This document is part of the Research and Innovation Action "European Live Translator (ELITR)".
This project has received funding from the European Union's Horizon 2020 Research and Innovation Programme under Grant Agreement No 825460.

Deliverable D4.3

Final Report on Multi-Lingual MT

Dominik Macháček (CUNI), Rico Sennrich (UEDIN),
Philip Williams (UEDIN), Dávid Javorský (CUNI)

Dissemination Level: Public

Final (Version 1.0), 31st March, 2022

Grant agreement no.	825460
Project acronym	ELITR
Project full title	European Live Translator
Type of action	Research and Innovation Action
Coordinator	doc. RNDr. Ondřej Bojar, PhD. (CUNI)
Start date, duration	1st January, 2019, 39 months
Dissemination level	Public
Contractual date of delivery	31st March, 2022
Actual date of delivery	31st March, 2022
Deliverable number	D4.3
Deliverable title	Final Report on Multi-Lingual MT
Type	Report
Status and version	Final (Version 1.0)
Number of pages	33
Contributing partners	CUNI, UEDIN, KIT
WP leader	UEDIN
Author(s)	Dominik Macháček (CUNI), Rico Sennrich (UEDIN), Philip Williams (UEDIN), Dávid Javorský (CUNI)
EC project officer	Luis Eduardo Martinez Lafuente
The partners in ELITR are:	▪ Univerzita Karlova (CUNI), Czech Republic ▪ University of Edinburgh (UEDIN), United Kingdom ▪ Karlsruher Institut für Technologie (KIT), Germany ▪ PerVoice SPA (PV), Italy ▪ alfatraining Bildungszentrum GmbH (AV), Germany
Partially-participating party	▪ Nejvyšší kontrolní úřad (SAO), Czech Republic

For copies of reports, updates on project activities and other ELITR-related information, contact:

doc. RNDr. Ondřej Bojar, PhD., ÚFAL MFF UK
Malostranské náměstí 25
118 00 Praha, Czech Republic

bojar@ufal.mff.cuni.cz
Phone: +420 951 554 276
Fax: +420 257 223 293

Copies of reports and other material can also be accessed via the project's homepage:

`http://www.elitr.eu/`

Contents

1 Executive Summary

This deliverable summarizes the progress in WP4 Multi-Lingual MT during the final third year of the project. The work that has been carried out in two previous years is reported in the initial deliverable D4.1 and in the intermediate D4.2. In this deliverable, we only briefly refer to the previous work, and report all necessary details of new work.

The work package consists of 5 tasks:

T4.1 Baseline MT Models was planned and carried out during the first 6 months of the project. It provided MT systems to the rest of the main processing pipeline, so that integration and technical testing could start. All the details regarding our baseline models are in the previous Deliverable D4.1: Initial Report on Multi-Lingual MT.

T4.2 Document-Level Translation is a research goal somewhat more independent of the remaining tasks. The aim is to substantially improve the practice of handling document-level context across MT processing stages: training, translation and evaluation. In Section 2, we report a new study on the post-processing strategy to improve document-level coherence, and a study for simulated document-level end-to-end evaluation of simultaneous speech-to-text translation.

T4.3 Multi-Target Translation explores the setup most needed for ELITR central event, the EUROSAI congress, where a single speech needs to be translated into up to 43 target languages. In Section 3, we report an update of previously reported English-to-many NMT.

T4.4 Multi-Source Translation aims to improve translation quality by considering other language versions of the same content. The task was scheduled for the third year of the project and could consider both written or spoken multi-source. We created an evaluation test set of parallel speeches and simultaneous interpreting from the European Parliament, and analyzed features, costs and benefits of using either the original, or interpreting as a source for speech translation. More is in Section 4.

T4.5 Flexible Multi-Lingual MT aimed to propose and evaluate NMT system architectures that can translate from one or more language versions provided simultaneously into one or more target languages. The practical limits of neural network capacity had to examined. However, we started by analyzing more fundamental bottleneck that limits the practical usability of multi-lingual NMT: robustness against unstable and unreliable sources in simultaneous SLT. We propose a simple rule-based system for selecting the most reliable source at real-time, e.g. from the original and one or more simultaneous interpreting. See details in Section 5.

2 Task T4.2 Document-Level Machine Translation (CUNI, UEDIN)

2.1 Post-Editing MT for Document-Level Coherence

Following our success with automatic post-editing in increasing document-level consistency (Voita et al., 2019), reported in D4.2, we extended this work in two ways:

- we draw on related research in (sentence-based) automatic post-editing (Pal et al., 2019), extending it to the document-level. Different from our previous experiments, state-of-the-art post-editing systems use not only the translation candidate, but also the original source text as its input.
- we perform a human evaluation focused on adequacy and fluency, confirming that monolingual automatic post-editing only increases fluency, whereas source-based post-editing can increase both.

Our experiments show that document-level automatic post-editing benefits from access to the source text. However, this increases data requirements, requiring some amount of parallel document-level data, whereas we were able to train a monolingual post-editing system based purely on sentence-level parallel data and document-level monolingual data. Details can be found in Wang et al. (2021), which is attached as Appendix A.

2.2 End-to-end Evaluation of Subtitling Comprehension

In CUNI, we examined the readers' ability to comprehend outputs of simultaneous speech translation by using user evaluation for document-level translation quality and subtitles presentation. Our approach relies on users' continuous rating and a follow-up questionnaire. In the continuous rating, the judges express their satisfaction with the subtitles by pressing buttons while following the subtitles. The follow-up questionnaire used direct factual questions and general inquiry of the overall quality in various aspects. Both evaluation methods aimed at testing full user comprehension, i.e. the text fluency, consistency and coherence.

We showed that user comprehension depends on system latency and sometimes the allowed level of rewriting. With flicker, the subtitles were presented immediately as available, but with frequent rewriting, which discomforts the reader. For comparison without flicker, we presented only the final translations without rewriting, but with a large latency. It showed that more source-language experienced users achieved higher comprehension with flicker than without. It means that users have different preference of coherence and fluency in translation when their source language proficiency differs.

We also related comprehension and the reported continuous feedback. The results showed that there was a statistically significant dependence between user comprehension and continuous rating of the judges who have at least intermediate knowledge of the source language. Relying on the follow-up questionnaire is a costly bottleneck that prevents scalability to larger volumes. We thus suggest that in future, after some follow-up works confirm and measure the correlation between continuous rating and comprehension, the continuous rating itself can be used as a cheap and scalable method for manual evaluation of subtitling quality.

Our study is attached as Appendix B.

3 Task T4.3 Multi-Target MT (CUNI, UEDIN, KIT)

We reported on v2.0 of the ELITR OPUS Corpus in Deliverable 1.5. UEDIN have since made some minor updates to the corpus in order to address some issues with language support that were observed in testing of the PV platform. Subsequently, we re-trained the production model replacing the multi-target English-to-41 model with a new English-to-44 model.

3.1 The ELITR OPUS Corpus v2.1

Version 2.1 was generated using the same methodology as v2.0. Since the corpus generation scripts were re-run from scratch, including the scripts that download the individual source corpora, v2.1 benefits from the additions made to OPUS in the intervening period.

The following additional changes were made in this version:

- Support for the Catalan language was added.
- The Serbian portion of the corpus was separated into Cyrillic and Latin parts.
- The Norwegian portion of the corpus was separated into Bokmål and Nynorsk parts.

3.2 The English-to-44 Model

We re-trained our production one-to-many model on a subset of OPUS v2.1 that includes all sentence pairs with English. Catalan and the variants of Serbian and Norwegian are supported

via new tags that can be prepended to the source sentence, adding <ca>, <sr_cyrillic>, <sr_latin>, <nn>, and <nb>, while removing <sr> and <no>. In addition, we trained a deep encoder version of the model that uses 12 layers instead of the standard 6. We evaluated the model using the auto-mt test sets from the ELITR test set. See the description in deliverable D1.6: Year 3 Test Data and in Ansari et al. (2021). Results are given in Table 1.

Target	en-to-41	en-to-44	en-to-44-deep
Arabic	3.9	3.8	3.6
Bosnian	23.5	27.3	28.6
Catalan	-	16.8	17.5
Czech	23.8	25.2	25.8
Danish	27.5	28.8	29.8
German	24.5	26.0	26.8
Spanish	34.4	35.3	36.0
Croatian	13.1	15.1	15.7
Hungarian	16.4	17.0	17.7
Dutch	28.4	29.4	29.9
Polish	22.0	23.0	23.8
Romanian	22.0	22.3	22.8
Russian	13.5	14.6	15.3
Slovak	10.2	11.3	12.0
Serbian	14.7	14.6	13.6

Table 1: BLEU scores for the English-to-44 model on the elitr-test auto-mt test sets.

With the exceptions of Arabic and Serbian, performance is consistently improved over the previous model and the deep model outperforms the standard model. For Serbian, the elitr-test test set uses Latin script. We noticed that the model still sometimes mixes Latin and Cyrillic (though <sr_cyrillic> predominantly generates Cyrillic and <sr_latin> predominantly generates Latin). We believe that this is due to an inadequate separation of scripts in the preparation of the data. Rather than needing to repeat the full process of corpus preparation and model training, this could be addressed with targeted fine-tuning on a cleaned-up version of the dataset. Similarly, we believe that the poor translation quality for Arabic could be addressed through targeted fine-tuning.

4 Task T4.4 Multi-Source MT (CUNI, UEDIN, KIT)

We decided to primarily focus on multi-source simultaneous speech translation. It can be applied to events with simultaneous interpreting. The machine translation should ideally combine the original speech in the first language and one or more simultaneous interpreting into another languages. All these sources should be used for translation into the target languages, into which it is not interpreted, e.g. from financial and capacity reasons. The motivation is higher quality due to lexical disambiguation and independent speech recognition whose errors may complement each other. The intended cost is larger latency due to interpreting delay.

To be able to proceed with this goal, we created ESIC: Europarl Simultaneous Interpreting Corpus. It consists of 10 hours of authentic English speeches given from European Parliament 2008-12, with simultaneous interpreting into Czech and German. The corpus contains 3 audio tracks with manual transcriptions, metadata and parallel translations. It was used in ELITR test set and already described in deliverable D1.6: Year 3 Test Sets. It was published with the paper by Macháček et al. (2021), see Appendix C.

Furthermore, in Macháček et al. (2021) we analyzed the features, benefits and costs of using the original, or the simultaneous interpreting as source of speech translation on ESIC corpus from following point of views:

- **Latency**. We measured the latency of ST following the interpreter and of the ST directly from source. ST following the interpreter is approximately twice slower than ST directly from the source. It can be comparable to interpreting through a pivot language which is feasible for users.
- **Length** of interpretese versus translationese. Interpreting strategy involves shortening and summarization, so that the verbal production is easier for the interpreter as well as for the user to perceive. In average, interpreting is significantly shorter than translation. It can be beneficial also for speech translation users. We showed how speech translation can be designed for shortening to the same average length as interpreting.
- **Complexity**. We found out that interpretese contains simpler language than translationese. It can be therefore better perceivable for users.
- **Content preservation**. In a short manual analysis, we found out that more information from the source is lost via interpreting than via direct speech translation. It can be caused by removing redundancies or by instability and unreliability of human interpreters. The reason has to be analyzed.

The results of analysis might be useful for further research in multi-source speech translation that will actually combine the sources, as well as for event organizers for considering the source for speech translation.

5 Task T4.5 Flexible Multi-Lingual MT (CUNI, UEDIN, KIT)

Based on our experience from ELITR dry-run sessions and on the EUROSAI Congress, we found out that very important bottleneck in multi-lingual NMT is not the network capacity or mix of languages in many-to-many models, but rather in the area between designing, deploying and operating the complex end-to-end speech translation system with multiple source and many target languages. The system has to flexibly respond to situations when the sources or intermediate components are unreliable or produce noise.

The experience from EUROSAI congress and current state-of-the-art solution is described in deliverable D6.2: Report on ELITR at EUROSAI Congress and in Bojar et al. (2021). There are multiple parallel speech sources, the original in one language, and multiple simultaneous interpreting tracks, each into a different language. Our current speech translation system has to use only one source at a time. The preference for the source may change over time, even within the duration of the speech. The proposed solution in Bojar et al. (2021) relies on human operator who continuously monitors the available sources and keeps selecting one of them in real-time. Such manual monitoring and selection are however demanding and imperfect. We therefore aim to automate it, to enable the system to flexibly select the optimal source.

In CUNI, we therefore implemented a tool called "Auto Switcher". It implements rules that detect and disable empty and unreliable sources and order the remaining ones by preference. The rules take into account the underlying ASR and MT not working properly, e.g. due to large lagging, "hallucinations" of NN (a single subword or syllable repeated many times), missing or wrong punctuation or capitalization, not delivering output for some time, networking problems, unexpected language on the source, etc. Auto Switcher is fully integrated into the ELITR pipeline. It is extendable by custom rules depending on the specifics of the used session. The rules can use both text and audio features, e.g. noise detection or language identification from speech, or external ASR quality estimation.

References

Ebrahim Ansari, Ondřej Bojar, Barry Haddow, and Mohammad Mahmoudi. SLTev: Comprehensive Evaluation of Spoken Language Translation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Demo Papers*, Kyiv, Ukraine, April 2021. Association for Computational Linguistics.

Ondřej Bojar, Vojtěch Srdečný, Rishu Kumar, Otakar Smrž, Felix Schneider, Barry Haddow, Phil Williams, and Chiara Canton. Operating a complex SLT system with speakers and human interpreters. In *Proceedings of the 1st Workshop on Automatic Spoken Language Translation in Real-World Settings (ASLTRW)*, pages 23–34, Virtual, August 2021. Association for Machine Translation in the Americas. URL `https://aclanthology.org/2021.mtsummit-asltrw.3`.

Dominik Macháček, Matúš Žilinec, and Ondřej Bojar. Lost in Interpreting: Speech Translation from Source or Interpreter? In *Proc. Interspeech 2021*, pages 2376–2380, 2021. doi: 10.21437/Interspeech.2021-2232.

Santanu Pal, Hongfei Xu, Nico Herbig, Antonio Krüger, and Josef van Genabith. USAAR-DFKI – the transference architecture for English–German automatic post-editing. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, pages 124–131, Florence, Italy, August 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-5414. URL `https://aclanthology.org/W19-5414`.

Elena Voita, Rico Sennrich, and Ivan Titov. Context-aware monolingual repair for neural machine 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 876–885, Hong Kong, China, November 2019. Association for Computational Linguistics. URL `https://www.aclweb.org/anthology/D19-1081`.

Chaojun Wang, Christian Hardmeier, and Rico Sennrich. Exploring the importance of source text in automatic post-editing for context-aware machine translation. In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 326–335, Reykjavik, Iceland (Online), May 31–2 June 2021. Linköping University Electronic Press, Sweden. URL `https://aclanthology.org/2021.nodalida-main.34`.

A Exploring the Importance of Source Text in Automatic Post-Editing for Context-Aware Machine Translation

Exploring the Importance of Source Text in Automatic Post-Editing for Context-Aware Machine Translation

Chaojun Wang[1] **Christian Hardmeier**[2,3] **Rico Sennrich**[4,1]

[1]School of Informatics, University of Edinburgh
[2]Department of Computer Science, IT University of Copenhagen
[3]Department of Linguistics and Philology, Uppsala University
[4]Department of Computational Linguistics, University of Zurich
zippo_wang@foxmail.com, chrha@itu.dk, sennrich@cl.uzh.ch

Abstract

Accurate translation requires document-level information, which is ignored by sentence-level machine translation. Recent work has demonstrated that document-level consistency can be improved with automatic post-editing (APE) using only target-language (TL) information. We study an extended APE model that additionally integrates source context. A human evaluation of fluency and adequacy in English–Russian translation reveals that the model with access to source context significantly outperforms monolingual APE in terms of adequacy, an effect largely ignored by automatic evaluation metrics. Our results show that TL-only modelling increases fluency without improving adequacy, demonstrating the need for conditioning on source text for automatic post-editing. They also highlight blind spots in automatic methods for targeted evaluation and demonstrate the need for human assessment to evaluate document-level translation quality reliably.

1 Introduction

Neural machine translation (NMT) has significantly improved the state of the art in MT (Sutskever et al., 2014; Bahdanau et al., 2015; Vaswani et al., 2017) on the sentence level. However, accurate translation requires looking at larger units than individual sentences (Hardmeier, 2014), and context-aware NMT has recently become a popular research direction (Miculicich et al., 2018; Scherrer et al., 2019; Junczys-Dowmunt, 2019).

One approach to discourse-level processing in NMT is automatic post-editing of the output of a sentence-level system. DocRepair (Voita et al., 2019a) is a monolingual sequence-to-sequence model to correct inconsistencies in groups of adjacent sentence-level translations, showing improvements for specific discourse-level phenomena such as the generation of inflections in elliptic sentences.

The hypotheses explored in this work are (1) that the coherence of the translation can be further improved by exploiting context in the source language, and (2) that the omission of source context disproportionately affects adequacy in a way that is not measured adequately by the existing automatic evaluation procedures.

Our post-editing model is a document-level adaptation of Transference (Pal et al., 2019), a successful three-way transformer architecture from the WMT 2019 Automatic Post-Editing (APE) task (Chatterjee et al., 2019). To keep the model from over-correcting the hypothesis, we use data weighting (Junczys-Dowmunt, 2018) and a conservativeness penalty (Junczys-Dowmunt and Grundkiewicz, 2016). We evaluate on the same training and evaluation sets as Voita et al. (2019a), including a general test set validated by BLEU score and contrastive sets for several discourse phenomena.

Our experimental results confirm both hypotheses. Despite similar BLEU, human evaluation demonstrates that our Transference model significantly outperforms DocRepair in terms of adequacy, whilst both models show a comparable improvement in fluency over a baseline without APE. The automatic evaluation on discourse-specific test sets suggests that source-side information is particularly useful for predicting omitted verb phrases; however, even the targeted discourse-specific evaluation does not reflect the adequacy gain found by human evaluators. This is especially true since some of the discourse-specific test sets of Voita et al. (2019a) have a very narrow focus on problems for which source context is unlikely to help.

2 Transference

Transference (Pal et al., 2019) (Figure 1) is a multi-source transformer (Vaswani et al., 2017) architec-

ture which exploits both source *src* and the MT output *mt* to predict the reference *ref*. It is composed of (1) a source encoder (enc_{src}) to generate the *src* representation, (2) a second encoder ($enc_{src \to mt}$) which is a standard transformer decoder architecture without mask to produce the representation of *mt* incorporating *src* information, and (3) a decoder (dec_{ref}) which captures the final representation from $enc_{src \to mt}$ via cross-attention.

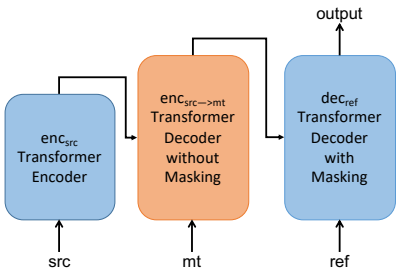

Figure 1: Transference architecture for multi-source document-level repair model.

If document-level APE is trained on a small subset of the parallel data, or only synthetic data, and therefore presumably weaker as a general model of translation than the sentence-level main model, we need to control how aggressively APE can modify *mt* to prevent over-correction. We adopt two strategies from the APE literature to achieve this. A *conservativeness penalty* (Junczys-Dowmunt and Grundkiewicz, 2016), denoted c, penalises the score of each prediction that is not in *src* or *mt*. Formally, let $V_c = V_{src} \cup V_{mt}$ be the subset of the full vocabulary V that occurs in an input segment. Given a $|V|$-sized vector of candidates h_t at time step t, the score of each candidate v is defined as:

$$h_t(v) = \begin{cases} h_t(v) - c & \text{if } v \in V \backslash V_c \\ h_t(v) & \text{otherwise.} \end{cases} \quad (1)$$

Second, similar to Lopes et al. (2019), we apply a *data weighting strategy* during training. We assign each training sample a weight that is defined as $\text{BLEU}_{\text{smooth}}(mt, ref)$ (Lin and Och, 2004) to upweight samples that require little post-editing.

3 Data and Preprocessing

We use all of the English-to-Russian data released by Voita et al. (2019a)[1], including: (1) 6M context-agnostic and 1.5M context-aware (4 consecutive sentences in each sample) data from the OpenSubtitles2018 corpus (Lison et al., 2018); (2) Russian monolingual data in 30M groups of 4 consecutive sentences gathered by Voita et al. (2019a). We reuse the synthetic training data for APE generated by Voita et al. (2019a), treating Russian monolingual data as *ref*, a sentence-level English back-translation as *src*, and the Russian roundtrip translation as *mt*. The evaluation data consists of general test sets extracted from the training data and four contrastive test sets to evaluate specific contextual phenomena.

[1] https://github.com/lena-voita/good-translation-wrong-in-context

Model	Deixis	Lex.c.	Ell.infl.	Ell.VP	BLEU
Results reported by Voita et al. (2019a):					
Baseline	50.0	45.9	53.0	28.4	32.41
DocRepair	91.8	80.6	86.4	75.2	34.60
Our experiments:					
DocRepair	88.6	70.5	83.8	69.0	32.69
DocRepair (+P)	87.6	67.6	82.2	71.8	32.38
Transference	86.8	62.9	81.6	73.0	30.56
Transference (+P)	87.8	65.4	84.8	82.8	32.53

Experiments marked +P use the ParData corpus.

Table 1: BLEU score on general test set and accuracy on contrastive test sets (deixis, lexical consistency, ellipsis (inflection), and VP ellipsis).

The four contrastive test sets have a narrow focus on specific discourse-level phenomena. The "Deixis" set targets consistent use of formal and informal second-person pronouns (T-V distinction) in Russian (however without regard to the social acceptability of the selected form). "Lexical cohesion" targets the consistent transliteration of proper names into Cyrillic script. These two sets are independent of source context by design, as the model is only evaluated on the generation of consistent repetitions of a form it has committed to, regardless of its adequacy in the context. The "Ellipsis VP" set targets elliptic verb phrases, where Russian requires the production of a lexical verb form not found in English. The "Ellipsis inflection" set tests the generation of noun inflections in sentences where the governing verb has been elided.

The training data is tokenised and truecased with Moses (Koehn et al., 2007), and encoded using byte-pair encoding (Sennrich et al., 2016b) with source and target vocabularies of 32000 tokens. Like Voita et al. (2019a), we report lowercased, tokenised BLEU (Papineni et al., 2002) with *multi-bleu.perl* from the Moses toolkit.

4 Model

The sentence-level baselines (EN→RU) and model used for RU→EN back-translation are Transformer base models (Vaswani et al., 2017).

For document-level APE, DocRepair is a Transformer base model that operates on groups of adjacent sentences, mapping from *mt* to *ref*. We use the Nematus toolkit (Sennrich et al., 2017) for DocRepair and our implementation of the Transference architecture, using the same configuration as Pal et al. (2019).[2] Detailed hyperparameters are listed in Appendix A. We train our document-level models on the 30M pairs of synthetic data. For some models, we also include the subset of the parallel data (1.5M pairs) for which context sentences are available, referred to as *ParData*. The *mt* part of *ParData* is generated by randomly sampling 20 translations with our EN→RU baseline system.

In preliminary experiments, adding noise to the training data improved model generalisation. We generated noise with two strategies. Following Voita et al. (2019a), *mt* in both synthetic data and *ParData* is randomly selected from 20 translations, and noise is added by making random token substitutions with probability of 10%. Following Edunov et al. (2018), noise is added to the *src* in synthetic data by three operations: (1) replacing a token; (2) deleting a token; (3) swapping adjacent token pairs, with a probability of 10%.

5 Automatic evaluation

Table 1 shows the results in terms of accuracy on the contrastive test sets and BLEU on the general test set. For DocRepair, we were unable to replicate the exact results of Voita et al. (2019a). Our conclusions are based on our own implementation.

On the general test set, trained on only synthetic training data, Transference achieves about 2 BLEU points less than DocRepair. We suspect that this derives from the mismatch of the training and test data for Transference. Specifically, during training, the "source" seen by Transference is the result of noisy back-translation from Russian, whereas at test time, the source is an original English sentence. When *ParData* is included, Transference and DocRepair achieve comparable BLEU.

In accuracy on the test sets for T/V pronouns ("deixis") and transliteration consistency ("lexical cohesion"), Transference does not improve over DocRepair, which is unsurprising considering how those test sets are constructed. However, adding source knowledge does improve results on both ellipsis test sets, for VP ellipsis even without adding the *ParData* data. The improvement is generally greater for VP ellipsis than for noun inflection.

[2] Code available at `https://github.com/zippotju/Context-Aware-Bilingual-Repair-for-Neural-Machine-Translation`

6 Human evaluation

To gain a better picture of the merits of the different systems, we conducted a manual evaluation. We randomly selected 720 sentences from the general test set and 100 sentences from the discourse test set and had them evaluated separately for adequacy and fluency by two native speakers of Russian. To avoid priming between the fluency and adequacy conditions, the test set was split between the annotators, and no sentence was annotated for adequacy and fluency by the same annotator. To determine the inter-annotator agreement, there are 100 overlapping sentences for two annotators. Table 5 shows inter-annotator agreement results while Table 4 shows the intra-annotator agreement. According to Landis and Koch (1977), all groups of human evaluation results are fair ($\kappa > 0.2$).

The sentences were presented to the annotators in random order along with 3 sentences of preceding context. The sentence to be evaluated was highlighted, and the Russian translations of the three systems (Baseline, DocRepair (+*ParData*) and Transference (+*ParData*)) were displayed next to each other, ordered randomly. In the adequacy condition only, the English source text was also shown. The annotators received instructions according to Table 2 and were told to assign the same rank if two translations were of equal quality. Once the annotation was complete, the rankings were converted into pairwise comparisons. Duplicate assessments from the inter- and intra-annotator sets were counted once if their annotations agreed, and discarded if they disagreed.

Table 3 shows the outcome of pairwise comparisons between the systems, including the number of times the output of one system was preferred over that of the other by the annotator. The results were tested for significance with a sign test. We find the same pattern of results for both test sets. In the *Fluency* evaluation, both monolingual DocRepair and bilingual Transference significantly improve over the Baseline. The comparison between DocRepair and Transference is not significant in this condi-

Adequacy: Please rank the three translations according to *how adequately the translation of the last sentence reflects the meaning of the source, given the context.*

Fluency: Please rank the three translations according to *how fluent the last sentence is, in terms of grammaticality, naturalness and consistency, taking into account the context of the previous sentences.*

Table 2: Instructions to human annotators

		Preference			
System A	*System B*	*A*	*B*	*Ties*	
Fluency					
General corpus:					
Baseline	**DocRepair**	30 < 62		612	($p < 0.005$)
Baseline	**Transference**	51 < 89		547	($p < 0.005$)
DocRepair	Transference	70	78	542	(n. s.)
Discourse corpus:					
Baseline	**DocRepair**	12 < 28		138	($p < 0.05$)
Baseline	**Transference**	15 < 34		120	($p < 0.01$)
DocRepair	Transference	23	25	121	(n. s.)
Adequacy					
General corpus:					
Baseline	DocRepair	24	31	655	(n. s.)
Baseline	**Transference**	34 < 67		592	($p < 0.005$)
DocRepair	**Transference**	39 < 66		592	($p < 0.05$)
Discourse corpus:					
Baseline	DocRepair	16	20	140	(n. s.)
Baseline	**Transference**	9 < 46		117	($p < 0.001$)
DocRepair	**Transference**	11 < 43		117	($p < 0.001$)

n. s. = not significant
Significance threshold: $p < 0.05$

Table 3: Human evaluation results. Winning systems in pairwise comparisons marked in bold.

tion. In the *Adequacy* evaluation, the comparison between DocRepair and the Baseline is not significant, but Transference significantly outperforms both DocRepair and the Baseline, demonstrating that knowledge of the source is essential for APE to improve the accuracy of the translations.

One of the evaluators provided qualitative comments on 32 pairs of DocRepair and Transference outputs sampled from those sentences for which the two systems were ranked differently in the human evaluation. The comments show that both

Per annotator:		
Annotator 1		91.1%
Annotator 2		83.9%
Per dataset:		
Fluency	General	90.0%
Fluency	Discourse	86.7%
Adequacy	General	90.0%
Adequacy	Discourse	78.3%

Table 4: Intra-annotator agreement of human evaluation

		κ	Pct.
Fluency	General	0.234	5
Fluency	Discourse	0.352	55
Adequacy	General	0.301	27
Adequacy	Discourse	0.471	93

Table 5: Inter-annotator agreement in terms of Cohen's κ (Cohen, 1960). The last column shows the percentile of our κ value in the context of a series of similar evaluations carried out at WMT 2012–2016 (Bojar et al., 2016, Table 4).

systems tend to produce imperfect output for the same sentences, but the winning system often manages to fix errors partially. Both systems make a wide range of errors in terms of morphology and lexical choice, but the source information permits Transference to correct certain recurring problems more reliably, such as agreement errors, mistranslations of proper names (e.g., Lena as Sarah), or the incorrect use or omission of subjunctive mood in conditional sentences.

7 Related Work

Our work draws on two strands of research: automatic post-editing and context-aware MT.

Automatic post-editing has a long history in MT (Knight and Chander, 1994), with regular shared tasks (Bojar et al., 2015, 2016, 2017). Neural multi-source APE systems as first proposed by Pal et al. (2016) and Junczys-Dowmunt and Grundkiewicz (2016), some of them including source language information (Junczys-Dowmunt and Grundkiewicz, 2017; Chatterjee et al., 2017; Libovický and Helcl, 2017), have come to dominate APE. We take inspiration from the top-performing systems at the WMT19 shared task for architectures and training/decoding tricks (Chatterjee et al., 2019), and make heavy use of synthetic training data (Sennrich et al., 2016a; Junczys-Dowmunt and Grundkiewicz, 2016; Freitag et al., 2019).

Neural context-aware MT can be achieved by integrating context into the main translation model (Jean et al., 2017; Tiedemann and Scherrer, 2017; Bawden et al., 2018, inter alia). Two-stage models with a sentence-level first pass and document-level second pass have been explored for scenarios with asymmetric training data. Voita et al. (2019b) introduces a two-pass model where, unlike in APE, the second-pass is tightly integrated with the first-pass model, reusing its hidden representations. Apart

from Voita et al. (2019a), the model closest to ours is by Junczys-Dowmunt (2019), who explored document-level APE, but only manually evaluated its efficacy as part of a large model ensemble.

8 Conclusion

Our human evaluation shows that monolingual APE oriented towards consistency beyond the sentence level improves fluency, but not adequacy, while multi-source APE with source context improves both adequacy and fluency. This shortcoming of monolingual APE in terms of adequacy was not easily visible with a consistency-focused automatic evaluation, highlighting the need for human evaluation to avoid such blind spots and reinforcing earlier findings about the inadequacy of automatic evaluation methods for discourse-level MT (Guillou and Hardmeier, 2018).

Clearly, a two-stage process with sentence-level translation and multi-sentence APE is a viable approach in asymmetric data settings with little document-level parallel data. However, we still required some actual document-level parallel data, and were unable to match the success of monolingual repair when using only synthetic data. Exploring the data requirements of document-level APE, and devising ways to reduce them, are worth further study.

Acknowledgments

Chaojun Wang was supported by the UK Engineering and Physical Sciences Research Council (EPSRC) fellowship grant EP/S001271/1 (MTStretch). Christian Hardmeier was supported by the Swedish Research Council under grant 2017-930. This project has received funding from the European Union's Horizon 2020 research and innovation programme (ELITR, grant agreement no 825460), and the Royal Society (NAF\R1\180122).

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Rachel Bawden, Rico Sennrich, Alexandra Birch, and Barry Haddow. 2018. Evaluating Discourse Phenomena in Neural Machine Translation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1304–1313, New Orleans, USA. Association for Computational Linguistics.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Barry Haddow, Matthias Huck, Chris Hokamp, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Carolina Scarton, Lucia Specia, and Marco Turchi. 2015. Findings of the 2015 workshop on statistical machine translation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 1–46, Lisbon, Portugal. Association for Computational Linguistics.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Shujian Huang, Matthias Huck, Philipp Koehn, Qun Liu, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Raphael Rubino, Lucia Specia, and Marco Turchi. 2017. Findings of the 2017 conference on machine translation (wmt17). In *Proceedings of the Second Conference on Machine Translation, Volume 2: Shared Task Papers*, pages 169–214, Copenhagen, Denmark. Association for Computational Linguistics.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Aurelie Neveol, Mariana Neves, Martin Popel, Matt Post, Raphael Rubino, Carolina Scarton, Lucia Specia, Marco Turchi, Karin Verspoor, and Marcos Zampieri. 2016. Findings of the 2016 conference on machine translation. In *Proceedings of the First Conference on Machine Translation*, pages 131–198, Berlin, Germany. Association for Computational Linguistics.

Rajen Chatterjee, M. Amin Farajian, Matteo Negri, Marco Turchi, Ankit Srivastava, and Santanu Pal. 2017. Multi-source neural automatic post-editing: FBK's participation in the WMT 2017 APE shared task. In *Proceedings of the Second Conference on Machine Translation*, pages 630–638, Copenhagen, Denmark. Association for Computational Linguistics.

Rajen Chatterjee, Christian Federmann, Matteo Negri, and Marco Turchi. 2019. Findings of the WMT 2019 shared task on automatic post-editing. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, pages 11–28, Florence, Italy. Association for Computational Linguistics.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. Understanding back-translation at scale. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 489–500, Brussels, Belgium. Association for Computational Linguistics.

Markus Freitag, Isaac Caswell, and Scott Roy. 2019. Ape at scale and its implications on mt evaluation biases. In *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, pages 34–44, Florence, Italy. Association for Computational Linguistics.

Liane Guillou and Christian Hardmeier. 2018. Automatic reference-based evaluation of pronoun translation misses the point. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4797–4802, Brussels, Belgium. Association for Computational Linguistics.

Christian Hardmeier. 2014. *Discourse in Statistical Machine Translation*, volume 15 of *Studia Linguistica Upsaliensia*. Acta Universitatis Upsaliensis, Uppsala.

Sebastien Jean, Stanislas Lauly, Orhan Firat, and Kyunghyun Cho. 2017. Does Neural Machine Translation Benefit from Larger Context? In *arXiv:1704.05135*. ArXiv: 1704.05135.

Marcin Junczys-Dowmunt. 2018. Microsoft's submission to the WMT2018 news translation task: How I learned to stop worrying and love the data. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 425–430, Belgium, Brussels. Association for Computational Linguistics.

Marcin Junczys-Dowmunt. 2019. Microsoft translator at WMT 2019: Towards large-scale document-level neural machine translation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 225–233, Florence, Italy. Association for Computational Linguistics.

Marcin Junczys-Dowmunt and Roman Grundkiewicz. 2016. Log-linear combinations of monolingual and bilingual neural machine translation models for automatic post-editing. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 751–758, Berlin, Germany. Association for Computational Linguistics.

Marcin Junczys-Dowmunt and Roman Grundkiewicz. 2017. An exploration of neural sequence-to-sequence architectures for automatic post-editing. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 120–129, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Kevin Knight and Ishwar Chander. 1994. Automated postediting of documents. In *AAAI*, volume 94, pages 779–784.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. 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*, pages 177–180, Prague, Czech Republic. Association for Computational Linguistics.

J. Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.

Jindřich Libovický and Jindřich Helcl. 2017. Attention strategies for multi-source sequence-to-sequence learning. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 196–202, Vancouver, Canada. Association for Computational Linguistics.

Chin-Yew Lin and Franz Josef Och. 2004. ORANGE: a method for evaluating automatic evaluation metrics for machine translation. In *COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics*, pages 501–507, Geneva, Switzerland. COLING.

Pierre Lison, Jörg Tiedemann, and Milen Kouylekov. 2018. OpenSubtitles2018: Statistical rescoring of sentence alignments in large, noisy parallel corpora. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

António V. Lopes, M. Amin Farajian, Gonçalo M. Correia, Jonay Trénous, and André F. T. Martins. 2019. Unbabel's submission to the WMT2019 APE shared task: BERT-based encoder-decoder for automatic post-editing. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, pages 118–123, Florence, Italy. Association for Computational Linguistics.

Lesly Miculicich, Dhananjay Ram, Nikolaos Pappas, and James Henderson. 2018. Document-level neural machine translation with hierarchical attention networks. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2947–2954, Brussels, Belgium. Association for Computational Linguistics.

Santanu Pal, Sudip Kumar Naskar, Mihaela Vela, and Josef van Genabith. 2016. A neural network based approach to automatic post-editing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 281–286, Berlin, Germany. Association for Computational Linguistics.

Santanu Pal, Hongfei Xu, Nico Herbig, Antonio Krüger, and Josef van Genabith. 2019. USAAR-DFKI – the transference architecture for English–German automatic post-editing. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, pages 124–131, Florence, Italy. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Yves Scherrer, Jörg Tiedemann, and Sharid Loáiciga. 2019. Analysing concatenation approaches to document-level NMT in two different domains. In *Proceedings of the Fourth Workshop on Discourse in Machine Translation (DiscoMT 2019)*, pages 51–61, Hong Kong, China. Association for Computational Linguistics.

Rico Sennrich, Orhan Firat, Kyunghyun Cho, Alexandra Birch, Barry Haddow, Julian Hitschler, Marcin Junczys-Dowmunt, Samuel Läubli, Antonio Valerio Miceli Barone, Jozef Mokry, and Maria Nădejde. 2017. Nematus: a toolkit for neural machine translation. In *Proceedings of the Software Demonstrations of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, pages 65–68, Valencia, Spain. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016a. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016b. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks.

In *Advances in Neural Information Processing Systems 27*, pages 3104–3112. Curran Associates, Inc.

Jörg Tiedemann and Yves Scherrer. 2017. Neural Machine Translation with Extended Context. In *Proceedings of the Third Workshop on Discourse in Machine Translation*, DISCOMT'17, pages 82–92, Copenhagen, Denmark. Association for Computational Linguistics.

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 *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.

Elena Voita, Rico Sennrich, and Ivan Titov. 2019a. Context-aware monolingual repair for neural machine 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 877–886, Hong Kong, China. Association for Computational Linguistics.

Elena Voita, Rico Sennrich, and Ivan Titov. 2019b. When a good translation is wrong in context: Context-aware machine translation improves on deixis, ellipsis, and lexical cohesion. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1198–1212, Florence, Italy. Association for Computational Linguistics.

A Appendix

A.1 Hyperparameter Search and Validation Performance

The following hyperparameters were manually tuned:

- The percentage of *ParData* mixed with the synthetic training data. of Transference.
- The conservativeness penalty.
- The decision whether to add the conservativeness penalty to the probability estimates or to the logits of the model.

The tuning bounds are shown in Table 7 in curly braces for each tuned hyperparameter. After 18 hyperparameter search trials, the best-performing models were selected considering both BLEU score on the general validation set and the accuracy on the contrastive validation sets. The validation results are shown in Table 6, and the hyperparameter configurations in Table 7.

Model	Deixis	Lex.c.	CE.loss	BLEU
DocRepair	89.0	68.0	58.2	32.01
DocRepair (+ParData)	88.8	68.8	56.3	31.63
Transference	86.0	62.2	61.0	30.37
Transference (+ParData)	85.4	64.8	50.7	31.99

Table 6: Validation performance of tested systems (CE represents Cross Entropy).

A.2 Training Time and Model Size

The two sentence-level baselines and the DocRepair model have approximately 72 million parameters each. The baseline systems are trained for around 72 hours each on a GeForce GTX 1080 Ti GPU. DocRepair and DocRepair (+*ParData*) are trained for approximately 216 hours on four TITAN X (Pascal) GPUs and 192 hours on a GeForce RTX 2080 Ti GPU, respectively.

The Transference model has around 119 million parameters. Transference and Transference (+*ParData*) were trained for around 192 and 288 hours, respectively, on three GeForce GTX 1080 Ti GPUs.

	DocRepair	Transference	Tuning bounds
Common hyperparameters			
Embedding layer size	512		
Hidden state size	512		
Tied encoder/decoder embeddings	yes	no	
Tie decoder embeddings	yes		
Loss function	per-token cross-entropy		
Label smoothing	0.1		
Optimizer	Adam		
Learning schedule	Transformer		
Warmup steps	8000		
Gradient clipping threshold	1.0		
Maximum sequence length	500		
Token batch size	15000		
Length normalization alpha	0.6		
Encoder depth	6		
Decoder depth	6		
Feed forward num hidden	2048		
Number of attention heads	8		
Embedding dropout	0.1		
Residual dropout	0.1		
ReLU dropout	0.1		
Attention weights dropout	0.1		
Beam size	4		
Percentage of ParData in training	0.3		$\{0.2, 0.3, 0.4\}$
Transference-specific hyperparameters			
Tied second encoder/decoder embeddings	yes		
Second encoder depth	6		
Conservativeness penalty	(0.2, probability)		$\{0.1, 0.2, 0.3\} \times \{\text{probability}, \text{logit}\}$

Table 7: Hyperparameter configurations for best-performing DocRepair and Transference models, and hyperparameter tuning bounds.

B Comprehension of Subtitles from Re-Translating Simultaneous Speech Translation

ACL-IJCNLP 2021 Submission ***. Confidential Review Copy. DO NOT DISTRIBUTE.

Comprehension of Subtitles from Re-Translating Simultaneous Speech Translation

Anonymous ACL-IJCNLP submission

Abstract

In simultaneous speech translation, one can vary the size of the output window, system latency and sometimes the allowed level of rewriting. The effect of these properties on readability and comprehensibility has not been tested with modern neural translation systems. In this work, we propose an evaluation method and investigate the effects on comprehension and user preferences. It is a pilot study with 14 users on 2 hours of German documentaries or speeches with online translations into Czech. We collect continuous feedback and answers on factual questions. Our results show that the subtitling layout or flicker have a little effect on comprehension, in contrast to machine translation itself and individual competence. Other results show that users with a limited knowledge of the source language have different preferences to stability and latency than the users with zero knowledge. The results are statistically insignificant, however, we show that our method works and can be reproduced in larger volume.

1 Introduction

Simultaneous speech translation is a technology that assists users to understand and follow a speech in a foreign language in real-time. The users may need such an assistance because of limited knowledge of the source language, the speaker's non-native accent, or the topic and vocabulary. The technology can be used for the target languages, for which human interpretation is unavailable, e.g. due to capacity reasons.

The candidate systems for simultaneous speech translation differ in quality of translation, latency and the approach to stability. Some are streaming, only adding more words (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), some allow re-translation as more input arrives (Müller et al., 2016b; Niehues et al., 2016; Dessloch et al., 2018; Niehues et al., 2018; Arivazhagan et al., 2020). Finally, subtitle presentation options (size of subtitling window, layout, allowed reading time, font size, etc.) also affect users' impression. The re-translating speech-to-text translation systems can offer lower latency by producing partial text hypotheses, which are however often withdrawn and replaced by new, more accurate versions. The combination of the re-translating approach and limited space for subtitles is challenging because of "flicker" by which we mean all the re-translations of the text that a user is reading at the moment, has already read, or that has been scrolled away. In this case, the subtitling options impact the reading comfort and delay and may affect the general usability.

The evaluation of the traditional, text-to-text machine translation (MT) has been researched for many years (see e.g. Han, 2018 or developments and discussion within the series of WMT, Barrault et al., 2020). It targets only the translation quality.

Simultaneous speech translation evaluation faces new challenges: simultaneity, latency, and readability to humans. Evaluating only selected aspects in isolation is reasonable (as quality in Elbayad et al., 2020), however, a complete evaluation must be end-to-end, from sound acquisition to subtitling and testing whether the users received the information.

We propose a method for human evaluation of simultaneous translation on simulated live events. We focus on the evaluation of subtitling layouts and measuring comprehension effectively. We demonstrate our method on 14 users and 15 video or audio documents (115 minutes in total) in German with one online translation system into Czech. We collect the users' feedback on the quality of subtitles during watching, and ask them to answer questions on information from the video to measure their

comprehension.

We have no prior estimate on the statistical significance of results with the limited number of participants and documents. In this pilot study, we test the significance and give the estimate for further, more extensive studies.

Our results showed that our speech translation system preserves on average 80% of information from the source, when used in offline mode, i.e. when the user has unlimited time to browse the translation. An average single person is able to find around 33% of information in online mode. Next, we found an optimal subtitling layout, and found that its difference from a suboptimal, but reasonable layout is small and insignificant. Finally, we tested if the evaluation can be simplified by using judges with a knowledge of the source language without comprehension questionnaires.

2 Related Work

Hamon et al. (2009) propose user evaluation of speech-to-speech simultaneous translation. To test the adequacy and intelligibility, they prepared questionnaires with factual questions from the source speech. The judges listened either to the interpreter, or the machine, and answered the questions. They evaluated the offline mode, the judges were allowed to stop and replay the audio while answering. This way the authors measured the comprehension loss caused by the automatic translation or interpretation. Each sample was processed by multiple judges, to eliminate human errors. Fluency was assessed by the judges on a scale.

Macháček and Bojar (2020) propose a technique for collecting continuous user rating while the user watches video and simultaneous subtitles. The user is asked to express the satisfaction with the subtitles at any moment by pressing one of four buttons as the rating changes.

Müller et al. (2016a) analyzed the feedback from foreign students using KIT Lecture Translator within two semesters. Such a long-term and informal evaluation differs considerably from judging in controlled conditions. On one hand, it summarizes the real-life situation with all the variables and corner cases that a lab test could only approximate or omit. On the other hand, the users may not be motivated to give the feedback, and can give only personal opinions that may be biased. This way it is also difficult to compare multiple system candidates.

3 Evaluation Campaign

In our evaluation, we simulate live events at which participants need assistance with understanding the spoken language. We prepared a web application presenting video or audio documents equipped with live subtitles. The judges see each document for their first time, only once, with source sound and without interruptions, to simulate the live setting. While watching, they press buttons to indicate their current satisfaction with the subtitles. Afterwards, they fill a questionnaire with comprehension and summary questions. We distribute different versions of subtitling setups among the judges for contrastive analysis.

The source and target languages in our study are German and Czech, respectively. This is an interesting example of two neighbouring countries, distinct language families and yet a relatively well studied pair with sufficient direct training data.

3.1 Translation System

We use the ASR system originally prepared for German lectures (Cho et al., 2013). It is a hybrid HMM-DNN model emitting partial hypotheses in real time, and correcting them as more context is available. The same system was used also by KIT Lecture Translator (Müller et al., 2016b).

The system is connected in a cascade with a tool for removing disfluencies and inserting punctuations (Cho et al., 2012), and with a German–Czech NMT system.

The machine translation is trained on 8M sentence pairs from Europarl and Open Subtitles (Koehn, 2005; Lison and Tiedemann, 2016), the only public parallel corpora of German and Czech, and validated on newstest. The Transformer-based (Vaswani et al., 2017) system runs in Marian (Junczys-Dowmunt et al., 2018) and reaches 18.8 cased BLEU on WMT newstest-2019.

Despite the translations are pre-recorded and only played back in our simulated setup, we ensured we keep the original timing as emitted by the online speech translation system.

3.2 Selection of Documents

We selected German videos or audio resources that fulfilled following conditions: 1) Length 5 to 10 minutes (with few exceptions). 2) The translations had to be of a sufficient quality. Based on a manual check, we discarded several candidate documents: a math lecture and broadcast news due

domain	type	docs.	duration	description
EP	TP	3	18:08	From European Parliament
DG	TP	3	17:34	From DG SCIC repository for interpretation training
Mock Int	A	3	27:52	From a mock interpreted conference at interpretation school
Maus	V	2	14:43	Educative videos for children
DW	A	2	18:48	Audio for intermediate learners of German
Dinge	V	2	16:09	Educative video for teenagers and grown-ups
All		15	114:52	

Table 1: Summary of domains of selected documents. "Type" distinguishes audio only (A), talking person only (TP) and video (V) with illustrative or informative content. Duration is reported in minutes and seconds.

to many mistranslated technical terms and named entities. Another group of documents was mistranslated and discarded because they were not long-form speeches, but isolated utterances with long pauses. 3) Informative content. We intend to measure adequacy and comprehension by asking the judges complementary questions. We thus excluded the documents where the speaker is not giving information by speech, but uses mostly paralinguistic means, e.g. singing, poetry, or non-verbal communication. 4) Non-technicality. We expect the judges answer in several plain words in their mother tongue. They may lack knowledge of any specialized vocabulary.

We selected audios, videos with informative or illustrative content, and videos of talking persons, to compare user feedback for these types of documents.

Table 1 summarizes the selected documents.

3.3 Questionnaires

We decided to use direct factual questions in our study, instead of yes/no questions to exclude guessing. We asked a Czech teacher of German to prepare the questions and an answer key from the original German documents, regardless of the machine translation. The teacher wrote the questions in Czech, and was instructed to prepare one question from every 30 seconds of the stream and distribute them evenly, if possible. The questions had to be answerable only after listening to the document, and not from the general knowledge. The complexity of the questions was targeted on the level that an ordinary high-school student could answer after listening to the source document once, if the student would not have any obstacles in understanding German. To reduce the effect of limited memory, the judges had an option in the questionnaire to indicate they knew the answer but forgot it. Furthermore, they had to fill, from which source they knew the answer: from the subtitles, from the speech, from an image on the video, or from their previous knowledge.

level	count	group	total
0	5	non-German speaking	10
A1	5		
A2	1	German speaking	4
B1	2		
B2	1		
All			14

Table 2: The judges by their German proficiency levels on CEFR scale and their distribution to groups.

After the factual questions, all the questionnaires had a common part where we asked the judges on their general impression of translation fluency, adequacy, stability and latency, overall quality, video watching comfort, and a summary comment. Each judge spent in total 2 hours on watching and 3 hours on the questionnaires.

Finally, we evaluated the factual questions manually against the key, rating them at three levels: correct, incorrect, and partially correct.

3.4 Judges

We selected 14 native Czech judges. Their self-reported knowledge of German had to be between zero and B2 on the CEFR[1] scale, to ensure they need some level of assistance with understanding German. We also ensured they do not have knowledge of any other language which could help them understanding German. The summary of their proficiency in German is in Table 2. For further analyses in our study, we divided them into two groups. For brevity further in the paper, we denote the 10 judges

[1]Common European Framework of Reference for Languages

Figure 1: A detail of the default layout with the video document "Dinge Erklärt: Impfen..." (https://youtu.be/4E0dwFS72gk). The video is at the top, below are two lines of subtitles in Czech, followed by buttons for the continuous quality rating. The button labels are "1 = worse", "2 = average", "3 = OK", "0 = I do not understand at all". The order 1, 2, 3, 0 matches the keyboard layout; users were encouraged to use keyboard shortcuts.

with zero or A1 level (beginners) as "non-German speaking", and the others as "German speaking". Because we have a small amount of German speaking ones, we do not classify them in more detail.

The judges were paid for participation in the study. They watched the videos at their homes on their own devices. They were asked to customize their screen resolution and eye-screen distance to suit their comfort.

3.5 Subtitler: Subtitle Presentation

The Subtitler is our implementation of the algorithm by Macháček and Bojar (2020) extended with automatic adaptive reading speed in addition to the "flicker" parameter as defined in the paper. The speed varies between 10 and 25 characters per second depending on the current size of the incoming buffer. The default font size is 4.8 mm. The default subtitling window is 2 lines high and 163 mm wide. By default, we use the maximum flicker and the lowest delay (presenting all translation hypotheses, not filtering out the partial and possibly unstable ones), no colour highlighting, and smooth slide-up animation while scrolling. The example of the setup can be seen in Figure 1.

With the default subtitling window, 90% of the words in the test documents are finalized in subtitles at most 3 seconds after translation. In 99%, it is at most 7 seconds.

Type	w. avg±std	t-test
Offline+voting	0.81±0.11	
Offline	0.59±0.16	***
Online, without flicker	0.36±0.16	***
Online, flicker, top layout	0.33±0.13	
Online, flicker, least preferred	0.31±0.16	

Table 3: Comprehension scores on all documents and judges. The average weighted by number of questions in document. *** denote the statistically significant difference (p-value<0.01) between the current and previous line.

4 Results

4.1 Comprehension

In our study, we assume that comprehension can be assessed as a proportion of correctly answered questions. We assume the following model: A person without any language barrier and with non-restricted access to the document during answering the questionnaire can answer all questions correctly. With a language barrier and offline machine translation (unlimited perusal of the document while answering), some information may be lost in machine translation. More information is lost with one-shot access to online machine translation because of forgetting and temporal inattention. Some more information may be lost because of flicker, and some more because of suboptimal subtitling layout.

Our results confirmed the assumed hierarchy of comprehension levels. Moreover, we noticed that even the judges with offline MT gave inconsistent answers. When we combined them and counted as correct if at least one was correct, they achieved higher scores. We explain it by insufficient attention.

Table 3 summarizes the results on all documents. We measured that on average, 81% of information was preserved by machine translation (Offline+voting, i.e. one of two judges answered correctly). A single judge could find 59% of information (Offline). In an oracle experiment without flicker, when the machine translation gives the final hypotheses with the timing of the partial ones (i.e. as if it knew the best translation of the upcoming sentence), a single judge could answer 36%. In real setup with flicker and the most preferred subtitling layout (Online, flicker, top layout), 33% information was found, and 31% with less preferred. The standard deviation is between 11 and 16%.

We found statistically significant difference (two-sided t-test) between offline MT with voting and

	German≥A2 #	avg±std	German<A2 #	avg±std	t-test
flicker	3	**0.59±0.15**	10	0.30±0.15	$p < 0.05$
no flicker	4	0.40±0.06	10	**0.34±0.07**	insig.
t-**test**	$p < 0.10522$		insignificant		

Table 4: Comprehension scores on two documents on a setup with and without flicker, as rated by judges whose German competence is between A2 and B2 on CEFR scale (elementary to upper intermediate), or below A2 (zero or beginner). Number of samples is denoted as "#", higher scores bolded.

without it, and between offline MT and online. The difference caused by flicker or layout was insignificant.

4.2 Preferences by Language Skills

We assume that the user behaviour differs by knowledge of the source language. The users with zero knowledge read all subtitles all the time and do not pay attention to the speech. They do not mind large latency, but demand high quality translation, and comfortable reading without flicker. On the other hand, the users with a limited, but nonzero knowledge of the source language listen to the speech, try to understand on their own, and look at the subtitles only occasionally, when they are temporarily uncertain or need assistance with an unfamiliar word. They need low latency, and do not mind slightly lower quality.

To empirically test our hypothesis, we prepared two setups: With flicker, the subtitles are presented immediately as available, but with frequent rewriting, which discomforts the reader. For comparison without flicker, we present only the final translations without rewriting, but with a large latency. We selected two videos and distributed these setups uniformly between German speaking and non-German speaking judges.

The results of comprehension are in Table 4. It shows that German-speaking users achieve higher comprehension with flicker than without. We consider the difference as close to statistically significant (p-value < 0.10522), although we had only 4 and 10 German and non-German speaking judges, respectively. The non-German speakers understood better without flicker (34% vs 30%), but this difference is statistically insignificant. The other types of feedback (weighted average of continuous rating and the overall rating at the end of questionnaire) confirm the trend of comprehension, but have larger variance and the differences are insignificant.

		Side	Below
Final rating	audio	3 2.00 ±0.82	6 2.00 ±0.82
	talking	4 2.25 ±0.83	3 **2.67 ±0.94**
	video	1 1.00 ±0.00	1 1.00 ±0.00
	sum, avg	8 2.00 ±0.87	10 **2.10 ±0.94**
Comprehension	audio	3 **0.27 ±0.13**	6 0.21 ±0.13
	talking	4 0.22 ±0.12	3 **0.28 ±0.26**
	video	1 0.18 ±0.00	1 **0.33 ±0.00**
	sum, avg	8 0.23 ±0.12	10 **0.24 ±0.18**
Avg. cont. rating	audio	3 **1.18 ±0.76**	6 0.76 ±0.54
	talking	4 1.20 ±0.79	3 **1.76 ±0.47**
	video	1 0.23 ±0.00	1 **0.77 ±0.00**
	sum, avg	8 **1.07 ±0.79**	10 1.06 ±0.67
Watching comfort	talking	4 2.75 ±0.83	3 **3.00 ±0.82**
	video	1 2.00 ±0.00	1 **3.00 ±0.00**
	sum, avg	5 2.60 ±0.80	4 **3.00 ±0.71**

Table 5: Results of the contrastive experiments of the non-German speaking judges for side vs below layout. The three numbers in each row and cell are the number of experiments, average and standard deviation. The higher score, the better. Comprehension rate is between 0 and 1, average continuous rating is between 0 and 3, the others on a discrete scale 1 to 5. Higher score in each row bolded.

4.3 Subtitling Layout

We analyzed effects of distinct subtitling features by contrastive experiments differing only at one feature. We distributed them randomly among the judges, regardless of their German skills. We can draw conclusions only on non-German speaking judges due to insufficient number of observations for the German-speaking group.

In all cases, the results show a slight insignificant preference towards one variant of the feature in all three types of feedback (comprehension, weighted average of continuous rating, and overall rating at the end of video).

4.3.1 Side vs Below

For videos and videos with a talking person, we consider two locations for the subtitle window: on the left side of the video, or below. The side window can be high but narrow (17 lines of 60 mm width, to match the height of the video), while the window underneath is short and wide (2 lines of 163 mm width). The first is more comfortable for reading, the latter for watching video.

The results are in Table 5. "Final rating" and "Watching comfort" summarize the responses in the final section of the questionnaire, where judges answered on a discrete scale 1 (worst) to 5 (best). "Comprehension" and "Average continuous rating" are, as above, results from correctness of answers and from the feedback button clicks, resp. The

		Below	Overlay
Final rating	talking	9 2.33 ±1.05	9 **2.78 ±1.13**
	video	5 1.40 ±0.80	8 **2.38 ±0.86**
	sum, avg	14 2.00 ±1.07	17 **2.59 ±1.03**
Compre-hension	talking	9 0.29 ±0.25	9 **0.39 ±0.20**
	video	5 0.26 ±0.14	8 **0.37 ±0.11**
	sum, avg	14 0.28 ±0.21	17 **0.38 ±0.17**
Avg. cont. rating	talking	9 1.65 ±0.52	9 1.65 ±0.99
	video	5 1.11 ±0.50	8 **1.15 ±0.77**
	sum, avg	14 **1.47 ±0.57**	17 1.42 ±0.93
Watching comfort	talking	9 3.43 ±0.73	9 **4.11 ±0.74**
	video	5 2.20 ±1.60	8 **3.00 ±1.00**
	sum, avg	14 2.92 ±1.32	17 **3.59 ±1.03**

Table 6: Results of the experiments on "overlay" vs "below" layout, for non-German speaking judges. Description of numbers and ratings as in Table 5.

Size [lines,mm width]	18×250 ("Large")	
Highlighting	No	Yes
Final rating	14 2.93 ±0.80	13 **3.31 ±1.14**
Comprehension	14 0.25 ±0.15	13 **0.30 ±0.12**
Avg. cont. rating	14 1.32 ±0.82	13 **1.42 ±0.74**

Size [lines,mm width]	5×200 ("Medium")	
Highlighting	No	Yes
Final rating	2 2.50 ±0.50	1 **4.00 ±0.00**
Comprehension	2 **0.44 ±0.18**	1 0.39 ±0.00
Avg. cont. rating	2 **2.19 ±0.50**	1 2.12 ±0.00

Table 7: Results of highlighting experiments on audio documents. Description of numbers as in Table 5.

results show statistically insignificant difference in all measures. There is a slight overall preference for the layout "below", except audio-only documents.

4.3.2 Overlay vs Below

The subtitling window can be placed over the video, as in films, or below. In the first case, the subtitles possibly hide an informative image content, in the latter case, there is a larger distance between the image and the subtitles. The results on non-German speaking judges are insignificantly in favor of overlay, see Table 6.

4.3.3 Highlighting Flicker Status

The underlying rewriting speech translation system distinguishes three levels of status for segments (automatically identified sentences): "Finalized" segments means no further changes are possible. "Completed" segments are sentences which received a punctuation mark. They can be changed by a new update and the prediction of the punctuation may also change or disappear. They usually flicker once in several seconds. "Expected" segments are incomplete sentences, to which new translated words are still appended. They flicker several times per second.

Size [lines,mm width]		2×163	5×200
Final rating	audio	10 1.80 ±0.87	8 **2.75 ±0.97**
	talking	9 2.33 ±1.05	5 **2.80 ±1.60**
	video	5 1.40 ±0.80	3 **2.33 ±0.47**
	sum, avg	24 1.92 ±1.00	16 **2.69 ±1.16**
Compre-hension	audio	10 0.25 ±0.15	8 **0.31 ±0.15**
	talking	9 0.29 ±0.25	5 **0.40 ±0.21**
	video	5 0.26 ±0.14	3 **0.28 ±0.05**
	sum, avg	24 0.26 ±0.19	16 **0.33 ±0.16**
Avg. cont. rating	audio	10 0.90 ±0.71	8 **1.66 ±0.95**
	talking	9 **1.65 ±0.52**	5 1.09 ±0.78
	video	5 1.11 ±0.50	3 **1.35 ±0.31**
	sum, avg	22 1.21 ±0.70	16 **1.42 ±0.85**
Watching comfort	talking	7 **3.43 ±0.73**	5 2.80 ±0.98
	video	5 2.20 ±1.60	3 **2.33 ±1.25**
	sum, avg	12 **2.92 ±1.32**	8 2.62 ±1.11

Size [lines,mm width]		18×250	5×200
Final rating	audio	11 **2.91 ±0.79**	8 2.75 ±0.97
Comprehension	audio	11 0.23 ±0.14	8 **0.31 ±0.15**
Avg. cont. rat.	audio	11 1.50 ±0.79	8 **1.66 ±0.95**

Table 8: Results of the experiments with subtitling window. Descriptions as in Table 5.

It is a user interface question if the status of the segments should be indicated by highlighting, or if this piece of information would be rather disturbing. We experimented only with colouring text background in large and medium subtitling window for audio-only documents.

Our experiments show that the judges prefer highlighting flicker status in the large window. For the medium window, this inclination is less clear, see Table 7.

4.3.4 Size of Subtitling Window

The subtitling window can be of any size. If the window is short and narrow, there is a short gap between an image and subtitles, which simplifies focus switching. On the other hand, a small window contains short history, so the user can miss translation content if it disappears while paying attention to the video. A small window may also cause a long subtitling delay if the translation was updated in scrolled away part of text, so that Subtitler has to return and repeat it (a very disturbing "reset"). With a large window, there is a larger distance between the end of subtitles and the image. The content stays longer, but it is more complicated to find a place where the user stopped reading before the last focus switch.

Depending on spatial constraints, it is always recommended to use as large window as possible, especially for documents without visual information, where focus switching between an image and

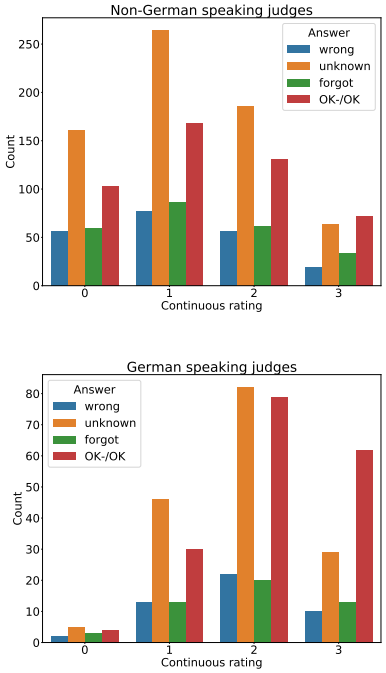

Figure 2: The distribution of the continuous rating and results of answers for non-German (upper) and German speaking (lower) judges.

subtitles is not expected. We tested two pairs of sizes on the same documents. The results are in Table 8. As we expected, the window with 5 lines was rated insignificantly better than with 2 lines, but the 2-line was more comfortable for watching. The judges rated it with average 2.92 in final section of the questionnaire, while the 5-line average was 2.62.

For an audio-only document, we also tested the large (18 lines) vs. medium (5 lines) window, observing users' reported preference for the large one but slightly higher comprehension and continuous feedback for the medium one, see the lower part of Table 8.

4.4 Relating Comprehension and Continuous Rating

We collected continuous rating of the overall quality of subtitles at given times, with four levels, where 0 means the worst and 3 the best. For every

	χ^2-test p-values			
answers	Non-G. sp. j.		Germ. sp. judges	
wrong	0.53	insig.	0.81	insig.
unknown	0.28	insig.	**0.09**	sign. $p < 0.1$
forgot	0.69	insig.	0.61	insig.
OK/OK-	0.12	insig.	**0.03**	sign. $p < 0.05$

Table 9: The results of χ^2-test for statistical significance of the independence of the distribution of continuous ratings and answer correctness.

comprehension question, we know the time when the necessary piece of information is uttered in the source speech document. Based on this timing information, we can relate comprehension and the reported continuous feedback. In Figure 2, we plot the number of Continuous rating button clicks divided according to whether the information at that time was understood acceptably ("OK/OK-"), spotted but forgotten ("forgot"), missed by the user ("unknown"), or misunderstood ("wrong"). This data aggregates observations for all documents and all setups excluding the offline MT and the oracle online MT without flicker.

We use the χ^2-test to measure whether the distribution of answer results and continuous rating are independent or not. The results are in Table 9. For the non-German speaking judges, the distributions are independent, while for the German speaking there is a statistically significant dependence between unknown answers and ratings, and correct answers and ratings. It means that if we know the ratings of the German speaking judges, we can predict their comprehension with a higher precision than without it. This observation could be used as the basis for a less time-consuming evaluation, e.g. when several translation systems need to be compared. Judges with elementary to upper intermediate knowledge of the source language could only watch the subtitles and provide continuous feedback, instead of the comprehension questions. The questions are laborious to both prepare and answer.

Forgetting and wrong answers are found to be independent on the continuous feedback. It is possible that the wrong answers are caused by inadequacies in the machine translation that non-German speakers can not observe, which are distributed uniformly regardless the flicker, latency or fluency.

From the χ^2 test results, we conclude that for the non-German speaking judges, their comprehen-

sion is probably independent of their continuous rating, because they have no competence for rating the adequacy. Their ratings are based only on fluency, readability and flicker. The German-speaking judges probably included the adequacy factor into the rating, which the non-German speakers could not do. This fact could be used in the future works. The judges could be used for comparison of multiple translation candidates. The judges who speak the source language could assess the adequacy only by the continuous rating without the need for questionnaires, which are laborious to prepare, answer and evaluate. The non-German speaking judges could skip the continuous rating and only fill out the questionnaire for adequacy.

5 Scalability

The evaluation method described in this paper requires manual work to select the documents, prepare, fill and evaluate the questionnaires. The amount of work is feasible in small number of documents and judges, but the results are insignificant. Re-scaling to large volumes may be costly. Therefore, in this section we propose ways to reduce the manual work in future evaluations.

It is advisable to target only on the documents, on which the speech translation achieves sufficient quality, because the users' impression will be equally bad with low-quality translations. The quality can be estimated by automatic MT metrics (e.g. BLEU, METEOR, etc.), if the reference translations are available.

We hypothesize that the questionnaires can be avoided, if future works confirm correlation of continuous rating of bilingual judges with adequacy. To measure the correlation and limits of significancy, experiments with large amounts of manual work are necessary, similarly as when finding the evidence for correlation of BLEU to human judgements (Reiter, 2018).

6 Conclusion

We proposed a method for end-to-end user evaluation of simultaneous speech translation, relying on users' continuous feedback and a follow-up questionnaire. The method can be used for measuring comprehension and evaluating subtitling parameters. We test the method in an evaluation campaign using 14 judges and 115 minutes of video and audio documents. Each of the judges spent 2 hours watching the documents and 3 hours answering the questionnaires. We observed that with the judges knowing the source language, it could be possible to omit the questionnaires because they seem to be able to assess adequacy in continuous rating.

The most preferred subtitling parameters are two lines of subtitles placed over the video, if the video has informative content. In case of video with a talking person or audio document, the most preferable is a large subtitling window with colour indication of whether the segment is final or still can change.

The users with a knowledge of the source language prefer low latency for sake of stability, while the users without language knowledge have no preference.

We did not find a statistically significant evidence on the impact of the differences in subtitling parameters to comprehension. We hypothesize that if the parameters are reasonable and do not cause a large delay, then the effect is close to zero. The largest effect on comprehension can be attributed to the individual competence and machine translation.

We successfully tested the method on limited number of participants and documents, and got statistically insignificant results. We conclude that our work may be used for an estimate of significance for further, more extensive studies.

References

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.

Loïc Barrault, Magdalena Biesialska, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Matthias Huck, Eric Joanis, Tom Kocmi, Philipp Koehn, Chi-kiu Lo, Nikola Ljubešić, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Santanu Pal, Matt Post, and Marcos Zampieri. 2020. Findings of the 2020 conference on machine translation (WMT20). In *Proceedings of the Fifth Conference on Machine Translation*, pages

1–55, Online. Association for Computational Linguistics.

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.

Eunah Cho, J. Niehues, and Alexander H. Waibel. 2012. Segmentation and punctuation prediction in speech language translation using a monolingual translation system. In *IWSLT*.

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.

Maha Elbayad, Michael Ustaszewski, Emmanuelle Esperança-Rodier, Francis Brunet-Manquat, Jakob Verbeek, and Laurent Besacier. 2020. Online versus offline NMT quality: An in-depth analysis on English-German and German-English. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5047–5058, Barcelona, Spain (Online). International Committee on Computational Linguistics.

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.

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.

Lifeng Han. 2018. Machine translation evaluation resources and methods: a survey. In *IPRC – Irish Postgraduate Research Conference*, Dublin, Ireland.

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.

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. 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 Ondřej Bojar. 2020. Presenting simultaneous translation in limited space. In *Proceedings of the 20th Conference Information Technologies – Applications and Theory (ITAT 2020), Hotel Tyrapol, Oravská Lesná, Slovakia, September 18-22, 2020*, volume 2718 of *CEUR Workshop Proceedings*, pages 34–39. CEUR-WS.org.

Markus Müller, Sarah Fünfer, Sebastian Stüker, and Alex Waibel. 2016a. Evaluation of the KIT lecture translation system. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1856–1861, Portorož, Slovenia. 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. 2016b. 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.

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 transla-

tion. 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.

Ofir Press and Noah A. Smith. 2018. You may not need attention. *CoRR*, abs/1810.13409.

Ehud Reiter. 2018. A structured review of the validity of BLEU. *Computational Linguistics*, 44(3):393–401.

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.

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

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.

C Lost in Interpreting: Speech Translation from Source or Interpreter?

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.