *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, p. 13340–13356, Toronto, Canada, July 2023b. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.745. Available at: `https://aclanthology.org/2023.acl-long.745`.

Papi, S. – Turchi, M. – Negri, M. AlignAtt: Using Attention-based Audio-Translation Alignments as a Guide for Simultaneous Speech Translation. In *Proc. INTERSPEECH 2023*, p. 3974–3978, 2023c. doi: 10.21437/Interspeech.2023-170.

Papineni, K. – Roukos, S. – Ward, T. – Zhu, W.-J. Bleu: a Method for Automatic Evaluation of Machine Translation. In Isabelle, P. – Charniak, E. – Lin, D. (Ed.) *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, p. 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics. doi: 10.3115/1073083.1073135. Available at: `https://aclanthology.org/P02-1040`.

Park, T. J. – Kanda, N. – Dimitriadis, D. – Han, K. J. – Watanabe, S. – Narayanan, S. A review of speaker diarization: Recent advances with deep learning. *Computer Speech & Language*. 2022, 72, p. 101317. ISSN 0885-2308. doi: https://doi.org/10.1016/j.csl.2021.101317. Available at: `https://www.sciencedirect.com/science/article/pii/S0885230821001121`.

Paszke, A. et al. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Advances in Neural Information Processing Systems 32*, p. 8024–8035. Curran Associates, Inc., 2019. Available at: `http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf`.

Paulik, M. – Stuker, S. – Fugen, C. – Schultz, T. – Schaaf, T. – Waibel, A. Speech translation enhanced automatic speech recognition. In *IEEE Workshop on Automatic Speech Recognition and Understanding, 2005.*, p. 121–126, 2005. doi: 10.1109/ASRU.2005.1566488.

Paulik, M. – Waibel, A. Automatic translation from parallel speech: Simultaneous interpretation as MT training data. In *2009 IEEE Workshop on Automatic Speech Recognition & Understanding*, p. 496–501, 2009. doi: 10.1109/ASRU.2009.5372880.

Paulik, M. – Waibel, A. Extracting clues from human interpreter speech for spoken language translation. In *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*, p. 5097–5100, 2008. doi: 10.1109/ICASSP.2008.4518805.

Pearson, K. X. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*. July 1900, 50, 302, p. 157–175. doi: 10.1080/14786440009463897. Available at: https://doi.org/10.1080/14786440009463897.

Polák, P. Long-form Simultaneous Speech Translation: Thesis Proposal. In *Proceedings of the 3nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 13th International Joint Conference on Natural Language Processing: Student Research Workshop*, Bali, Indonesia, November 2023. Association for Computational Linguistics.

Polák, P. – Sagar, S. – Macháček, D. – Bojar, O. CUNI Neural ASR with Phoneme-Level Intermediate Step for~Non-Native~SLT at IWSLT 2020. In Federico, M. – Waibel, A. – Knight, K. – Nakamura, S. – Ney, H. – Niehues, J. – Stüker, S. – Wu, D. – Mariani, J. – Yvon, F. (Ed.) *Proceedings of the 17th International Conference on Spoken Language Translation*, p. 191–199, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.iwslt-1.24. Available at: https://aclanthology.org/2020.iwslt-1.24.

Polák, P. – Pham, N.-Q. – Nguyen, T. N. – Liu, D. – Mullov, C. – Niehues, J. – Bojar, O. – Waibel, A. CUNI-KIT System for Simultaneous Speech Translation Task at IWSLT 2022. In Salesky, E. – Federico, M. – Costa-jussà, M. (Ed.) *Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT 2022)*, p. 277–285, Dublin, Ireland (in-person and online), May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.iwslt-1.24. Available at: https://aclanthology.org/2022.iwslt-1.24.

Polák, P. – Liu, D. – Pham, N.-Q. – Niehues, J. – Waibel, A. – Bojar, O. Towards Efficient Simultaneous Speech Translation: CUNI-KIT System for Simultaneous Track at IWSLT 2023. In Salesky, E. – Federico, M. – Carpuat, M. (Ed.) *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023)*, p. 389–396, Toronto, Canada (in-person and online), July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.iwslt-1.37. Available at: https://aclanthology.org/2023.iwslt-1.37.

Polák, P. – Yan, B. – Watanabe, S. – Waibel, A. – Bojar, O. Incremental Blockwise Beam Search for Simultaneous Speech Translation with Controllable Quality-Latency Tradeoff. In *Proc. INTERSPEECH 2023*, p. 3979–3983, 2023. doi: 10.21437/Interspeech.2023-2225.

Popović, M. chrF deconstructed: beta parameters and n-gram weights. In Bojar, O. et al. (Ed.) *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, p. 499–504, Berlin, Germany, August 2016. Association for Computational Linguistics. doi: 10.18653/v1/W16-2341. Available at: https://aclanthology.org/W16-2341.

Popović, M. chrF++: words helping character n-grams. In Bojar, O. – Buck, C. – Chatterjee, R. – Federmann, C. – Graham, Y. – Haddow, B. – Huck, M. – Yepes, A. J. – Koehn, P. – Kreutzer, J. (Ed.) *Proceedings of the Second Conference on Machine Translation*, p. 612–618, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-4770. Available at: `https://aclanthology.org/W17-4770`.

Post, M. A Call for Clarity in Reporting BLEU Scores. In Bojar, O. et al. (Ed.) *Proceedings of the Third Conference on Machine Translation: Research Papers*, p. 186–191, Brussels, Belgium, October 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-6319. Available at: `https://aclanthology.org/W18-6319`.

Povey, D. – Ghoshal, A. – Boulianne, G. – Goel, N. – Hannemann, M. – Qian, Y. – Schwarz, P. – Stemmer, G. The kaldi speech recognition toolkit. In *In IEEE 2011 workshop*, 2011.

Pratap, V. et al. Scaling Speech Technology to 1,000+ Languages. *arXiv*. 2023.

Przybyl, H. – Lapshinova-Koltunski, E. – Menzel, K. – Fischer, S. – Teich, E. EPIC UdS - Creation and Applications of a Simultaneous Interpreting Corpus. In Calzolari, N. – Béchet, F. – Blache, P. – Choukri, K. – Cieri, C. – Declerck, T. – Goggi, S. – Isahara, H. – Maegaard, B. – Mariani, J. – Mazo, H. – Odijk, J. – Piperidis, S. (Ed.) *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, p. 1193–1200, Marseille, France, June 2022. European Language Resources Association. Available at: `https://aclanthology.org/2022.lrec-1.127`.

Purchartová, P. Vybrané aspekty lingvistické analýzy výchozího textu z pohledu simultánního tlumočení a strojového překladu mluvené řeči z angličtiny do češtiny. Master thesis, Institute of Translation Studies, Faculty of Arts, Charles University, 2023. Supervised by Mgr. Věra Kloudová, Ph.D.

Radford, A. – Wu, J. – Child, R. – Luan, D. – Amodei, D. – Sutskever, I. Language Models are Unsupervised Multitask Learners. 2019.

Radford, A. – Kim, J. W. – Xu, T. – Brockman, G. – McLeavey, C. – Sutskever, I. Robust Speech Recognition via Large-Scale Weak Supervision, 2022.

Rei, R. – Stewart, C. – Farinha, A. C. – Lavie, A. Unbabel's Participation in the WMT20 Metrics Shared Task. In Barrault, L. et al. (Ed.) *Proceedings of the Fifth Conference on Machine Translation*, p. 911–920, Online, November 2020. Association for Computational Linguistics. Available at: `https://aclanthology.org/2020.wmt-1.101`.

Rei, R. – Treviso, M. – Guerreiro, N. M. – Zerva, C. – Farinha, A. C. – Maroti, C. – Souza, J. G. – Glushkova, T. – Alves, D. – Coheur, L. – Lavie, A. – Martins, A. F. T. CometKiwi: IST-Unbabel 2022 Submission for the Quality Estimation Shared Task. In Koehn, P. et al. (Ed.) *Proceedings of the Seventh Conference on Machine Translation (WMT)*, p. 634–645, Abu Dhabi, United Arab Emirates (Hybrid), December 2022. Association for Computational Linguistics. Available at: `https://aclanthology.org/2022.wmt-1.60`.

Reichel, U. D. Language-independent grapheme-phoneme conversion and word stress assignment as a web service. In Hoffmann, R. (Ed.) *Elektronische Sprachverarbeitung 2014*, 71. Dresden, Germany: TUDpress, 2014. p. 42–49.

Ren, Y. – Liu, J. – Tan, X. – Zhang, C. – Qin, T. – Zhao, Z. – Liu, T.-Y. SimulSpeech: End-to-End Simultaneous Speech to Text Translation. In Jurafsky, D. – Chai, J. – Schluter, N. – Tetreault, J. (Ed.) *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, p. 3787–3796, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.350. Available at: `https://aclanthology.org/2020.acl-main.350`.

Rouvier, M. – Dupuy, G. – Gay, P. – Khoury, E. – Merlin, T. – Meignier, S. An open-source state-of-the-art toolbox for broadcast news diarization. In *Proc. Interspeech 2013*, p. 1477–1481, 2013. doi: 10.21437/Interspeech.2013-383.

Rubino, R. – Fujita, A. – Marie, B. NICT Kyoto Submission for the WMT'21 Quality Estimation Task: Multimetric Multilingual Pretraining for Critical Error Detection. In Barrault, L. et al. (Ed.) *Proceedings of the Sixth Conference on Machine Translation*, p. 941–947, Online, November 2021. Association for Computational Linguistics. Available at: `https://aclanthology.org/2021.wmt-1.99`.

Ruiz, N. – Federico, M. Assessing the impact of speech recognition errors on machine translation quality. In Al-Onaizan, Y. – Simard, M. (Ed.) *Proceedings of the 11th Conference of the Association for Machine Translation in the Americas: MT Researchers Track*, p. 261–274, Vancouver, Canada, October 22-26 2014. Association for Machine Translation in the Americas. Available at: `https://aclanthology.org/2014.amta-researchers.20`.

Ruiz, N. – Gangi, M. A. D. – Bertoldi, N. – Federico, M. Assessing the Tolerance of Neural Machine Translation Systems Against Speech Recognition Errors. In *Proc. Interspeech 2017*, p. 2635–2639, 2017. doi: 10.21437/Interspeech.2017-1690.

Ryšlink, V. Methods of Input Segmentation for Simultaneous Speech Translation. Master thesis, Institute of Formal and Applied Linguistics, Faculty of Mathematics and Physics, Charles University, 2022. Supervised by doc. RNDr. Ondřej Bojar, Ph.D., consultant Mgr. Aleš Tamchyna, Ph.D.

SANDRELLI, A. – BENDAZZOLI, C. Tagging a Corpus of Interpreted Speeches: the European Parliament Interpreting Corpus (EPIC). In CALZOLARI, N. – CHOUKRI, K. – GANGEMI, A. – MAEGAARD, B. – MARIANI, J. – ODIJK, J. – TAPIAS, D. (Ed.) *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy, May 2006. European Language Resources Association (ELRA). Available at: `http://www.lrec-conf.org/proceedings/lrec2006/pdf/174_pdf.pdf`.

SCAO, T. L. – FAN, A. – AKIKI, C. – PAVLICK, E. – ILIĆ, S. – HESSLOW, D. – CASTAGNÉ, R. – LUCCIONI, A. S. – YVON, F. – GALLÉ, M. – OTHERS. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*. 2022.

SCHWENK, H. – WENZEK, G. – EDUNOV, S. – GRAVE, E. – JOULIN, A. – FAN, A. CCMatrix: Mining Billions of High-Quality Parallel Sentences on the Web. In ZONG, C. – XIA, F. – LI, W. – NAVIGLI, R. (Ed.) *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, p. 6490–6500, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.507. Available at: `https://aclanthology.org/2021.acl-long.507`.

SILERO. Silero VAD: pre-trained enterprise-grade Voice Activity Detector (VAD), Number Detector and Language Classifier. `https://github.com/snakers4/silero-vad`, 2021.

SOKY, K. – LI, S. – MIMURA, M. – CHU, C. – KAWAHARA, T. Leveraging Simultaneous Translation for Enhancing Transcription of Low-resource Language via Cross Attention Mechanism. In *Interspeech*, 2022.

SPERBER, M. – PAULIK, M. Speech Translation and the End-to-End Promise: Taking Stock of Where We Are. In JURAFSKY, D. – CHAI, J. – SCHLUTER, N. – TETREAULT, J. (Ed.) *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, p. 7409–7421, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.661. Available at: `https://aclanthology.org/2020.acl-main.661`.

SRIDHAR, V. – CHEN, J. – BANGALORE, S. Corpus analysis of simultaneous interpretation data for improving real time speech translation. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*. 01 2013, p. 3468–3472.

SRIVASTAVA, A. – OTHERS. Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*. 2023. ISSN 2835-8856. Available at: `https://openreview.net/forum?id=uyTL5Bvosj`.

STEWART, C. – VOGLER, N. – HU, J. – BOYD-GRABER, J. – NEUBIG, G. Automatic Estimation of Simultaneous Interpreter Performance. In GUREVYCH, I. – MIYAO, Y. (Ed.) *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, p. 662–666, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-2105. Available at: `https://aclanthology.org/P18-2105`.

SZYMAŃSKI, P. – ŻELASKO, P. – MORZY, M. – SZYMCZAK, A. – ŻYŁA-HOPPE, M. – BANASZCZAK, J. – AUGUSTYNIAK, L. – MIZGAJSKI, J. – CARMIEL, Y. WER we are and WER we think we are. In COHN, T. – HE, Y. – LIU, Y. (Ed.) *Findings of the Association for Computational Linguistics: EMNLP 2020*, p. 3290–3295, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.295. Available at: https://aclanthology.org/2020.findings-emnlp.295.

TAMURA, A. – WATANABE, T. – SUMITA, E. Recurrent Neural Networks for Word Alignment Model. In TOUTANOVA, K. – WU, H. (Ed.) *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, p. 1470–1480, Baltimore, Maryland, June 2014. Association for Computational Linguistics. doi: 10.3115/v1/P14-1138. Available at: https://aclanthology.org/P14-1138.

TEDESCHI, S. – BOS, J. – DECLERCK, T. – HAJIČ, J. – HERSHCOVICH, D. – HOVY, E. – KOLLER, A. – KREK, S. – SCHOCKAERT, S. – SENNRICH, R. – SHUTOVA, E. – NAVIGLI, R. What's the Meaning of Superhuman Performance in Today's NLU? In ROGERS, A. – BOYD-GRABER, J. – OKAZAKI, N. (Ed.) *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, p. 12471–12491, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.697. Available at: https://aclanthology.org/2023.acl-long.697.

TEMNIKOVA, I. – ABDELALI, A. – HEDAYA, S. – VOGEL, S. – AL DAHER, A. Interpreting Strategies Annotation in the WAW Corpus. In TEMNIKOVA, I. – ORASAN, C. – PASTOR, G. C. – VOGEL, S. (Ed.) *Proceedings of the Workshop Human-Informed Translation and Interpreting Technology*, p. 36–43, Varna, Bulgaria, September 2017. Association for Computational Linguistics, Shoumen, Bulgaria. doi: 10.26615/978-954-452-042-7_005. Available at: https://doi.org/10.26615/978-954-452-042-7_005.

TIEDEMANN, J. Word Alignment Step by Step. In NORDGÅRD, T. (Ed.) *Proceedings of the 12th Nordic Conference of Computational Linguistics (NODALIDA 1999)*, p. 216–227, Trondheim, Norway, December 2000. Department of Linguistics, Norwegian University of Science and Technology, Norway. Available at: https://aclanthology.org/W99-1022.

TIEDEMANN, J. – NYGAARD, L. The OPUS Corpus - Parallel and Free: http://logos.uio.no/opus. In LINO, M. T. – XAVIER, M. F. – FERREIRA, F. – COSTA, R. – SILVA, R. (Ed.) *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal, May 2004. European Language Resources Association (ELRA). Available at: http://www.lrec-conf.org/proceedings/lrec2004/pdf/320.pdf.

Tiedemann, J. – Thottingal, S. OPUS-MT – Building open translation services for the World. In Martins, A. – Moniz, H. – Fumega, S. – Martins, B. – Batista, F. – Coheur, L. – Parra, C. – Trancoso, I. – Turchi, M. – Bisazza, A. – Moorkens, J. – Guerberof, A. – Nurminen, M. – Marg, L. – Forcada, M. L. (Ed.) *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, p. 479–480, Lisboa, Portugal, November 2020. European Association for Machine Translation. Available at: `https://aclanthology.org/2020.eamt-1.61`.

Tsiamas, I. – Gállego, G. I. – Fonollosa, J. A. R. – Costa-jussà, M. R. SHAS: Approaching optimal Segmentation for End-to-End Speech Translation. In *Proc. Interspeech 2022*, p. 106–110, 2022. doi: 10.21437/Interspeech.2022-59.

Valk, J. – Alumäe, T. VoxLingua107: a Dataset for Spoken Language Recognition. In *Proc. IEEE SLT Workshop*, 2021.

Varga, D. – Németh, L. – Halácsy, P. – Kornai, A. – Trón, V. – Nagy, V. Parallel corpora for medium density languages. In *Proceedings of the RANLP 2005*, p. 590–596, 2005.

Vaswani, A. – Shazeer, N. – Parmar, N. – Uszkoreit, J. – Jones, L. – Gomez, A. N. – Kaiser, u. – Polosukhin, I. Attention is All You Need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, p. 6000–6010, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.

Veysov, A. – Voronin, D. One Voice Detector to Rule Them All. *The Gradient.* 2022.

Wang, C. – Pino, J. – Wu, A. – Gu, J. CoVoST: A Diverse Multilingual Speech-To-Text Translation Corpus. In Calzolari, N. – Béchet, F. – Blache, P. – Choukri, K. – Cieri, C. – Declerck, T. – Goggi, S. – Isahara, H. – Maegaard, B. – Mariani, J. – Mazo, H. – Moreno, A. – Odijk, J. – Piperidis, S. (Ed.) *Proceedings of the Twelfth Language Resources and Evaluation Conference*, p. 4197–4203, Marseille, France, May 2020. European Language Resources Association. Available at: `https://aclanthology.org/2020.lrec-1.517`. ISBN 979-10-95546-34-4.

Wang, C. – Riviere, M. – Lee, A. – Wu, A. – Talnikar, C. – Haziza, D. – Williamson, M. – Pino, J. – Dupoux, E. VoxPopuli: A Large-Scale Multilingual Speech Corpus for Representation Learning, Semi-Supervised Learning and Interpretation. In Zong, C. – Xia, F. – Li, W. – Navigli, R. (Ed.) *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, p. 993–1003, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.80. Available at: `https://aclanthology.org/2021.acl-long.80`.

WANG, M. – GUO, J. – LI, Y. – QIAO, X. – WANG, Y. – LI, Z. – SU, C. – CHEN, Y. – ZHANG, M. – TAO, S. – YANG, H. – QIN, Y. The HW-TSC's Simultaneous Speech Translation System for IWSLT 2022 Evaluation. In SALESKY, E. – FEDERICO, M. – COSTA-JUSSÀ, M. (Ed.) *Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT 2022)*, p. 247–254, Dublin, Ireland (in-person and online), May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.iwslt-1.21. Available at: `https://aclanthology.org/2022.iwslt-1.21`.

WICKS, R. – POST, M. A unified approach to sentence segmentation of punctuated text in many languages. In ZONG, C. – XIA, F. – LI, W. – NAVIGLI, R. (Ed.) *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, p. 3995–4007, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.309. Available at: `https://aclanthology.org/2021.acl-long.309`.

WOLF, T. et al. Transformers: State-of-the-Art Natural Language Processing. In LIU, Q. – SCHLANGEN, D. (Ed.) *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, p. 38–45, Online, October 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-demos.6. Available at: `https://aclanthology.org/2020.emnlp-demos.6`.

XU, W. – YIN, Y. – MA, S. – ZHANG, D. – HUANG, H. Improving Multilingual Neural Machine Translation with Auxiliary Source Languages. In MOENS, M.-F. – HUANG, X. – SPECIA, L. – YIH, S. W.-t. (Ed.) *Findings of the Association for Computational Linguistics: EMNLP 2021*, p. 3029–3041, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-emnlp.260. Available at: `https://aclanthology.org/2021.findings-emnlp.260`.

XUE, H. – FENG, Y. – GU, S. – CHEN, W. Robust Neural Machine Translation with ASR Errors. In WU, H. – CHERRY, C. – HUANG, L. – HE, Z. – LIBERMAN, M. – CROSS, J. – LIU, Y. (Ed.) *Proceedings of the First Workshop on Automatic Simultaneous Translation*, p. 15–23, Seattle, Washington, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.autosimtrans-1.3. Available at: `https://aclanthology.org/2020.autosimtrans-1.3`.

YUDES, C. – MACIZO, P. – BAJO, T. The Influence of Expertise in Simultaneous Interpreting on Non-Verbal Executive Processes. *Frontiers in Psychology*. 2011, 2. ISSN 1664-1078. doi: 10.3389/fpsyg.2011.00309. Available at: `https://www.frontiersin.org/articles/10.3389/fpsyg.2011.00309`.

ZERVA, C. – BLAIN, F. – REI, R. – LERTVITTAYAKUMJORN, P. – SOUZA, J. G. – EGER, S. – KANOJIA, D. – ALVES, D. – ORĂSAN, C. – FOMICHEVA, M. – MARTINS, A. F. T. – SPECIA, L. Findings of the WMT 2022 Shared Task on Quality Estimation. In KOEHN, P. et al. (Ed.) *Proceedings of the Seventh Conference on Machine Translation (WMT)*, p. 69–99, Abu Dhabi, United Arab Emirates (Hybrid), December 2022. Association for Computational Linguistics. Available at: `https://aclanthology.org/2022.wmt-1.3`.

ZHANG, B. – WILLIAMS, P. – TITOV, I. – SENNRICH, R. Improving Massively Multilingual Neural Machine Translation and Zero-Shot Translation. In JURAFSKY, D. – CHAI, J. – SCHLUTER, N. – TETREAULT, J. (Ed.) *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, p. 1628–1639, Online, July 2020a. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.148. Available at: `https://aclanthology.org/2020.acl-main.148`.

ZHANG, T. – KISHORE, V. – WU, F. – WEINBERGER, K. Q. – ARTZI, Y. BERTScore: Evaluating Text Generation with BERT. In *International Conference on Learning Representations*, 2020b. Available at: `https://openreview.net/forum?id=SkeHuCVFDr`.

ZHAO, J. – ARTHUR, P. – HAFFARI, G. – COHN, T. – SHAREGHI, E. It Is Not As Good As You Think! Evaluating Simultaneous Machine Translation on Interpretation Data. In MOENS, M.-F. – HUANG, X. – SPECIA, L. – YIH, S. W.-t. (Ed.) *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, p. 6707–6715, Online and Punta Cana, Dominican Republic, November 2021a. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.537. Available at: `https://aclanthology.org/2021.emnlp-main.537`.

ZHAO, M. – WU, H. – NIU, D. – WANG, Z. – WANG, X. Verdi: Quality Estimation and Error Detection for Bilingual Corpora. In *Proceedings of the Web Conference 2021*, WWW '21, p. 3023–3031, New York, NY, USA, 2021b. Association for Computing Machinery. doi: 10.1145/3442381.3449931. Available at: `https://doi.org/10.1145/3442381.3449931`. ISBN 9781450383127.

ZIEMSKI, M. – JUNCZYS-DOWMUNT, M. – POULIQUEN, B. The United Nations Parallel Corpus v1.0. In *International Conference on Language Resources and Evaluation*, 2016.

ZOPH, B. – KNIGHT, K. Multi-Source Neural Translation. In KNIGHT, K. – NENKOVA, A. – RAMBOW, O. (Ed.) *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, p. 30–34, San Diego, California, June 2016. Association for Computational Linguistics. doi: 10.18653/v1/N16-1004. Available at: `https://aclanthology.org/N16-1004`.

List of Acronyms

ASR automatic speech recognition

CAI computer assisted interpreting

CAT computer assisted translation

CR Continuous Rating

EU European Union

EUROSAI European Organisation of Supreme Audit Institutions

MT machine translation

NE Normalized Erasure

NLP natural language processing

NMT neural machine translation

PBMT phrase-based machine translation

QE quality estimation

SI simultaneous interpreting

SMT statistical machine translation

SST simultaneous speech translation

ST speech translation

TRL Technology Readiness Level

UN United Nations

VAD voice activity detection

WER word error rate