Multi-Lingual Machine Translation
Ph.D. Thesis Proposal

Dominik Macháček
Charles University, Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics
machacek@ufal.mff.cuni.cz

Abstract

Neural machine translation (NMT) has capability to translate from several parallel inputs in different languages. Current simultaneous speech translation has sometimes issues with quality, especially when the source speech is unclear. We see an opportunity to use multiple parallel speech signals, the original and simultaneous interpreting, as sources for translation, to achieve higher quality. We plan to achieve this goal primarily by experimental exploration of methods in multi-lingual NMT and simultaneous speech translation, and by inspecting features of simultaneous interpreting.

1 Introduction

The neural machine translation (NMT) has capability to handle more source or target languages at once (Johnson et al., 2017; Dabre et al., 2020). The goal of this thesis is to experimentally explore the area of multi-lingual machine translation and propose and evaluate variations of NMT model architectures, training data layout or training methods to achieve gains in quality or efficiency.

The work may primarily focus on one of the following use cases: (i) Multi-target machine translation (MT), where the source is available in one source language, and has to be translated into multiple target languages in a short time. The desired savings are in terms of memory and computation resources (one NMT model serves multiple languages). There has to be as little quality loss as possible. (ii) Multi-source MT, where the source is available in more than one source language. The expected gains are in translation quality, the enriched source should be used for disambiguation. The proposed method should handle the situation where the additional parallel source is unavailable.

We decided to primarily focus only on the use case (ii), multi-source MT.

1.1 Goal: Simultaneous Speech Multi-Source

There are events with multi-lingual audience that use simultaneous interpreting by human experts. However, the set of interpreting target languages is often smaller than what would the audience need. An existing technology of automatic simultaneous speech translation (Müller et al., 2016; Niehues et al., 2018) may increase the set of target languages. Unfortunately, the simultaneous speech translation from one source is often imperfect, due to speaker variance, background noise, non-native accent, etc. A machine could be processing multiple parallel speech signals as source for translation, e.g. the main speaker, and the interpreters. We hypothesize that multi-sourcing may bring benefit to translation quality.

The primary goal of our dissertation is to investigate and propose methods for automatic simultaneous multi-source speech translation. We plan to investigate combination of currently existing methods for multi-source text translation (Section 2.2) with methods for simultaneous speech translation (Section 3).

Namely, we plan to use supervised training on multi-parallel or multiple pairs of bi-parallel data. We plan to use an evaluation corpus of interpreting from European Parliament ESIC (Macháček et al., 2021) that we created. We will primarily experiment with English and German as source languages and Czech as a the target. For training, we plan to use existing corpora. See Section 4.1.

The other basic components for our research are automatic speech recognition (Section 4.2), MT framework (Section 4.3), alignments (Section 4.4), and simultaneous interpreting (Section 5).

In Section 6, we describe one planned experiment with training multi-source without explicit word alignment. We summarize our strategy in Section 7 and conclude in Section 8.

2 Overview of Multi-Lingual MT

In this section, we briefly overview the use of multiple languages in MT. However, we refer to Dabre et al. (2020) for a comprehensive survey of multilingual MT. We extend it in Kocmi et al. (2021), in preparation.

More than two languages in MT are motivated by efficacy (one multi-way model instead of many bilingual ones), and by quality: zero-shot, pivoting and transfer learning especially for low-resourced languages, and parallel multi-source.

2.1 Multi-Way NMT

Recent multi-way multi-lingual NMT uses massive joint models (Aharoni et al., 2019; Arivazhagan et al., 2019a), the same architecture as for the bilingual model (contrary to models with multiple encoders and decoders, one for each source and target language; Firat et al., 2016a). The massive multi-lingual NMT models are capable to translate 100 language pairs at once. When averaging the quality over languages, there are quality gains, however, there may be loss for the top-resourced languages in comparison to the bilingual baseline, due to negative task inference or capacity bottleneck. There may be also wrong target language in zero-shot directions, especially while trianing on an unbalanced English-centric dataset for 100 languages (OPUS-100). Zhang et al. (2020) improve the quality of zero-shot directions by backtranslating randomly selected training data and language pair during training.

The challenge of negative inference can be addressed by controlling parameter sharing (Zhang et al., 2021; Sachan and Neubig, 2018), or by clustering related languages (Tan et al., 2019).

2.2 Parallel Text Multi-Source

In some international organizations, there may be a need to translate text into many target languages in a short time. If the text is already translated into multiple languages and revised by humans, all the revised parallel language variants may be used to increase the translation quality into another target language. A machine translation system that we denote *parallel text multi-source* translates multiple parallel sentences in various source languages.

Zoph and Knight (2016) propose a double-source model model with two encoders and one decoder. They train the model on 2.3M tri-parallel sentences in English, German and French. They report increased quality over baseline trained with the same amount of bilingual sentences. However, we assume that their baseline is weak, and that scaling to larger model and larger data is questionable.

Firat et al. (2016b) propose another double-source model with 2 encoders, one for Spanish, the other for French. The target is English. They train it with two high-resource bi-parallel corpora: 35M sentences for Spanish-English and 66M for French-English. They evaluate several methods for merging the doubled source information in one NMT model. In early averaging, they average the context vectors from both encoders, and feed them into attention and decoder. In late averaging, they average the decoder outputs, before softmax layer. It is the same way as ensembling isolated models. The last and best performing option is ensembling both the early and late averaging.

Dabre et al. (2017) experiment with universal single-encoder model for all source languages. It is the same as single-source NMT, except that the source sequence is a concatenation of sentences in multiple parallel languages. Dabre et al. (2017) compare it with multi-encoder model, one encoder for each language, and with late averaging as in Firat et al. (2016b). They test it on three corpora: United Nations (Es, En, Fr, Ru, Ar), IWSLT (Fr, De, Ar, Cs, En), and with ILCI corpus (4 Indian languages and English). Their results show that multi-sourcing outperforms single source. Moreover, they show that the joint model outperforms late averaging with close source languages, e.g. with French and German into English. With three, less related source languages, e.g. French, German and Arabic, multi-encoder model has the best performance.

2.2.1 Discussion

The three works of Zoph and Knight (2016), Firat et al. (2016b), and Dabre et al. (2017) are the only publications we were able to find that report beneficial usage of multiple parallel sources in text translation. However, these works are 4 or 5 years old, and they use limited training data. We hypothesize that a recent baseline single-source NMT model trained on large bilingual data could outperform the multi-source. There may be too few words and sentences that require two language sources for translation, and not only one, e.g. the closer, or more informative language. Simple heuristics, e.g. detecting the the optimal source from multiple variants, may be very effective, and may not be an adequate result of dissertation.

3 Simultaneous Speech Translation

Simultaneous speech translation is a technology that aims to assist humans with understanding speech in a foreign language. It is primarily for users who need assistance e.g. due to their zero or limited knowledge of the language of the speech, or the speaker's non-native accent, or the specific in-domain vocabulary.

3.1 Translation vs Interpreting

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.

Interpreting involves more than just translating words. The interpreters provide also the intercultural 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 understand. Furthermore, the interpreters handle inappropriate words and offenses in a suitable way, they comment the actions on the stage and provide organizational comments, when necessary. They use not only speech on their input, but complete audio-visual information (metalinguistics, who is addressed by a gesture, etc.), and meta-information, such as current time, location, the event schedule, slides and other relevant documents, etc.

We use the term *speech translation (ST)* for translating speech to text or speech, without the functions that the human interpreters provide in addition to translation. Speech translation is a sub-task of interpreting.

There exists a related term *spoken language translation (SLT)*. Some authors use it as synonymum to ST, however, in our work we distinguish 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 that were prepared for spoken production, or transcribed and normalized speech.

3.2 Long-form Monologue Speech

Our research of simultaneous speech translation primarily focuses on long-form monologue speech, because it would be potentially useful for many users, and it is more challenging due to segmentation to translation units. The alternatives to long-form monologues are conversations and short utterances. They usually contain enough clues for sufficient segmentation to meaning coherent segments (e.g. turns in conversations may correspond to sentences) than long monologues, e.g. a lecture, TED talk, or a speech at parliamentary plenary session.

A speech might be read or spontaneous. It might be very smooth and fluent, or it might contain disfluencies (false starts, repetitions, hesitations, filler 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. The disfluency features are often undesirable for the target audience, especially in translation.

There are tools for speech normalization (inserting punctuation and casing) and speech reconstruction. The speech reconstruction involves detecting and removing disfluencies (Češka, 2009; Chen et al., 2020).

3.3 Simultaneous vs Offline

Speech translation can be used in several modes by latency, each requiring different strategies. In *offline* mode, there are no restrictions on the processing time and efficacy (refer to the offline ST task at IWSLT 2020: Ansari et al., 2020). In *simultaneous translation* (known also as *online* or *low-latency*), there has to be small additive delay, so that the outputs are delivered simultaneously with the source, and the users can interact with the original speaker in real-time (Niehues et al., 2018). In simultaneous mode, the efficacy of computational resources is important. There may be need for translation into several target languages at once, and it might be unfeasible to have a dedicated hardware component for each target.

A hybrid between offline and simultaneous mode is *incremental MT*[1]. It is an algorithm that handles

[1] Many authors use the term *simultaneous translation* when they mean *incremental*. In our opinion, it is necessary to distinguish them.

gradually growing input sequence, processing one input token at a time (Ma et al., 2019a; Zheng et al., 2020; Arivazhagan et al., 2019c; Zheng et al., 2019). The latency of the algorithm is expressed by number of tokens behind the source (Ma et al., 2020a), not by time. Processing time is not important, so we alternatively denote it as *simultaneous, but offline*. Varying speech pace is not assumed, and therefore the implementation is unusable in low-latency mode. However, implementation for a practical low-latency application is possible.

We primarily focus our research to simultaneous mode, and test it in end-to-end setup, from audio to translation. However, we may validate our methods on the other, less complex modes first, if it will be reasoned by simplicity and more universal reproducibility without specifying hardware.

3.4 Cascaded vs End-to-End ST

The automatic speech translation can be cascaded, or end-to-end. Sperber and Paulik (2020) overview and compare these approaches, and describe also their hybrids.

The cascade is a pipeline of individual systems for processing intermediate tasks, e.g.: (i) automatic speech recognition (ASR), (ii) normalization, which may include the speech reconstruction, punctuation prediction, and truecasing, and (iii) machine translation (MT). The advantage of cascaded approach is possibility of distributed development. We may rely on ASR and normalization tools provided by other researchers, and use them as a black-box. We primarily focus our research to improvements of the MT component in cascaded ST, because we have an access to ELITR ST pipeline (Bojar et al., 2021; Franceschini et al., 2020), with ASR and speech normalization for English, German and Czech (and other languages). We also have experience with operating it from IWSLT 2020 (Macháček et al., 2020).

The disadvantage of cascaded ST is error propagation between the sub-systems. In the alternative end-to-end approach, the ASR and MT is 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.

3.5 Re-Translating vs Streaming

There are two main approaches for simultaneous translation. The *re-translating* approach allows output revisions, as the system receives more context (Niehues et al., 2018; Arivazhagan et al., 2019b; Weller et al., 2021)[2], while the *streaming* (Iranzo-Sánchez et al., 2021; Ma et al., 2019a,b) does not. The re-translating approach offers maximal quality that may be comparable to offline ST with enough processing time and context, but it sacrifices the stability. If the stability is low, then the output in form of subtitles may be unreadable. Streaming systems offer maximal stability by producing stable stream of translations. However, they face quality and latency trade-off.

Re-translating systems are often cascaded, as in ELITR. The ASR component tends to make revisions only within fixed-sized processing window at the end of the output stream. The early hypotheses at the end of output may flicker, but they stabilize with more content. With enough space for presenting the subtitles, the user may decide whether to read the unstable parts immediately, or wait for the stabilized parts later. With limited space for subtitling, e.g. with only two lines below the slides presented on the main screen, along the space for other languages, the subtitling brings additional delay while waiting for the finalized hypotheses (Macháček and Bojar, 2020).

With re-translating approach, it is easy to integrate any MT system that is designed for text translation, into the cascaded pipeline (see Section 3.4). Therefore, we first plan to experiment with re-translating ST, but we also consider streaming ST in our research.

3.6 Speech vs Text as Output Modality

Automatic speech translation can be delivered to human users either as text, or as audio with speech. Both options have their advantages and challenges.

In our research, we will primarily focus on ST with an unspecified output modality, whenever the targeted objective (e.g. MT quality, or translation latency) may be measured without delivering the translations to the real users at live or simulated session.

Secondarily, we will focus on simultaneous

[2] Weller et al. (2021) use terminology that is inconsistent to ours. They propose re-translating simultaneous end-to-end ST that provides transcripts and translations at the same time, but describe it by other words. As their *streaming*, they mean the simultaneous mode.

speech translation for live subtitling. We have experience with re-translating ST subtitling in limited space (Macháček and Bojar, 2020), and ELITR system is designed for text output. We have also collaborated on a pilot study of human comprehension with subtitling, with a limited number of human testers and documents (Javorský et al., 2021).

3.7 Evaluation

Evaluation of simultaneous ST is a complex multi-faceted problem. The basic dimensions are translation quality, stability, and latency.

Universal comparison across approaches, architectures, and systems is complicated by their diversity. Real-time end-to-end simulation on a test set is time-consuming. Software and hardware dependencies bring another layer of complexity. A practical, but not universal way, is evaluation that takes into account the specifics or shared characteristics of the evaluated systems. For example, if two cascaded ST systems use the same ASR, but different MT component, it is advisable to evaluate only the MTs.

The evaluation method should also take into account the primary purpose of the evaluation. For example, while selecting a system for subtitling in a limited space, it is advisable to express the stability rate in number of erased characters, and not in tokens (Macháček and Bojar, 2020).

3.7.1 Automatic Quality

The quality of ST may be evaluated by an automatic metric that compares the candidate translation to reference, similarly as in standard text-to-text MT. However, the MT metrics are usually based on the assumption that source and candidate target sentences are aligned one-to-one. In long-form monologue ST, the sentence segmentation is a complex problem. The candidate does not have to preserve the sentence segmentation of the source. Therefore, it is advisable to use a metric that does not rely on source-target sentence alignment, e.g. a variant of BLEU where each document in the test set is treated as one sentence, instead of set of individual sentences. The second, less advisable option, is to estimate the sentence alignment to reference prior to evaluation, e.g. by a tool mwerSegmenter (Matusov et al., 2005). However, the alignment quality may affect the score.

In the annual IWSLT evaluation campaign (Ansari et al., 2020), human evaluation is not used at all. The campaign relies only on the automatic metrics. IWSLT 2021 proposes metrics implementations SimulEval (Ma et al., 2020a) for incremental ST. SLTev SLTev (Ansari et al., 2021) is applicable for simultaneous ST.

3.7.2 Human Evaluation

Human evaluation of simultaneous ST is a complex research problem. In offline ST, the only measurable feature is the quality, contrary to simultaneous ST that involves also stability, latency, and unrepeatability. Therefore, the methods for human evaluation of document-level MT (Castilho, 2020) are applicable to offline ST. The methods consist of blind presentation of the candidates, and e.g. direct assessment.

To best of our knowledge, the only published human evaluation of ST was performed in TC-STAR project (Hamon et al., 2009; Mostefa et al., 2006). It focused on speech-to-speech translation. It was evaluated in the offline mode, so that the judges were able to access the source and candidate repeatedly.

Human evaluation of simultaneous ST in a simulated online mode, with only one access to the candidate, is insufficiently explored research problem. There is a challenge to reduce the effects of varying memory competence between the human evaluators. The initial study by Javorský et al. (2021) showed results with a limited number of human evaluators and documents. The results were mostly insignificant. However, it was shown that the individual competence has the largest impact on comprehension. The online presentation and single access has the second largest impact, followed by presentation options.

3.7.3 Our Plans on Evaluation

We plan to use primarily the evaluation methods that may not be universal, but will be transparent, reasonable, and suitable for our specific purposes. We consider using SimulEval or SLTev, or our own implementation as in the work on ESIC (Macháček et al., 2021). We do not plan to use large scale human evaluation, but we may consider limited human evaluation in the offline mode, similarly to the one in Macháček et al. (2021).

3.8 Multi-Source ST

To the best of our knowledge, Wang et al. (2020a) is the only work where any authors publish results of parallel multi-source speech translation. They publish multi-parallel speech-to-text corpus CoVoST, from 11 languages into English. It contains isolated sentences that volunteers read and recorded for a massively multi-lingual Common Voice corpus (Ardila et al., 2020). The sentences are optimally segmented (one sentence per recording) and aligned, even in the test set. Therefore, the test set may not be applicable to long-form monologue speech.

The authors report improvement of double-source model over 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 of the benefit. As Kocmi et al. (2021) refer also on other works, there may be benefits not because of the multi-lingual source, but because the longer input that makes the encoder wider. Despite of these questions, their work may serve as inspiration for our research. We may also use their published implementation in fairseq ST (Wang et al., 2020b).

4 Basics of Planned Research

In this section, we describe the basics and starting points for our research. However, we will continuously follow new advances in research, and consider other directions, if reasonable.

4.1 Data

We need data for training and evaluation of simultaneous multi-source ST.

4.1.1 ESIC: Europarl Simultaneous Interpreting Corpus

For the evaluation of simultaneous multi-source ST, we need multi-parallel speech corpus in at least three languages. Two of them as sources, and one as target. It has to be authentic monologue long-form speech, and not clean single sentences, for evaluation of real events. Therefore, we can not use evaluation subset of CoVoST (Wang et al., 2020a).

A suitable resource of such authentic data are multi-parallel simultaneously interpreted events, for example the plenary sessions of the European Parliament. Since there did not exist any multi-parallel corpus of interpreting, we created it on our own. Our work results in ESIC: Europarl Simultaneous Interpreting Corpus (Macháček et al., 2021). It is a 10-hour evaluation corpus of 370 speeches given at European Parliament Plenary Sessions in the period 2008 to 2011. The speeches are given originally in English. The corpus contains also simultaneous interpreting into Czech and German, with manual transcripts and word-level timestamps, revised parallel text translations, and metadata about the speakers, time, date and location, and the topic. The speakers are mostly members of European Parliament, both native and non-native speakers of English. The metadata contain the information, whether they read, or speak spontaneously.

ESIC is split into evaluation and validation subsets. Each subset has around 5 hours, or 2 000 sentences. It is comparable by size to standard MT evaluation sets at WMT and IWSLT.

4.1.2 ESIC Extension for Training?

European Parliament is an extensive resource of multi-lingual data. We decided to first collect and process data from the period 2008-11 because at that times, the voices of interpreters were published together with text translations into all EU official languages. The translations may be used as targets for supervised training. It is possible to collect much more speeches from the period 2008-2011 than the 10 hours that we included to the first release of ESIC. Manual transcripts are costly, however, it is possible to process them at least by ASR. The ASR may be improved specifically for ESIC domain by monolingual texts and ESIC validation set. ESIC validation set may be used also for semi-supervision of automatic segmentation to single speeches. Moreover, there is a large amount of data from the new period. Although they are not equipped with translations, the target side may be synthesized by MT.

However, the dataset preparation is very laborious task. There may be other usable training corpora. Therefore, we consider extending ESIC with the training subset only for later.

4.1.3 MT Training Data

We primarily focus only on training the MT component of the cascaded MT, instead of end-to-end ST. For training, we can use either tri-parallel English-German-Czech dataset (because our evaluation corpus ESIC is for these languages), or two bi-parallel corpora for MT. For English-Czech and English-German, we may use the data from WMT shared task: e.g. CzEng 2.0 (Kocmi et al., 2020) that contains also back-translated news. For German-Czech, we may use Europarl (Koehn, 2005) and Open Subtitles (Lison and Tiedemann, 2016). We may also back-translate any data to make them tri-parallel (Choi et al., 2018), or use a model architecture that enables multi-source training on bi-parallel data as Firat et al. (2016b). The architecture options for tri-parallel vs two bi-parallel corpora were described in Section 2.2.

4.1.4 Speech-to-Text Training Data

If we focused on end-to-end ST, or if we would like to finetune the MT part of cascaded ST to accommodate it to ASR errors, we would need speech-to-text training data. Such data are e.g. referred by IWSLT: Europarl-ST (Iranzo-Sánchez et al., 2020) and CoVoST (Wang et al., 2020d) for English-German, and MuST-C (Di Gangi et al., 2019) also for English-Czech. However, all ST corpora except CoVoST are not multi-parallel. They are also often limited in size.

4.2 Automatic Speech Recognition

We have an access to low-latency re-translating ASR systems from ELITR project for English, German, Czech, and other languages. We may use them as the ASR component for our cascaded multi-source ST.

English low-latency ASR (Nguyen et al., 2021) is neural end-to-end model. It achieves super-human quality on conversational speech. In ELITR, it is adapted and tested also on long-form monologue lectures. It is connected with a tool for online speech reconstruction, truecasing and punctuation insertion. It was used also in KIT IWSLT 2020 shared task, both in online and offline mode (Pham et al., 2020).

For German, there is an older hybrid HMM-DNN model trained in Janus Recognition Toolkit which features a single pass decoder. It is described in Cho et al. (2013). There is also a newer end-to-end system, parallel to the English one. We will prefer the system that shows higher quality and sufficient robustness and latency in validation. Both German ASR systems are connected to speech reconstruction and normalization tool.

For Czech, ELITR has a hybrid online ASR in Kaldi (Povey et al., 2011), and an offline end-to-end model (Polák and Bojar, 2021; Polák et al., 2020). There is no speech reconstruction tool for Czech. Speech normalization tool (punctuation prediction and truecasing) for Czech is there, but it is an early version with relatively low quality. However, we may not need the Czech ASR at all if we will focus on Czech as the target language. More critical are English and German for us.

4.2.1 Mock ASR for Evaluation

Based on our initial experience, the English and German ASRs achieve very low quality on the ESIC domain. The ASR quality is critical in cascaded ST. We hypothesize that we may not observe influence of multi-sourcing while evaluating the MT component with very low quality ASR inputs, or with gold transcripts. It is possible that multi-sourcing is beneficial only with specific quality of the ASR sources between zero and optimum.

Therefore, we may evaluate it with “mock ASRs”. We synthesize them either by inserting simple random errors to gold transcripts, or inserting random correct words into realistic low-quality ASR transcript, so that the result has the WER (word error rate) that we choose. We then use it for simulating ASRs with similar quality. We inspect multi-sourcing with various ASR quality levels. The results may be used in future, when the ASRs achieve higher quality.

4.3 MT Framework

For training multi-source ST, we see an opportunity to use any standard NMT framework, e.g. Marian NMT, or any specialized framework for ST.

4.3.1 Marian NMT

We have experience with Marian NMT framework from WMT 2019 news translation task (Popel et al., 2019). Marian NMT (Junczys-Dowmunt et al., 2018) is a framework for fast text-to-text NMT training and inference. It has implementation of Transformer (Vaswani et al., 2017). We may use it for training the MT component of cascaded ST, especially a re-translating simultaneous model. Marian enables also multi-encoder model that may be suitable. We plan to start our experiments with Marian.

4.3.2 Fairseq

Fairseq (Ott et al., 2019) is a fast, extensible toolkit for sequence modeling. Fairseq S2T (Wang et al., 2020c) is an extention for end-to-end speech-to-text translation. Another extension (Ma et al., 2020b) is for simultaneous ST. It contains an implementation of recent streaming models.

Fairseq is connected with a community of developers and researchers. There are published source codes and training recipes. We consider training of simultaneous multi-source ST in this framework.

However, it may be challenging to adapt the source code for multi-sourcing.

4.3.3 Other ST Frameworks

We are aware of other frameworks that could be used for ST training: ESPnet-ST (Inaguma et al., 2020), RETURNN (Zeyer et al., 2018), Lingvo (Shen et al., 2019), and SLT.KIT (Zenkel et al., 2018). We do not have any experience with any of these systems, except one. We had serious issues with installing SLT.KIT.

4.4 Alignments

We plan to propose a double-source system that will translate an original speaker, and a simultaneous interpreter. We assume that we will face the problem of aligning the sources because the interpreters keep lagging behind the source. We attempt to solve it in an unsupervised way, by translating scrolling windows that will be large enough (Section 6). If it will not be possible, we may apply a word alignment tool to detect the non-parallel parts and truncate them. However, it may be challenging because the tools are also designed for limited length, mostly for individual sentences. Another challenge is the speech input. The tools are effective for normalized texts.

Usable word alignment tools are e.g. fast_align (Dyer et al., 2013), or recent neural aligners (Dou and Neubig, 2021).

5 Inspiration by Human Interpreting

We get information about interpreting mainly from the active interpreters and teachers of interpreting (Čeňková, 2008, 1988; Ešnerová, 2019; Olsen, 2020). They refer to theory in translation and interpreting studies, such as to Gile's effort model (Gile, 1995). Another resource is computational linguistics research (He et al., 2016; Stewart et al., 2018; Sridhar et al., 2013; Vogler et al., 2019). A very insightful is also the keynote by Welle (2020), the head of administration of European Parliament, world's biggest employer of interpreters.

In this section, we describe how we could potentially use inspiration by interpreting in our research.

5.1 Comparing ST to Interpreting

Simultaneous interpreting is widely used in international communication at least since 1945 (Čeňková, 2008). Humans are naturally used to it. Simultaneous speech translation is a relatively new technology. Therefore, we see an opportunity to estimate the adoption of simultaneous ST by users by comparison to interpreting.

For example, in Macháček et al. (2021), we compared the latency of ST to interpreters. We conclude that if interpreting latency is 4 seconds, interpreting through a pivot language may have latency 8 seconds. Therefore, we assume that human users may be accustomed to 8 second latency of ST.

5.2 Gold Truth Translation Units

Segmentation to translation units is one of the largest challenges in simultaneous ST. Too short units lead to low quality, and large units to large latency. If we assume that the interpreters choose their translation units optimally, then we can use interpreting data for supervised training of segmentation for ST. The translation units can be analyzed on ESIC. Moreover, ESIC can be extended by manually detected translation units.

5.3 Redundancy Detection

Čeňková (2008) writes about redundancy detection and elision. It is a subtask of interpreting that enables the simultaneity. Redundancy detection could be useful in practical applications such as in automatic summarization, speech reconstruction, text compression and simplification, and also in speech translation. Short translation outputs with less redundancies may be better comprehensible.

We hypothesize that ESIC or other simultaneous interpreting dataset could be potentially useful for supervised learning of redundancy detection.

5.4 Understanding Interpretese

When developing multi-source ST that uses interpreting as one of the sources, it might be beneficial to be aware of the limits and features of interpreting. It might give us insights for more successful ST development. Below, we list selected facts about interpreting that may be relevant for our research.

- Interpreting is not completely reliable. It is the art of possible (Olsen, 2020; Ešnerová, 2019). There may be outages in interpreting, e.g. when the interpreter is exhausted. The interpreters must make pauses and work in pairs.

- Interpreting is impossible, or possible with a reduced quality, when the speech pace is too high, when the interpreter does not receive the source audio or when the speech is unclear, or untranslatable in nature, e.g. mocking accent, poetry, and puns.

For simplicity, we assume that the source speech for our ST will always be translatable.

- If the speaker prepares the speech in advance as a text that is going to be read, then the interpreters prefer if they receive a copy of the text. Written language tends to be more complicated than spontaneous.

  In our opinion, the ST research should ideally primarily focus to non-read, spontaneous speech, because the read speech can be better translated using the text. However, it is difficult to find data for only spontaneous ST.

- A skilled interpreter can determine from listening, whether another interpreter used the text that the speaker was reading, or just the audio (Čeňková, 1988). We, as non-experts, are not able to determine it in ESIC.

- Interpreters use *offline* strategies to prepare for interpreting of an event in advance (Ešnerová, 2019). They keep themselves up to date about the news associated with the source and target language cultures. They study materials about the event. They anticipate the topics and prepare a vocabulary.

  The same backgrounds should be ideally taken into account in expert evaluation of ST. However, it may be complicated to study them for ESIC corpus because it contains facts that were valid in 2008-11.

- Interpreting strategies include keeping optimal delay behind the speaker, and vary it when reasonable. He et al. (2016) find out that English-Japanese interpreters tend to consolidate the word order between the source and target language by passivization. They also prefer other language constructs than the translators.

- It is preferred to interpret into native language of the interpreter over the non-native. For us, it is important to note when using ASR that may have difficulties with a non-native accent.

- Interpreting through a pivot language (so called relay interpreting) is sometimes used, when impossible directly. When using interpreting as gold truth, it should be taken into account. In ESIC, we can not determine for sure whether English-Czech or English-German interpreting was direct, or relay. However, we assume it was direct because these language pairs are frequent.

- There may be traces of the source language in interpreting, in nearly all levels of language: in the speech signal, phonetics, vocabulary, syntax, in the topic, etc.

  We expect the traces make interpretese a special domain for MT.

- Čeňková (2008) describe a principle of economy. The principle recommends the interpreters to use short variant of translation whenever possible, and to elide redundancies. It is recommended to use simple sentence structure over long dependencies that are difficult to remember and complete fluently.

  In Macháček et al. (2021), we found out that interpretese is by 20% shorter than translationese, and that interpretese contains simpler vocabulary.

6 Planned Experiment: Multi-Source with Unsupervised Source Alignment

We propose an experiment with a double-source NMT model as in Figure 1. It will have two sources.

The first source, the original English, is going to be translated word for word. It has two parts that will be separated by a special marker on input (solid vertical line in the figure): context (C words long), and translation window (W words long). The context is necessary because we translate fixed-size super-sentence windows. It is likely that the words close to window start require previous context for determining the correct form.

Second source, simultaneous interpreting into German, is used for enriching the first source. It has three unmarked parts (separated by a dashed vertical line on the figure): part to skip because it is parallel to the content that has been already scrolled away in the first source. Second part is parallel context, and the last is parallel to the part to translate. We assume that the NMT system detects the parts unsupervised.

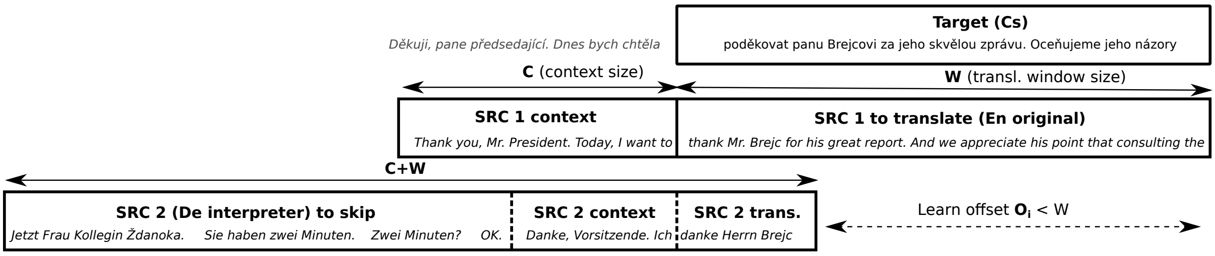
Target (Cs)
poděkovat panu Brejcovi za jeho skvělou zprávu. Oceňujeme jeho názory
Děkuji, pane předsedající. Dnes bych chtěla
C (context size)
W (transl. window size)
SRC 1 context
Thank you, Mr. President. Today, I want to
SRC 1 to translate (En original)
thank Mr. Brejc for his great report. And we appreciate his point that consulting the
C+W
SRC 2 (De interpreter) to skip
Jetzt Frau Kollegin Ždanoka. Sie haben zwei Minuten. Zwei Minuten? OK.
SRC 2 context
Danke, Vorsitzende. Ich
SRC 2 trans.
danke Herrn Brejc
Learn offset $\mathbf{O_i}$ < W

Figure 1: Example training instance for double-source simultaneous ST with unsupervised source alignment.

6.1 Estimating the Length Limits

NMT models have limits on their source and target length. If the latency of interpreting would be too large, it could exceed the limit. Therefore, we estimate whether the maximum latency fits.

Čeňková (2008) states that the majority of interpreters keep lagging time between 2 and 4 seconds. The lagging varies by speech content. Our analysis on ESIC (Macháček et al., 2021) confirms it. Moreover, we found that in 1% of source words, the interpreting latency may be very large, over 22 seconds.

Čeňková (2008) also states that AIIC[3] states maximum speech pace for simultaneous interpreting. It is 200 words per minute (wpm).[4]

Let us assume that the interpreter can be delayed by at most 30 seconds. With the maximum speech pace 200 wpm, the original to interpreting offset can be $W = 100$ words. If we select $C = 50$, then the model has 300 words on sources and 100 words on target. We assume that these sizes are feasible e.g. for Marian NMT.

7 Planned Tasks

We split our work into three tasks. We plan to switch focus between the tasks. We assume that as we will make progress, we may figure out that one task is blocked by another, and that the blocked task brings a clearer specification for unblocking by other task.

7.1 Task 1: Data

A large piece of work on data has been already accomplished by publishing the ESIC corpus (Macháček et al., 2021). We have an evaluation corpus. However, it has to be decided, what should be considered as reference English-Czech translation: interpreting, or translation?

The interpreting may contain outages, but we do not know how much. If it is low in amount, then we may ignore it or make it complete. The translation is, on the other hand, revised and normalized for reading, so it is partially not verbatim translation. We may need to analyze, how much editing was made in the revisions, and either ignore it, or edit it back to verbatim.

We may also evaluate, which source is more preferred by users. The evaluation in Macháček et al. (2021) were not focused on fluency. Fluency is an important aspect that may favor translation. The evaluation focused only on information loss. It showed that interpreting drops more information than translation. However, the reduced information could be redundant, so we do not know whether the reduction is a beneficial compression, or a substantial adequacy error.

The next step are training data. We prepare them according to the plans in Section 4.1, by specifications that will become apparent while working on the other tasks.

7.2 Task 2: Baselines

A baseline to multi-source is single-source. In Macháček et al. (2021), we compared two single-source systems that can be considered as baselines: shortening English-Czech MT and German-Czech MT. They are strong as offline SLT baselines. However, although they are deployed in online ST cascade, we can not consider them as strong online baselines, because they are not finetuned for stability, and are relatively unstable. Baseline stabilization method of re-translating online ST involves finetuning on sentences prefixes (Niehues et al., 2018).

7.3 Task 3: Simultaneous Multi-Source ST

We plan to approach our primary goal by investigating and exploring methods that we described in this

[3] International Association of Conference Interpreters. Its French acronym AIIC is used in all languages.

[4] According to Čeňková (2008), AIIC does not specify, for which language the limit holds. We can neglect this fact in our approximation.

thesis proposal. We start with a suboptimal, but feasible system, and gradually attempt to improve it.

Along the way, we may meet following milestones:

1. Offline multi-source ST
2. Simultaneous multi-source ST
3. Flexible and robust simultaneous multi-source ST.

8 Conclusion

In this dissertation thesis proposal, we presented our plan to primarily focus our research to simultaneous multi-source speech translation. We described the concepts of multi-lingual NMT, simultaneous speech translation, and simultaneous interpreting that we plan to use in our experiments. We also described our strategy regarding experimenting with data, architecture, implementation, evaluation, approaches to simultaneity, and alternative directions we could take.

In this proposal, we partially used results of work that we elaborated or on which we collaborated and that were reviewed and published (Macháček et al., 2021; Macháček and Bojar, 2020; Macháček et al., 2020; Franceschini et al., 2020; Polák et al., 2020; Bojar et al., 2021; Popel et al., 2019), as well us recent unpublished results (Javorský et al., 2021; Kocmi et al., 2021).

References

Roee Aharoni, Melvin Johnson, and Orhan Firat. 2019. Massively multilingual neural machine translation. In *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)*, pages 3874–3884, Minneapolis, Minnesota. Association for Computational Linguistics.

Ebrahim Ansari, Amittai Axelrod, Nguyen Bach, Ondřej Bojar, Roldano Cattoni, Fahim Dalvi, Nadir Durrani, Marcello Federico, Christian Federmann, Jiatao Gu, Fei Huang, Kevin Knight, Xutai Ma, Ajay Nagesh, Matteo Negri, Jan Niehues, Juan Pino, Elizabeth Salesky, Xing Shi, Sebastian Stüker, Marco Turchi, Alexander Waibel, and Changhan Wang. 2020. FINDINGS OF THE IWSLT 2020 EVALUATION CAMPAIGN. In *Proceedings of the 17th International Conference on Spoken Language Translation*, pages 1–34, Online. Association for Computational Linguistics.

Ebrahim Ansari, Ondřej Bojar, Barry Haddow, and Mohammad Mahmoudi. 2021. SLTEV: Comprehensive evaluation of spoken language translation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 71–79, Online. Association for Computational Linguistics.

R. Ardila, M. Branson, K. Davis, M. Henretty, M. Kohler, J. Meyer, R. Morais, L. Saunders, F. M. Tyers, and G. Weber. 2020. Common voice: A massively-multilingual speech corpus. In *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, pages 4211–4215.

Naveen Arivazhagan, Ankur Bapna, Orhan Firat, Dmitry Lepikhin, Melvin Johnson, Maxim Krikun, Mia Xu Chen, Yuan Cao, George F. Foster, Colin Cherry, Wolfgang Macherey, Zhifeng Chen, and Yonghui Wu. 2019a. Massively multilingual neural machine translation in the wild: Findings and challenges. *CoRR*, abs/1907.05019.

Naveen Arivazhagan, Colin Cherry, Te I, Wolfgang Macherey, Pallavi Baljekar, and George Foster. 2019b. Re-translation strategies for long form, simultaneous, spoken language translation.

Naveen Arivazhagan, Colin Cherry, Wolfgang Macherey, Chung-Cheng Chiu, Semih Yavuz, Ruoming Pang, Wei Li, and Colin Raffel. 2019c. Monotonic infinite lookback attention for simultaneous machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1313–1323, Florence, Italy. Association for Computational Linguistics.

Ondřej Bojar, Dominik Macháček, Sangeet Sagar, Otakar Smrž, Jonáš Kratochvíl, Peter Polák, Ebrahim Ansari, Mohammad Mahmoudi, Rishu Kumar, Dario Franceschini, Chiara Canton, Ivan Simonini, Thai-Son Nguyen, Felix Schneider, Sebastian Stüker, Alex Waibel, Barry Haddow, Rico Sennrich, and Philip Williams. 2021. ELITR multilingual live subtitling: Demo and strategy. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 271–277, Online. Association for Computational Linguistics.

Sheila Castilho. 2020. Document-level machine translation evaluation project: Methodology, effort and inter-annotator agreement. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 455–456, Lisboa, Portugal. European Association for Machine Translation.

Qian Chen, Mengzhe Chen, Bo Li, and Wen Wang. 2020. 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*, pages 8069–8073. IEEE.

Eunah Cho, C. Fügen, T. Hermann, K. Kilgour, Mohammed Mediani, C. Mohr, J. Niehues, Kay Rottmann, C. Saam, Sebastian Stüker, and A. Waibel. 2013. A real-world system for simultaneous translation of german lectures. pages 3473–3477.

Gyu-Hyeon Choi, Jong-Hun Shin, and Young-Kil Kim. 2018. Improving a multi-source neural machine translation model with corpus extension for low-resource languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Raj Dabre, Chenhui Chu, and Anoop Kunchukuttan. 2020. A survey of multilingual neural machine translation. *ACM Comput. Surv.*, 53(5).

Raj Dabre, Fabien Cromieres, and Sadao Kurohashi. 2017. Enabling multi-source neural machine translation by concatenating source sentences in multiple languages.

Mattia Antonino Di Gangi, Roldano Cattoni, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2019. 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.

Zi-Yi Dou and Graham Neubig. 2021. Word alignment by fine-tuning embeddings on parallel corpora. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2112–2128, Online. Association for Computational Linguistics.

Chris Dyer, Victor Chahuneau, and Noah A. Smith. 2013. A simple, fast, and effective reparameterization of IBM model 2. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 644–648, Atlanta, Georgia. Association for Computational Linguistics.

Kateřina Ešnerová. 2019. Hledáme dream job: Tlumočnice. Skautský institut, `https://youtu.be/f8z464rTC0Y`.

Orhan Firat, Kyunghyun Cho, and Yoshua Bengio. 2016a. Multi-way, multilingual neural machine translation with a shared attention mechanism. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 866–875, San Diego, California. Association for Computational Linguistics.

Orhan Firat, Baskaran Sankaran, Yaser Al-onaizan, Fatos T. Yarman Vural, and Kyunghyun Cho. 2016b. Zero-resource translation with multi-lingual neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 268–277, Austin, Texas. Association for Computational Linguistics.

Dario Franceschini, Chiara Canton, Ivan Simonini, Armin Schweinfurth, Adelheid Glott, Sebastian Stüker, Thai-Son Nguyen, Felix Schneider, Thanh-Le Ha, Alex Waibel, Barry Haddow, Philip Williams, Rico Sennrich, Ondřej Bojar, Sangeet Sagar, Dominik Macháček, and Otakar Smrž. 2020. Removing European language barriers with innovative machine translation technology. In *Proceedings of the 1st International Workshop on Language Technology Platforms*, pages 44–49, Marseille, France. European Language Resources Association.

Daniel Gile. 1995. *Basic Concepts and Models for Interpreter and Translator Training*. John Benjamins, Amsterdam/Philadelphia.

Olivier Hamon, Christian Fügen, Djamel Mostefa, Victoria Arranz, Muntsin Kolss, Alex Waibel, and Khalid Choukri. 2009. End-to-end evaluation in simultaneous translation. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, pages 345–353, Athens, Greece. Association for Computational Linguistics.

He He, Jordan Boyd-Graber, and Hal Daumé III. 2016. Interpretese vs. translationese: The uniqueness of human strategies in simultaneous interpretation. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 971–976, San Diego, California. Association for Computational Linguistics.

Hirofumi Inaguma, Shun Kiyono, Kevin Duh, Shigeki Karita, Nelson Yalta, Tomoki Hayashi, and Shinji Watanabe. 2020. ESPnet-ST: All-in-one speech translation toolkit. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 302–311, Online. Association for Computational Linguistics.

Javier Iranzo-Sánchez, Joan Albert Silvestre-Cerdà, Javier Jorge, Nahuel Roselló, Adrià Giménez, Albert Sanchis, Jorge Civera, and Alfons Juan. 2020. 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)*, pages 8229–8233.

Javier Iranzo-Sánchez, Javier Jorge, Pau Baquero-Arnal, Joan Albert Silvestre-Cerdà, Adrià Giménez, Jorge Civera, Albert Sanchis, and Alfons Juan. 2021. Streaming cascade-based speech translation leveraged by a direct segmentation model. *Neural Networks*, 142:303–315.

Dávid Javorský, Dominik Macháček, and Ondřej Bojar. 2021. Comprehension of subtitles from re-translating simultaneous speech translation. Unpublished, under review.

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F. T. Martins, and Alexandra Birch. 2018. Marian: Fast neural machine translation in C++. In *Proceedings of ACL 2018, System Demonstrations*, pages 116–121, Melbourne, Australia. Association for Computational Linguistics.

Tom Kocmi, Dominik Macháček, and Ondřej Bojar. 2021. *More than Two Languages in Machine Translation (Unpublished draft)*. Institute of Formal and Applied Linguistics, Prague, Czech Republic.

Tom Kocmi, Martin Popel, and Ondrej Bojar. 2020. Announcing czeng 2.0 parallel corpus with over 2 gigawords. *arXiv preprint arXiv:2007.03006*.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *MT summit*, volume 5, pages 79–86. Citeseer.

Pierre Lison and Jörg Tiedemann. 2016. OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 923–929, Portorož, Slovenia. European Language Resources Association (ELRA).

Mingbo Ma, Liang Huang, Hao Xiong, Renjie Zheng, Kaibo Liu, Baigong Zheng, Chuanqiang Zhang, Zhongjun He, Hairong Liu, Xing Li, Hua Wu, and Haifeng Wang. 2019a. STACL: Simultaneous translation with implicit anticipation and controllable latency using prefix-to-prefix framework. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3025–3036, Florence, Italy. Association for Computational Linguistics.

Mingyu Derek Ma, Kevin Bowden, Jiaqi Wu, Wen Cui, and Marilyn Walker. 2019b. Implicit discourse relation identification for open-domain dialogues. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 666–672, Florence, Italy. Association for Computational Linguistics.

Xutai Ma, Mohammad Javad Dousti, Changhan Wang, Jiatao Gu, and Juan Pino. 2020a. SIMULEVAL: An evaluation toolkit for simultaneous translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 144–150, Online. Association for Computational Linguistics.

Xutai Ma, Juan Pino, and Philipp Koehn. 2020b. SimulMT to SimulST: Adapting simultaneous text translation to end-to-end simultaneous speech translation. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 582–587, Suzhou, China. Association for Computational Linguistics.

Dominik Macháček, Jonáš Kratochvíl, Sangeet Sagar, Matúš Žilinec, Ondřej Bojar, Thai-Son Nguyen, Felix Schneider, Philip Williams, and Yuekun Yao. 2020. ELITR non-native speech translation at IWSLT 2020. In *Proceedings of the 17th International Conference on Spoken Language Translation*, pages 200–208, Online. Association for Computational Linguistics.

Dominik Macháček and Ondřej Bojar. 2020. Presenting simultaneous translation in limited space. In *Proceedings of the 20th Conference Information Technologies - Applications and Theory (ITAT 2020)*, pages 32–37, Košice, Slovakia. Tomáš Horváth.

Dominik Macháček, Matúš Žilinec, and Ondřej Bojar. 2021. Lost in interpreting: Speech translation from source or interpreter? In *Interspeech 2021, 22nd Annual Conference of the International Speech Communication Association*. ISCA.

Evgeny Matusov, Gregor Leusch, Oliver Bender, and Hermann Ney. 2005. Evaluating machine translation output with automatic sentence segmentation. In *International Workshop on Spoken Language Translation*, pages 148–154, Pittsburgh, PA, USA.

Djamel Mostefa, Olivier Hamon, and Khalid Choukri. 2006. Evaluation of automatic speech recognition and speech language translation within TC-STAR:results from the first evaluation campaign. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy. European Language Resources Association (ELRA).

Markus Müller, Thai Son Nguyen, Jan Niehues, Eunah Cho, Bastian Krüger, Thanh-Le Ha, Kevin Kilgour, Matthias Sperber, Mohammed Mediani, Sebastian Stüker, and Alex Waibel. 2016. Lecture translator - speech translation framework for simultaneous lecture translation. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 82–86, San Diego, California. Association for Computational Linguistics.

Thai-Son Nguyen, Sebastian Stueker, and Alex Waibel. 2021. Super-human performance in online low-latency recognition of conversational speech.

Jan Niehues, Ngoc-Quan Pham, Thanh-Le Ha, Matthias Sperber, and Alexander H. Waibel. 2018. Low-latency neural speech translation. In *INTERSPEECH*.

Barry Slaughter Olsen. 2020. Human interpreter training and practice: Insights for simultaneous machine translation research. Invited talk at workshop AutoSimTrans 2020 at ACL.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53, Minneapolis, Minnesota. Association for Computational Linguistics.

Ngoc-Quan Pham, Felix Schneider, Tuan-Nam Nguyen, Thanh-Le Ha, Thai Son Nguyen, Maximilian Awiszus, Sebastian Stüker, and Alexander Waibel. 2020. KIT's IWSLT 2020 SLT translation system. In *Proceedings of the 17th International Conference on Spoken Language Translation*, pages 55–61, Online. Association for Computational Linguistics.

Peter Polák, Sangeet Sagar, Dominik Macháček, and Ondřej Bojar. 2020. CUNI neural ASR with phoneme-level intermediate step for˜Non-Native˜SLT at IWSLT 2020. In *Proceedings of the 17th International Conference on Spoken Language Translation*, pages 191–199, Online. Association for Computational Linguistics.

Peter Polák and Ondřej Bojar. 2021. Coarse-to-fine and cross-lingual ASR transfer. In *Proceedings of the 20th Conference Information Technologies - Applications and Theory (ITAT 2021)*.

Martin Popel, Dominik Macháček, Michal Auersperger, Ondřej Bojar, and Pavel Pecina. 2019. English-Czech systems in WMT19: Document-level transformer. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 342–348, Florence, Italy. Association for Computational Linguistics.

Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Nagendra Goel, Mirko Hannemann, Yanmin Qian, Petr Schwarz, and Georg Stemmer. 2011. The kaldi speech recognition toolkit. In *In IEEE 2011 workshop*.

Devendra Sachan and Graham Neubig. 2018. Parameter sharing methods for multilingual self-attentional translation models. pages 261–271.

Jonathan Shen, Patrick Nguyen, Yonghui Wu, Zhifeng Chen, Mia X. Chen, Ye Jia, Anjuli Kannan, Tara Sainath, Yuan Cao, Chung-Cheng Chiu, Yanzhang He, Jan Chorowski, Smit Hinsu, Stella Laurenzo, James Qin, Orhan Firat, Wolfgang Macherey, Suyog Gupta, Ankur Bapna, Shuyuan Zhang, Ruoming Pang, Ron J. Weiss, Rohit Prabhavalkar, Qiao Liang, Benoit Jacob, Bowen Liang, HyoukJoong Lee, Ciprian Chelba, Sébastien Jean, Bo Li, Melvin Johnson, Rohan Anil, Rajat Tibrewal, Xiaobing Liu, Akiko Eriguchi, Navdeep Jaitly, Naveen Ari, Colin Cherry, Parisa Haghani, Otavio Good, Youlong Cheng, Raziel Alvarez, Isaac Caswell, Wei-Ning Hsu, Zongheng Yang, Kuan-Chieh Wang, Ekaterina Gonina, Katrin Tomanek, Ben Vanik, Zelin Wu, Llion Jones, Mike Schuster, Yanping Huang, Dehao Chen, Kazuki Irie, George Foster, John Richardson, Klaus Macherey, Antoine Bruguier, Heiga Zen, Colin Raffel, Shankar Kumar, Kanishka Rao, David Rybach, Matthew Murray, Vijayaditya Peddinti, Maxim Krikun, Michiel A. U. Bacchiani, Thomas B. Jablin, Rob Suderman, Ian Williams, Benjamin Lee, Deepti Bhatia, Justin Carlson, Semih Yavuz, Yu Zhang, Ian McGraw, Max Galkin, Qi Ge, Golan Pundak, Chad Whipkey, Todd Wang, Uri Alon, Dmitry Lepikhin, Ye Tian, Sara Sabour, William Chan, Shubham Toshniwal, Baohua Liao, Michael Nirschl, and Pat Rondon. 2019. Lingvo: a modular and scalable framework for sequence-to-sequence modeling.

Matthias Sperber and Matthias Paulik. 2020. Speech translation and the end-to-end promise: Taking stock of where we are. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7409–7421, Online. Association for Computational Linguistics.

V.K.R. Sridhar, J. Chen, and S. Bangalore. 2013. Corpus analysis of simultaneous interpretation data for improving real time speech translation. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, pages 3468–3472.

Craig Stewart, Nikolai Vogler, Junjie Hu, Jordan Boyd-Graber, and Graham Neubig. 2018. Automatic estimation of simultaneous interpreter performance. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 662–666, Melbourne, Australia. Association for Computational Linguistics.

Xu Tan, Jiale Chen, Di He, Yingce Xia, Tao Qin, and Tie-Yan Liu. 2019. Multilingual neural machine translation with language clustering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 963–973, Hong Kong, China. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, undefinedukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6000–6010, Red Hook, NY, USA. Curran Associates Inc.

Nikolai Vogler, Craig Stewart, and Graham Neubig. 2019. Lost in Interpretation: Predicting Untranslated Terminology in Simultaneous Interpretation. In *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), pages 109–118, Minneapolis, Minnesota. Association for Computational Linguistics.

Changhan Wang, Juan Pino, Anne Wu, and Jiatao Gu. 2020a. CoVoST: A diverse multilingual speech-to-text translation corpus. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4197–4203, Marseille, France. European Language Resources Association.

Changhan Wang, Yun Tang, Xutai Ma, Anne Wu, Dmytro Okhonko, and Juan Pino. 2020b. fairseq s2t: Fast speech-to-text modeling with fairseq. In *Proceedings of the 2020 Conference of the Asian Chapter of the Association for Computational Linguistics (AACL): System Demonstrations*.

Changhan Wang, Yun Tang, Xutai Ma, Anne Wu, Dmytro Okhonko, and Juan Pino. 2020c. Fairseq S2T: Fast speech-to-text modeling with fairseq. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 33–39, Suzhou, China. Association for Computational Linguistics.

Changhan Wang, Anne Wu, and Juan Pino. 2020d. Covost 2: A massively multilingual speech-to-text translation corpus.

Klaus Welle. 2020. Transformation of language services in the European Parliament. Keynote and QA at IWSLT 2020.

Orion Weller, Matthias Sperber, Christian Gollan, and Joris Kluivers. 2021. Streaming models for joint speech recognition and translation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2533–2539, Online. Association for Computational Linguistics.

Thomas Zenkel, Matthias Sperber, Jan Niehues, Markus Müller, Ngoc-Quan Pham, Sebastian Stüker, and Alex Waibel. 2018. Open Source Toolkit for Speech to Text Translation. *The Prague Bulletin of Mathematical Linguistics*, 111:125–135.

Albert Zeyer, Tamer Alkhouli, and Hermann Ney. 2018. RETURNN as a generic flexible neural toolkit with application to translation and speech recognition. In *Proceedings of ACL 2018, System Demonstrations*, pages 128–133, Melbourne, Australia. Association for Computational Linguistics.

Biao Zhang, Ankur Bapna, Rico Sennrich, and Orhan Firat. 2021. Share or not? learning to schedule language-specific capacity for multilingual translation. In *International Conference on Learning Representations (ICLR 2021)*. Ninth International Conference on Learning Representations 2021, ICLR 2021 ; Conference date: 04-05-2021 Through 07-05-2021.

Biao Zhang, Philip Williams, Ivan Titov, and Rico Sennrich. 2020. Improving massively multilingual neural machine translation and zero-shot translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1628–1639, Online. Association for Computational Linguistics.

Baigong Zheng, Kaibo Liu, Renjie Zheng, Mingbo Ma, Hairong Liu, and Liang Huang. 2020. Simultaneous translation policies: From fixed to adaptive. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2847–2853, Online. Association for Computational Linguistics.

Renjie Zheng, Mingbo Ma, Baigong Zheng, and Liang Huang. 2019. Speculative beam search for simultaneous translation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1395–1402, Hong Kong, China. Association for Computational Linguistics.

Barret Zoph and Kevin Knight. 2016. Multi-source neural translation. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 30–34, San Diego, California. Association for Computational Linguistics.

Ivana Čeňková. 1988. *Teoretické aspekty simultánního tlumočení: na materiálu rusko-českém a česko-ruském*, volume 99. Univerzita karlova.

Ivana Čeňková. 2008. *Úvod do teorie tlumočení*. Česká komora tlumočníků znakového jazyka, o.s., Praha. 2. opravené vydání.

Pavel Češka. 2009. Speech reconstruction - overview of state-of-the-art systems. In *WDS'09 Proceedings of Contributed Papers*, pages 11–15, Praha, Czechia. Matfyzpress, Charles University.