INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

Lost in Interpreting: Speech Translation from Source or Interpreter?

Dominik Macháček, Matúš Žilinec, Ondřej Bojar

Charles University, Faculty of Mathematics and Physics
Czech Republic
{surname}@ufal.mff.cuni.cz

Abstract

Interpreters facilitate multi-lingual meetings but the affordable set of languages is often smaller than what is needed. Automatic simultaneous speech translation can extend the set of provided languages. We investigate if such an automatic system should rather follow the original speaker, or an interpreter to achieve better translation quality at the cost of increased delay.

To answer the question, we release Europarl Simultaneous Interpreting Corpus (ESIC), 10 hours of recordings and transcripts of European Parliament speeches in English, with simultaneous interpreting into Czech and German. We evaluate quality and latency of speaker-based and interpreter-based spoken translation systems from English to Czech. We study the differences in implicit simplification and summarization of the human interpreter compared to a machine translation system trained to shorten the output to some extent. Finally, we perform human evaluation to measure information loss of each of these approaches.

Index Terms: speech translation, machine translation, simultaneous interpreting corpus, interpreting

1. Introduction

Multilingual events with participants without a common language are often simultaneously interpreted by humans. Automatic simultaneous speech translation can increase the language coverage where human interpreting is not available, e.g. because of capacity reasons. Assuming the presence of a human interpreter, speech translation can rely on the original speech as the source, or by translating the speech of the interpreter. In this work, we compare the features of these two options.

The direct source-to-target translation is supposed to be fast (no latency introduced by the interpreter), and more literal, and therefore very detailed. However, the verbosity might be uncomfortable for final users to follow, if the speech is too fast or disfluent. The indirect interpreter-to-target translation might benefit from the fact that interpreters tend to compress and simplify [1, 2], on the other hand, it could decrease adequacy.

In this work, we examine two possible sources and one target language. However, we put aside the effects of varying quality of speech recognition and machine translation. They can favor any option, depending on the specific version of the tools and other conditions. We focus on the evaluation of latency, shortening and simplification, and human assessment of information loss. We prepare a new evaluation corpus ESIC (Europarl Simultaneous Interpreting Corpus v.1.0) with 10 hours of English speeches with transcripts, translations and transcripts of simultaneous interpreting into Czech and German.

2. Related Work

The plenary sessions of European Parliament (EP) are a useful source of parallel data, known well from the multi-parallel text-to-text corpus Europarl [3]. The recent speech-to-text corpus Europarl-ST [4] is a collection of short audio-translation segments for bilingual or multi-target speech-to-text translation. It contains only the audio of original speakers, not the interpreters.

The corpora EPTIC [2], EPIC [5] and EPIC-Ghent [6] are small collections of transcribed interpretings from European Parliament created for analyses of interpreting. They contain only selected languages, not including English, German and Czech. They do not contain timestamps and audios of interpreting, and their accessibility is restricted. The other corpora of simultaneous interpreting [7, 8] focus on other languages.

Additionally, text simplification in the context of machine translation remains an open problem. The existing methods focus on augmenting the translation model with length tokens or positional encoding to control the length of the output text [9, 10]. For an overview, we refer the reader to Lakew [11].

3. ESIC: Corpus Composition

Since 2008, the EP is publishing the audios of simultaneous interpreting into all 22 EU official languages in that time. Until 2011, it was publishing the revised transcripts and translations into all EU languages. The period of 2008 to 2011 is a valuable resource containing parallel revised translations and simultaneous interpreting, which we decided to study.

We focus on English, the most common European lingua franca, as the source, and on simultaneous interpreting into German and Czech. German is a language with second most speakers in EU, and it often serves as interpreting target at many international events. Czech is an example target language into which it might be translated automatically.

We downloaded the data and aligned the revised transcripts and audio by metadata. We processed the speeches with automatic diarization [12] to roughly annotate their beginning and end timestamps in long recordings of the whole sessions. For simplicity, we decided to exclude the president because his or her utterances while chairing the sessions were often not transcribed, or not word-for-word. We also excluded speeches which we could not align due to error in metadata or in automatic processing, which were shorter than 30 seconds, or whose Czech translation or interpreting was missing.

Next, we selected 10 hours of speeches into validation and evaluation set. We decided to eliminate the potentially malicious overlap of ESIC dev-test with Europarl-ST train set. We identified the speakers of Europarl-ST English-German dev-test, found all their speeches in our data, and included them into ESIC dev-test. To cover full 10 hours, we added additional 28 randomly selected speeches, regardless the speakers in Europarl-ST. We marked them so that the users can be aware.

http://dx.doi.org/10.21437/Interspeech.2021-2232

Table 1: *Size statistics of ESIC corpus. The two numbers in each cell are the number of sentences (or documents, in the row of Verbatim transcription), and number of words.*

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

3.1. Manual Revisions

We manually revised the segmentation into individual speeches in all three tracks (English source, Czech and German interpreting) because the automatic diarization was inaccurate at beginnings and ends. In the next steps, we manually transcribed the interpreters following fixed annotation guidelines. Our annotators marked false starts, unintelligible words, short insertions in different languages and swapping voices, so that ESIC users can decide to handle them in a particular way. They transcribed and marked the segments which could not be easily transferred from orthography to verbatim, e.g. the non-canonical forms of numerals, dates, loaned named entities and acronyms. They inserted orthographic punctuation and spelling, but did not do any changes in syntax, even when the interpreter's syntax could be considered as ungrammatical. Hesitations were not marked. In sum, we ended up with three versions: Revised as downloaded from the web, Verbatim which does not include any punctuation, but does include false starts, and Ortho with punctuation and without false starts.

The transcripts of English sources were revised in the same way as those of interpreters', but the annotator re-used the transcripts from the web, which were manually revised and normalized by EP staff for comfortable reading. They often differed from the verbose ones in the way of addressing the president and Parliament at the introduction, in the correction of disfluencies and grammar, use of more formal named entities or decompressed acronyms, and removal of side and organizational comments. Also, the concluding "thank you" to the president was added by our revision.

Furthermore, our annotator marked, with the use of the video-recording, whether the speech was spontaneous, or read, because we believe it has a big impact on the grammar, style and complexity of translation. In rare cases, we excluded speeches given in another language than in English, but short code switching, e.g. the salutation of the president in his or her native language, were kept for authenticity.

Finally, we used MAUS forced aligner [13] for English, German and Czech to obtain the word-based timestamps. The corpus statistics are in Table 1.

3.2. Ethics

We received the authorisation to repackage and publish the texts and audios of the speakers on the EP plenary sessions, and the transcripts of interpreters[1]. Since the interpreters' voices are considered as personal data, we do not publish them together with the corpus. However, they are publicly available on the web of EP, and we can publish the links and instructions that every user of our corpus can follow to obtain them.

[1] Available at `http://hdl.handle.net/11234/1-3719`.

4. Translation Systems

In the next sections, we compare three options for translation of English speech into Czech: human interpreting into Czech (CS-INT), human interpreting into German (DE-INT) followed by a machine translation system into Czech (DE-CS), and a machine translation model directly into Czech, which was additionally trained to shorten source text (EN-CS).

4.1. Machine Translation

EN-CS is a Transformer-Base [14] machine translation model trained using Marian [15] on CzEng 1.7 [16] using the default hyperparameters. It was biased during training by providing training examples illustrating shortening. Specifically, sentence pairs from the parallel corpus were selected only if the Czech sentence had not more than 86% of the number of subword units compared to the English counterpart. Given that in the CzEng corpus, Czech sentences are on average 10% longer than their English translations in terms of subword units, our requirement corresponds to EN:CS compression factor of 1:0.78.

In comparison to an identical model trained on the full corpus, we observed a decrease in both mean length of the translation and BLEU score with the shortening model.

Furthermore, we observed that the model often performs shortening by replacing words and phrases with their synonyms with fewer subword units, but preserves the syntax, which does not significantly differ from the baseline non-shortening model's translation. This is in contrast to human interpreting strategy [1]. Human interpreters tend to segment the source sentence into small units and translate them as individual sentences. Furthermore, they use generalization and summarization of the whole clauses, and other techniques such as passivization to consolidate the word order between source and target.

DE-CS is trained on 8M sentence pairs from Europarl and Open Subtitles [3, 17], the only public parallel corpora of German and Czech, and validated on newstest. The Transformer-based system runs in Marian [15] and reaches 18.8 cased BLEU on WMT newstest-2019. It is not adapted for simultaneous translation which would need translation stability and partial translation for partial sentences [18, 19].

4.2. Low-Latency ASR

We use online German and English ASR systems originally prepared for lectures [20]. They emit partial hypotheses in real time, and correct them as more context is available. German is a hybrid HMM-DNN model (DE ASR). The same system was used also by KIT Lecture Translator [21]. English is neural sequence-to-sequence ASR [22]. They are connected in a cascade with a tool for removing disfluencies and inserting punctuation [23] and with the MT systems. The cascade is the same as the one of the ELITR project at IWSLT 2020 [24].

5. Latency

We aim to compare the latency of interpreting and machine translation. Note that the comparison is inevitably limited by different output modalities. The interpreters produce speech, and the machine translation text. We disregard the perception effects of hearing versus reading.

We need to assess the time when each word in source, interpreting and machine translation was produced. For the source and interpreting, we have word-based timestamps from forced alignment tool. For the re-translating machine translation, we use the finalization time of a target word as in [19]. It is the first time when the system produces the word, and the word

and all its preceding words remain unchanged until the end of the session. This definition is rather harsh because it penalizes subtle, cosmetic changes in translation output the same way as meaning-altering re-translations. It is possible that a real user reads the translation earlier than at finalization time, and does not notice short flicker in previous words. However, the finalization time is an upper bound for the word production time.

The "latency" is the difference of times of the source word and its "corresponding" word in the target. We assess the correspondence with automatic word alignment.

5.1. Word Alignment

We aligned English source transcripts and target interpreting or machine translation at the word level with fast_align [25] after tokenizing [26] and trimming them to 5 characters as a trivial form of lemmatization. We processed all 370 ESIC documents, treating each as a single sentence. We added relevant sentence-aligned texts to fast_align training data, to expand the vocabulary: revised translations of Europarl (around 4 thousands documents from the same period) for interpreting, and the source and target sentence prefixes for machine translation. We obtained forward and backward alignments, and removed those going back in time, assuming that the interpreters do not risk predicting content. Finally, we intersected them. Based on a small manual check, the resulting word alignments were reasonably good, despite that fast_align is designed for individual sentences and our documents were much longer.

5.2. Latency Comparison

The latency is summarized in Table 2. Both CS-INT and DE-INT have average latency around 4 seconds. In 90% of the source words that were aligned to any target word, the latency is below 7 seconds. In small number of cases, in around 1%, the latency is larger than 23 seconds. It can be caused either by interpreters using so long translation unit, or a rare error in the automatic alignment. The methodology is the same for all options, therefore we assume that the error rate is homogeneous, although unknown, so the results are comparable.

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

The systems thus translate much longer units than interpreters, and therefore have larger latency. We hypothesize that more advanced translation system could have latency comparable to the interpreter. Assuming that the interpreters always wait optimally for meaningful translation units, their latency is an upper bound for the waiting. Machine processing (speech recognition and translation) can take up to 1 second. ESIC corpus can serve for tuning the parameter k of wait-k models [27] for simultaneous translation, so the resulting latency of wait-k is the same as interpreters'.

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

Table 2: *Latency of interpreting and machine translation from English to Czech (white background), based on automatic word alignments, in seconds. Gray rows break down the two intermediate components of the indirect translation: English-to-German interpreter and German-to-Czech translation. The percentile indicates that, e.g. 90% of aligned words fit under 7 sec.*

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

6. Shortening and Complexity

We aim to compare the shortening and simplification capability of interpreting vs direct machine translation systems.

First, the translation length. Syllables are units independent on the orthography and phonemic inventory of the languages, and they are capable to express shortening rate of translation into multiple languages. Therefore, we used grapheme-to-phoneme and syllabification tool [28] for estimating the number of syllables in English, Czech and German source, interpreting and translation. The results are in Table 3. We also demonstrate that German uses more characters per syllable than Czech, due to smaller character inventory. This fact has to be considered especially in speech-to-text translation.

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

Next, we compare the vocabulary complexity. We rank Czech words from the CzEng corpus by frequencies, such that the most common word has rank 1, and the least common word has the rank of number of unique words. The "comma" and "full stop" characters were removed before the evaluation. Table 4 shows the mean and standard deviation of log ranks for each system across the documents in the test set. We test whether the mean log rank of EN-CS is statistically equal to that of DE-CS. Using the two-sample Z-test, we reject this hypothesis with $p < 0.01$. Thus, we conclude that the translations EN-CS (machine) and CS-REF (human), which do not contain any interpreter component, use a more complex vocabulary than both setups involving an interpreter, CS-INT and DE-CS.

7. Quality

We estimate the quality of machine translation with an automatic metric, and manually assess content preservation.

Table 3: *Length rate of source to target of ESIC test set. For example,* CS-REF *has 1.19-times more syllables than English source. There is average and standard deviation on all test documents.*

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

Table 4: *Mean and standard deviation of log word frequency ranks calculated from translations of the test set. The column "words" denotes the sample size (number of words in the translation). The proportion of out-of-vocabulary words is less than 0.5 % for each system.*

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

7.1. BLEU against two References

In Table 5, we provide the BLEU [29] score of the indirect translation of German interpreting (DE-CS) and the direct EN-CS translation. We measure the score against two possible references: the revised text translation, and transcript of Czech interpreting. The sources are gold transcripts, not ASR, therefore it is an upper bound for translation quality in a real event.

We expected that DE-CS will be closer to CS-INT reference than EN-CS, but it is not. It might be caused by different interpreting strategies, and variability of translation, and too literal translation from German. We however refrain from the interpretation that DE-CS is of lower quality, since it has been previously shown that BLEU negatively correlates with simplicity [30].

7.2. Content Preservation

To compare the difference in text simplification between machine translation and a human interpreter, we manually check the amount of information from the source text preserved in the translation. We employed two human annotators. They are both non-experts on the EP debates, non-native speakers of English, and native speakers of Czech. The first one, a professional translator, worked 5 hours and annotated 107 sentences. The second one, a computer linguist, contributed 20 sentences (1 hour). The annotators were provided with English revised transcripts of the whole document, and the translation candidates of automatic systems, interpreting and reference in Czech. They were all blinded and in random order. One random sentence from the source document was highlighted for assessment. The annotators were asked to express to what extent the information from the highlighted source sentence was preserved in the translation candidates, on a scale from 0 to 100. For comparability, they were asked to rate all the 6 candidates at once.

Table 6 indicates that EN-CS applied to the golden transcript preserves a similar amount of information as the manual translation. Involving any interpreter (DE-CS and CS-INT) leads to a considerable loss. ASR as the source for MT instead of gold transcripts significantly reduces translation quality, and loses further information (EN ASR+EN-CS and DE ASR+DE-CS).

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

Table 5: *BLEU score between* EN-CS, DE-CS *and both Czech reference translations. BLEU requires a 1-1 correspondence between candidate and reference segments. We either treat the whole test set as one segment ("BLEU agg") or each speech in the test set as one segment ("BLEU one").*

Reference	System	BLEU agg	BLEU one
CS-INT	EN-CS	21.4	13.8
CS-INT	DE-CS	19.9	10.4
CS-REF	EN-CS	27.6	22.6
CS-REF	DE-CS	21.1	13.2

Table 6: *Manual assessment of information preserved.*

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

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

8. Conclusion

In this work, we release ESIC 1.0, a corpus with 10 hours of European Parliament speeches in English with transcripts, translations, and transcripts of simultaneous interpreting into Czech and German. We make it available for future work in speech translation and other areas:

`http://hdl.handle.net/11234/1-3719`

We conclude that the automatic BLEU score is unable to distinguish whether the source-to-target or interpreter-to-target translation is better, due to the simplification feature of interpreting. We compare direct and indirect speech translation by latency, and show that the indirect option could be comparable to relay interpreting. On the other hand, interpreter-based translation leads to shorter targets with significantly less complex vocabulary. A limited human assessment shows that more information is preserved in direct translation than in interpreting-based translations, and that far more content survives in translation from gold transcripts than from online ASR.

9. Acknowledgements

The research was partially supported by the grants 19-26934X (NEUREM3) of the Czech Science Foundation, H2020-ICT-2018-2-825460 (ELITR) of the EU, and 398120 of the Grant Agency of Charles University.

10. References

[1] H. He, J. Boyd-Graber, and H. Daumé III, "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*. Association for Computational Linguistics, 2016, pp. 971–976.

[2] S. Bernardini, A. Ferraresi, and M. Milicevic, "From EPIC to EPTIC — Exploring simplification in interpreting and translation from an intermodal perspective," *Target*, vol. 28, pp. 61–86, 05 2016.

[3] P. Koehn, "Europarl: A Parallel Corpus for Statistical Machine Translation," in *Conference Proceedings: the tenth Machine Translation Summit*, AAMT. AAMT, 2005, pp. 79–86.

[4] J. Iranzo-Sánchez, J. A. Silvestre-Cerdà, J. Jorge, N. Roselló, A. Giménez, A. Sanchis, J. Civera, and A. Juan, "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)*, 2020, pp. 8229–8233.

[5] A. Sandrelli and C. Bendazzoli, "Tagging a corpus of interpreted speeches: the European parliament interpreting corpus (EPIC)," in *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*. European Language Resources Association (ELRA), 2006.

[6] B. Defrancq, "Corpus-based research into the presumed effects of short evs," *Interpreting*, vol. 17, 04 2015.

[7] I. Temnikova, A. Abdelali, S. Hedaya, S. Vogel, and A. Al Daher, "Interpreting strategies annotation in the WAW corpus," in *Proceedings of the Workshop Human-Informed Translation and Interpreting Technology*. Association for Computational Linguistics, Shoumen, Bulgaria, 2017, pp. 36–43.

[8] J. Pan, "The Chinese/English political interpreting corpus (CEPIC): A new electronic resource for translators and interpreters," in *Proceedings of the Human-Informed Translation and Interpreting Technology Workshop (HiT-IT 2019)*. Incoma Ltd., Shoumen, Bulgaria, 2019, pp. 82–88.

[9] Y. Kikuchi, G. Neubig, R. Sasano, H. Takamura, and M. Okumura, "Controlling output length in neural encoder-decoders," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2016, pp. 1328–1338.

[10] S. Takase and N. Okazaki, "Positional encoding to control output sequence length." Association for Computational Linguistics, 2019, pp. 3999–4004.

[11] S. M. Lakew, "Multilingual neural machine translation for low resource languages," Ph.D. dissertation, University of Trento, 2020.

[12] M. Rouvier, P. Gay, E. Khoury, T. Merlin, and S. Meignier, "An open-source state-of-the-art toolbox for broadcast news diarization," in *in Proc. of Interspeech*, 2013.

[13] T. Kisler, U. Reichel, and F. Schiel, "Multilingual processing of speech via web services," *Computer Speech & Language*, vol. 45, pp. 326 – 347, 2017.

[14] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems 30*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds. Curran Associates, Inc., 2017, pp. 6000–6010.

[15] M. Junczys-Dowmunt, R. Grundkiewicz, T. Dwojak, H. Hoang, K. Heafield, T. Neckermann, F. Seide, U. Germann, A. Fikri Aji, N. Bogoychev, A. F. T. Martins, and A. Birch, "Marian: Fast neural machine translation in C++," in *Proceedings of ACL 2018, System Demonstrations*. Association for Computational Linguistics, 2018, pp. 116–121.

[16] O. Bojar, O. Dušek, T. Kocmi, J. Libovický, M. Novák, M. Popel, R. Sudarikov, and D. Variš, "CzEng 1.6: Enlarged Czech-English Parallel Corpus with Processing Tools Dockered," in *Text, Speech, and Dialogue: 19th International Conference, TSD 2016*, no. 9924, Masaryk University. Springer International Publishing, 2016, pp. 231–238.

[17] P. Lison and J. Tiedemann, "OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles," in *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*. European Language Resources Association (ELRA), 2016, pp. 923–929.

[18] J. Niehues, N.-Q. Pham, T.-L. Ha, M. Sperber, and A. Waibel, "Low-latency neural speech translation," in *Proc. Interspeech 2018*, 2018, pp. 1293–1297. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2018-1055

[19] N. Arivazhagan, C. Cherry, I. Te, W. Macherey, P. Baljekar, and G. F. Foster, "Re-translation strategies for long form, simultaneous, spoken language translation," *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7919–7923, 2020.

[20] E. Cho, C. Fügen, T. Hermann, K. Kilgour, M. Mediani, C. Mohr, J. Niehues, K. Rottmann, C. Saam, S. Stüker, and A. Waibel, "A real-world system for simultaneous translation of german lectures," pp. 3473–3477, 01 2013.

[21] M. Müller, T. S. Nguyen, J. Niehues, E. Cho, B. Krüger, T.-L. Ha, K. Kilgour, M. Sperber, M. Mediani, S. Stüker, and A. Waibel, "Lecture translator - speech translation framework for simultaneous lecture translation." Association for Computational Linguistics, 2016, pp. 82–86.

[22] T.-S. Nguyen, S. Stueker, and A. Waibel, "Super-human performance in online low-latency recognition of conversational speech," 2021.

[23] E. Cho, J. Niehues, and A. H. Waibel, "Segmentation and punctuation prediction in speech language translation using a monolingual translation system," in *IWSLT*, 2012.

[24] D. Macháček, J. Kratochvíl, S. Sagar, M. Žilinec, O. Bojar, T.-S. Nguyen, F. Schneider, P. Williams, and Y. Yao, "ELITR nonnative speech translation at IWSLT 2020," in *Proceedings of the 17th International Conference on Spoken Language Translation*. Association for Computational Linguistics, 2020, pp. 200–208.

[25] C. Dyer, V. Chahuneau, and N. A. Smith, "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*. Association for Computational Linguistics, 2013, pp. 644–648.

[26] P. Koehn and al., "Moses: Open source toolkit for statistical machine translation," in *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*. Association for Computational Linguistics, 2007, pp. 177–180.

[27] M. Ma and al., "STACL: Simultaneous translation with implicit anticipation and controllable latency using prefix-to-prefix framework." Association for Computational Linguistics, 2019, pp. 3025–3036.

[28] U. D. Reichel, "Language-independent grapheme-phoneme conversion and word stress assignment as a web service," in *Elektronische Sprachverarbeitung 2014*, R. Hoffmann, Ed. Dresden, Germany: TUDpress, 2014, vol. 71, pp. 42–49.

[29] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2002, pp. 311–318.

[30] E. Sulem, O. Abend, and A. Rappoport, "BLEU is not suitable for the evaluation of text simplification," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2018, pp. 738–744.