ELITR Multilingual Live Subtitling: Demo and Strategy

Ondřej Bojar and **Dominik Macháček** and **Sangeet Sagar** and **Otakar Smrž** and
Jonáš Kratochvíl and **Peter Polák** and **Ebrahim Ansari** and **Mohammad Mahmoudi** and **Rishu Kumar**
Charles University; `<surname>@ufal.mff.cuni.cz` except `jkratochvil@ufal.mff.cuni.cz`

Dario Franceschini and **Chiara Canton** and **Ivan Simonini**
PerVoice

Thai-Son Nguyen and **Felix Schneider** and **Sebastian Stüker** and **Alex Waibel**
Karlsruhe Institute of Technology; `<firstname>.<lastname>@kit.edu`

Barry Haddow and **Rico Sennrich** and **Philip Williams**
University of Edinburgh

Abstract

This paper presents an automatic speech translation system aimed at live subtitling of conference presentations. We describe the overall architecture and key processing components. More importantly, we explain our strategy for building a complex system for end-users from numerous individual components, each of which has been tested only in laboratory conditions.

The system is a working prototype that is routinely tested in recognizing English, Czech, and German speech and presenting it translated simultaneously into 42 target languages.

1 Introduction

With the tremendous gains observed recently in automatic speech recognition (ASR) and machine translation (MT) quality, including methods of joint learning of both of the tasks, the goal of a practically usable simultaneous spoken language translation (SLT[1]) system is getting closer.

In this paper, we introduce the SLT system developed in the EU project ELITR (European Live Translator[2]) (Bojar et al., 2020) which aims at a distinct setting: real-time speech translation into many target languages.

2 Motivation

In the current globalized world, meetings with participants from a very wide spectrum of nations are common. Many multinational organizations, public or private, regularly run congresses and conferences where attendees do not have any language in common. Interpretation is a must at such meetings and the cost of interpretation services consumes a considerable portion of the budget. The number of provided languages is then kept as low as possible, even in cases when some of the attendees are not sufficiently fluent in any of them.

We primarily focus on the setting of such multinational congresses where one source speech needs to be translated into many target languages. While we are aware of the quality limitations of speech recognition and machine translation, we strongly believe that the technology has reached the level where it is becoming practically usable and related systems confirm that belief, see Section 3 below.

Even if the automatic translation of recognized speech is not perfect, it can serve as a valuable supportive material. For instance, a Czech attendee may have a fair knowledge of English and French, but may easily get lost due to pronunciation difficulties to follow, gaps in his or her grammar knowledge, general vocabulary or specific terminology. Following live subtitles in mother tongue while listening to the foreign language could be of great help. Some level of errors in the subtitles is acceptable *if* the subtitles are sufficiently simultaneous. Our main goal is thus *gist interpretation*, i.e. live supportive translation of speech into text.

Within the ELITR project, we focus on ASR for English, Czech, German, French, Spanish and later Russian and Italian, and targetting the set of 43 languages spoken in member countries of EUROSAI, the association of supreme audit institu-

---

[1] We use the term SLT to refer primarily to simultaneous systems, although off-line spoken language systems can also fall under the same acronym.

[2] http://elitr.eu/

tions of the EU and nearby countries. Experimentally, we include also other languages based on available systems among the research partners in our project, e.g. Hindi.

The scientific motivation for our efforts is to find an approach that allows to assemble laboratory system components to a practically usable product and to document the problems on this journey.

3   Related Systems

Live spoken language translation has been continuously studied for decades, see e.g. Osterholtz et al. (1992); Fügen et al. (2008); Bangalore et al. (2012). Recent systems differ in whether they provide revisions to their previous output (Müller et al., 2016; Niehues et al., 2016; Dessloch et al., 2018; Niehues et al., 2018; Arivazhagan et al., 2020), or whether they only append output tokens (Grissom II et al., 2014; Gu et al., 2017; Arivazhagan et al., 2019; Press and Smith, 2018; Xiong et al., 2019; Ma et al., 2019; Zheng et al., 2019).

Müller et al. (2016) were probably the first to allow output revision when they find a better translation. Zenkel et al. (2018) released a simpler setup as an open-source toolkit consisting of a neural speech recognition system, a sentence segmentation system, and an attention-based translation system providing also some pre-trained models for their tasks. (Zenkel et al., 2018) evaluated only the quality of the output translations using BLEU and WER metrics.

Zheng et al. (2019) proposed a new approach with a delay-based heuristic. The model decides to read more input (or wait for it) or write the translation to the output. Ma et al. (2019) introduced a simple wait-k heuristic: output is emitted after k words of input. Both works are limited to simultaneous *translation*, i.e. they start from text and only simulate the speech-like input by processing input word by word.

Arivazhagan et al. (2020) combine industry-grade ASR and MT and allow output revisions by re-translating the source from scratch as it grows to decrease the latency, providing acceptable translation quality at the price of a higher number of text revisions.

4   ELITR Flexible Architecture

We always strive for the best performance for each considered language pair. With the perpetual com-

petition in ASR and MT research, it is not surprising that there is no universally best solution. The interplay of available data, underlying method, the actual implementation as well as its adaptability to the domain of interest requires different choices for different languages.

Furthermore, the top-performing components are often available only at universities or research labs, as more or less stable research prototypes. Releasing any such system, let alone their combination so that they could be easily deployed by lay users is surely possible, but it would require considerable additional implementation resources.

The ELITR architecture (Franceschini et al., 2020) tackles this integration problem by means of a distributed connection-based client-server application. Research labs provide their components by connecting to a central point (the "mediator") which in turn uses these "workers" to satisfy users' stream processing requests. A technical benefit is that worker connection is issued *from* the secured networks of the labs so it usually does not run into firewall issues.

5   System Components

All our workers, except recent online sequence-to-sequence ASRs, have been described in our IWSLT 2020 shared task submission (Macháček et al., 2020). We briefly summarize them in following sections.

5.1   ASR Systems in ELITR

All our ASR systems provide online processing with low latency and hypotheses updates, as in KIT Lecture Translator (Müller et al., 2016). We use the hybrid ASR models based on Janus from KIT Lecture Translator, for German and English, as well as recent neural sequence-to-sequence ASR models trained on the same data (Nguyen et al., 2020). For Czech ASR, we use a Kaldi hybrid model trained on a Corpus of Czech Parliament Plenary Hearings (Kratochvíl et al., 2019). Czech sequence-to-sequence ASR is a work in progress.

5.2   MT Systems in ELITR

We use bilingual NMT models for some high resource and well-studied language pairs e.g. for English-Czech (Popel et al., 2019; Wetesko et al., 2019). For other targets, we use multi-target models, e.g. an English-centric universal model for

Index Name	Worker	Source Lang	Target Lang	sacreBLEU	WER	Words	Lines
auto-iwslt2020-antrecorp(ASR)	en-EU-lecture_KIT-s2s	EN	–	–	0.46	6634	571
auto-iwslt2020-antrecorp(MT)	rb-EU_fromEN-en_to_41_all	EN	CS	13.66	–	5345	571
auto-iwslt2020-antrecorp(MT)	rb-EU_fromEN-en_to_41_all	EN	DE	17.95	–	6119	571
auto-asr-english-auditing(ASR)	en-EU-lecture_KIT-s2s	EN	–	–	0.37	24530	2220
auto-asr-english-auditing(MT)	rb-EU_fromEN-en_to_41_all	EN	CS	16.45	–	43146	2170
auto-asr-english-auditing(MT)	rb-EU_fromEN-en_to_41_all	EN	DE	19.60	–	18616	2220
auto-iwslt2020-khanacademy(ASR)	en-EU-lecture_KIT-s2s	EN	–	–	0.55	4470	538

Table 1: An overview of WER, sacreBLEU scores on the ELITR test set domain and the size of gold transcript for reference.

42 languages (Johnson et al., 2017). The models are mostly Transformers (Vaswani et al., 2017) but we improve their performance in massively multi-lingual setting by extra depth (Zhang et al., 2020).

5.3 Interplay of ASR and MT

Connecting ASR and MT systems is not straightforward because MT systems assume input in the form of complete sentences. We follow the strategy of Niehues et al. (2016), first inserting punctuation into the stream of tokens coming from ASR (Tilk and Alumäe, 2016), breaking it up at full stops and sending individual sentences to MT, either as unfinished sentence prefixes, or complete sentences. We are using re-translation, as ASR or punctuation updates are received.

Currently, the main problem is that punctuation prediction does not have access to the sound any more, so intonation cannot be considered. Another problem is the information structure of translated sentences, where MT systems tend to "normalize" word order. The loss of topicalization reduces understandability of the stream of uttered sentences.

For the future, we consider three approaches: (1) training MT on sentence chunks, (2) including sound input in punctuation prediction, or (3) end-to-end neural SLT.

6 Evaluation

We evaluate our systems in multiple ways:
- The individual components are evaluated in isolation during deployment, and on a comparable test set. compared with baseline by the MT quality.
- English to Czech and German simultaneous translation of non-native speech was evaluated on a shared task at IWSLT 2020 (Ansari et al., 2020). We validated our candidate systems, and submitted the best one as Macháček et al. (2020). The results showed that the speech recognition of the non-native speech in the test

set was problematic, and resulted to inadequate translations. However, the systems were not yet adapted to non-natives or for the domain. It is a challenge for future work. It can be achieved by speaker adaptation of the ASR from a small sample of the speaker, by multi-lingual ASR, and by collecting non-native speech training data, as AMI corpus.
- We regurarly test our system end-to-end on linguistic seminars in Czech or English. The participants are Czech or English speakers and do not need any assistance with the language, so we can not receive relevant feedback about adequacy and fluency. However, we test our system in end-to-end fashion and face engineering problems and technical issues on all layers from sound acquisition through network connections, worker configuration to subtitle presentation.
- We are currently running a user study with non-German speakers watching German videos with our online subtitles, see Section 7.1. We aim to measure the comprehension loss caused by different subtitling options, latency or flicker.

For comparability across our project partners but also across external research labs, we publicly released a tool for evaluation, SLTev[3] (Ansari et al., 2021) and a test set.[4] The results of our currently best candidates on the testset are in Table 1.

It is important to realize that the evaluation for quality, latency and stability on a speech-to-text test set in lab conditions is necessary, but not sufficient for assessing the practical usability of the system. Practical usability has to include the presentation layer (Section 7) and tests in live sessions or rigorously controlled conditions.

---

[3]https://github.com/ELITR/SLTev
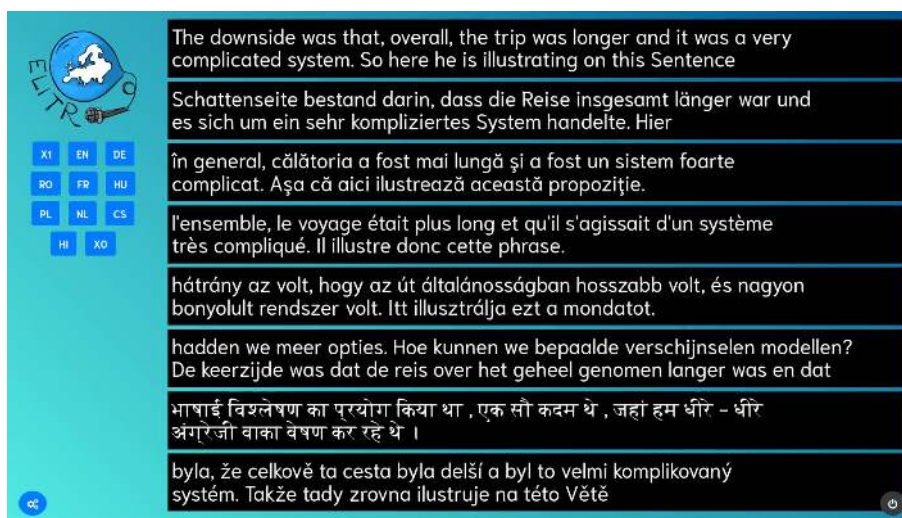[4]https://github.com/ELITR/elitr-testset

Figure 1: A screenshot of subtitle view from a presentation given in Czech (last row), automatically transcribed and translated to English (first row) and then from English into several other languages. The various processing and network delays lead to slightly different timing of each of the languages.

7 Presentation Techniques

The last step in an SLT system is the delivery of the translated content to the user. Our goal stops at the textual representation, i.e. we do not include speech synthesis and delivery of the sound, which would bring yet another set of design decisions and open problems, see e.g. Zheng et al. (2020).

We experiment with two different views for our text output, both implemented as web applications. The "subtitle view" is optimized toward minimal use of screen space. Only two lines of text are available which leaves room either for e.g. a streamed video of the session or the slides, or for many languages displayed at once, if the screen is intended for a multi-lingual audience. The "paragraph view" provides more textual context to the user.

7.1 Subtitle View

The subtitle view offers a simple interface with a HLS stream of the video or slides and one or more subtitles streams.

Section 7.1 presents one screenshot of this view, selected from a screencast. Instead of presenting the video, we use the screen space to show seven target languages, in addition to the live transcript of the source Czech.

We are probably the first to combine re-translation strategy with the presentation in such limited space. To limit text flicker as re-translations are arriving, we had to introduce a critical component after the MT output called Sub-

titler (Macháček and Bojar, 2020). The subtitler allows us to choose the level of updates, trading simultaneity for stability. A user study on the impact of this choice on comprehensibility is currently running. We believe that the ideal choice will depend also on the users' knowledge of the source and target languages and their reading speed.

Even if the flicker is avoided, there remains the main drawback of the subtitle view, the limited context. Both ASR and MT suffer from natural errors. Following the output of ASR (subtitles of the speakers' language) is easier, the erroneous hypotheses still somehow resemble the original sound, so the user can recover from recognition errors.

The output of MT causes a substantially bigger challenge for the user because the sentences are mostly rendered as fully fluent but containing unexpected words or information structure. With only two lines of text available, the user does not see sufficient number of words to let the brain "make up" or reconstruct the original meaning from pieces. The short-term memory of recently processed text does not seem to be sufficient for this type recovery, while seeing the words in larger context gives the user a better chance.

7.2 Paragraph View

We created the paragraph view primarily to improve the chances of recovery from translation errors. The added benefit is a clearer indication of which sentences are finished and which may still
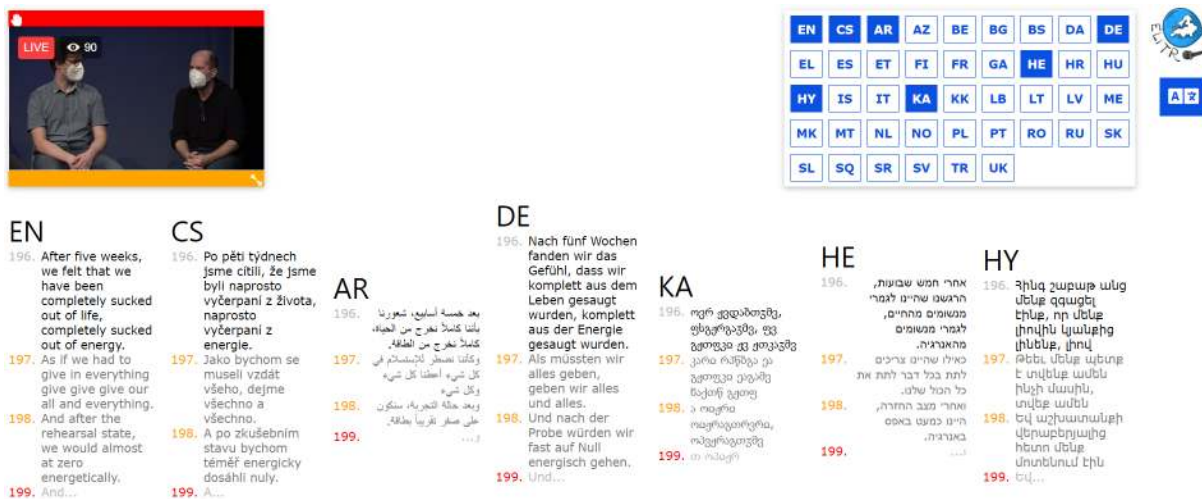
Figure 2: Sample screenshot from the paragraph view of simultaneous translation output on a live discussion of THEaiTRE project. The talk was given in Czech, interpreted into English by a human interpreter, automatically recognized (the leftmost EN column) and translated into 41 languages. Sentence indices correspond to each other across languages in all columns. Sentences in black are "stable", no update will arrive. Sentences in dark gray and with yellow index number are tentative, the segmentation (and thus translation) still may change. The last sentence (light gray) is still being uttered and is thus highly unstable.

change. Without any settings, users can simply decide if they want to read the less stable gray output, or rather wait for the stable segments.

The view is illustrated in Figure 2, with Czech as source and two more languages shown. More than three languages can be presented as well but they generally do not fit. The scrolling of the languages is not fully parallel by our design decision to prefer contiguous columns within each language over tabular synchronous presentation. One important aspect is however synchronized, and that is the stable "level" for finalized sentences: the completed text (shown in black) is aligned at the bottom across languages while the unstable hypotheses flicker below the level, varying in their length as needed.

A drawback of this interface is that all errors such as laughable or obscene words in MT output remain on screen for a long time, needlessly distracting the user.

8 Conclusion

We presented a complex system for live subtitling of conference speech into many target languages, composed of research prototype components but still serving in close-to-production setting. New and updated models and other components can be easily plugged in and tested in practice.

As of now, we are at a good starting point for gradual model improvement and field tests. One of them is very likely to be the META-FORUM 2021 but we are also searching for suitable events with more than one official communication language.

Demonstration videos from past sessions can be found in the blogposts at `https://elitr.eu/blog/`.

References

Ebrahim Ansari, Amittai Axelroad, Nguyen Bach, Ondrej 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, and Changhan Wang. 2020. Findings of the IWSLT 2020 Evaluation Campaign. In *Proceedings of the 17th International Conference on Spoken Language Translation (IWSLT 2020)*, Seattle, USA.

Ebrahim Ansari, Ondřej Bojar, Barry Haddow, and Mohammad Mahmoudi. 2021. SLTev: Comprehensive evaluation of spoken language translation. In *Proceedings of the System Demonstrations of*

the 16th Conference of the European Chapter of the Association for Computational Linguistics, Kyiv, Ukraine. Association for Computational Linguistics.

Naveen Arivazhagan, Colin Cherry, Wolfgang Macherey, Chung-Cheng Chiu, Semih Yavuz, Ruoming Pang, Wei Li, and Colin Raffel. 2019. 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.

Naveen Arivazhagan, Colin Cherry, Wolfgang Macherey, and George Foster. 2020. Re-translation versus streaming for simultaneous translation. In *Proceedings of the 17th International Conference on Spoken Language Translation*, pages 220–227, Online. Association for Computational Linguistics.

Srinivas Bangalore, Vivek Kumar Rangarajan Sridhar, Prakash Kolan, Ladan Golipour, and Aura Jimenez. 2012. Real-time incremental speech-to-speech translation of dialogs. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 437–445, Montréal, Canada. Association for Computational Linguistics.

Ondřej Bojar, Dominik Macháček, Sangeet Sagar, Otakar Smrž, Jonáš Kratochvíl, Ebrahim Ansari, Dario Franceschini, Chiara Canton, Ivan Simonini, Thai-Son Nguyen, Felix Schneider, Sebastian Stücker, Alex Waibel, Barry Haddow, Rico Sennrich, and Philip Williams. 2020. ELITR: European live translator. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 463–464, Lisboa, Portugal. European Association for Machine Translation.

Florian Dessloch, Thanh-Le Ha, Markus Müller, Jan Niehues, Thai-Son Nguyen, Ngoc-Quan Pham, Elizabeth Salesky, Matthias Sperber, Sebastian Stüker, Thomas Zenkel, and Alexander Waibel. 2018. KIT lecture translator: Multilingual speech translation with one-shot learning. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pages 89–93, Santa Fe, New Mexico. 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.

Christian Fügen, Alex Waibel, and Muntsin Kolss. 2008. Simultaneous translation of lectures and speeches. *Springer Netherlands, Machine Translation, MTSN 2008, Springer, Netherland*, 21(4).

Alvin Grissom II, He He, Jordan Boyd-Graber, John Morgan, and Hal Daumé III. 2014. Don't until the final verb wait: Reinforcement learning for simultaneous machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1342–1352, Doha, Qatar. Association for Computational Linguistics.

Jiatao Gu, Graham Neubig, Kyunghyun Cho, and Victor O.K. Li. 2017. Learning to translate in real-time with neural machine translation. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 1053–1062, Valencia, Spain. Association for Computational Linguistics.

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.

Jonáš Kratochvíl, Peter Polák, and Ondřej Bojar. 2019. Large corpus of czech parliament plenary hearings. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

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. 2019. 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.

Dominik Macháček and Onřej Bojar. 2020. Presenting simultaenous translation in limited space. In *Proceedings of the 20th Conference ITAT 2020: Workshop on Automata, Formal and Natural Languages (WAFNL 2020)*. To be published.

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.

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, Ngoc-Quan Pham, Sebastian Stueker, and Alex Waibel. 2020. High performance sequence-to-sequence model for streaming speech recognition. *arXiv preprint arXiv:2003.10022*.

Jan Niehues, Thai Son Nguyen, Eunah Cho, Thanh-Le Ha, Kevin Kilgour, Markus Müller, Matthias Sperber, Sebastian Stüker, and Alex Waibel. 2016. Dynamic transcription for low-latency speech translation. In *17th Annual Conference of the International Speech Communication Association, INTERSPEECH 2016*, volume 08-12-September-2016 of *Proceedings of the Annual Conference of the International Speech Communication Association. Ed. : N. Morgan*, pages 2513–2517. International Speech and Communication Association, Baixas.

Jan Niehues, Ngoc-Quan Pham, Thanh-Le Ha, Matthias Sperber, and Alex Waibel. 2018. Low-latency neural speech translation. In *Interspeech 2018*, Hyderabad, India.

L. Osterholtz, C. Augustine, A. McNair, I. Rogina, H. Saito, T. Sloboda, J. Tebelskis, and A. Waibel. 1992. Testing generality in janus: a multilingual speech translation system. In *[Proceedings] ICASSP-92: 1992 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 209–212 vol.1.

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.

Ofir Press and Noah A. Smith. 2018. You may not need attention.

Ottokar Tilk and Tanel Alumäe. 2016. Bidirectional recurrent neural network with attention mechanism for punctuation restoration. In *Interspeech 2016*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 6000–6010. Curran Associates, Inc.

Joanna Wetesko, Marcin Chochowski, Pawel Przybysz, Philip Williams, Roman Grundkiewicz, Rico Sennrich, Barry Haddow, Antonio Valerio Miceli Barone, and Alexandra Birch. 2019. Samsung and University of Edinburgh's System for the IWSLT 2019. In *IWSLT*.

Hao Xiong, Ruiqing Zhang, Chuanqiang Zhang, Zhongjun Hea, Hua Wu, and Haifeng Wang. 2019. Dutongchuan: Context-aware translation model for simultaneous interpreting.

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, NUMBER 11, p. 125-135.*

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, Renjie Zheng, Mingbo Ma, and Liang Huang. 2019. Simpler and faster learning of adaptive policies 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 1349–1354, Hong Kong, China. Association for Computational Linguistics.

Renjie Zheng, Mingbo Ma, Baigong Zheng, Kaibo Liu, Jiahong Yuan, Kenneth Church, and Liang Huang. 2020. Fluent and low-latency simultaneous speech-to-speech translation with self-adaptive training. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3928–3937, Online. Association for Computational Linguistics.