MockConf: A Student Interpretation Dataset: Analysis, Word- and Span-level Alignment and Baselines

Dávid Javorský[1] and **Ondřej Bojar**[1] and **François Yvon**[2]

[1]Charles University, Faculty of Mathematics and Physics, Prague, Czechia
[2]Sorbonne Université, CNRS, ISIR, Paris, France
{javorsky,bojar}@ufal.mff.cuni.cz francois.yvon@cnrs.fr

Abstract

In simultaneous interpreting, an interpreter renders a source speech into another language with a very short lag, much sooner than sentences are finished. In order to understand and later reproduce this dynamic and complex task automatically, we need dedicated datasets and tools for analysis, monitoring, and evaluation, such as parallel speech corpora, and tools for their automatic annotation. Existing parallel corpora of translated texts and associated alignment algorithms hardly fill this gap, as they fail to model long-range interactions between speech segments or specific types of divergences (e.g., shortening, simplification, functional generalization) between the original and interpreted speeches. In this work, we introduce *MockConf*, a student interpreting dataset that was collected from Mock Conferences run as part of the students' curriculum. This dataset contains 7 hours of recordings in 5 European languages, transcribed and aligned at the level of spans and words. We further implement and release *InterAlign*, a modern web-based annotation tool for parallel word and span annotations on long inputs, suitable for aligning simultaneous interpreting. We propose metrics for the evaluation and a baseline for automatic alignment. Dataset and tools are released to the community.

1 Introduction

Recent advances in speech and translation technologies offer new perspectives for the study of multilingual speech processing, a field that has its origins several decades ago (Waibel, 2004). This includes, for instance, the translation of speech transcripts for videos, to be used as captions in a video player, or the automatic generation of full-fledged subtitles for movies or TV shows. These processes have already been studied, and resources are available for a variety of genres and languages, enabling the development of automatic end-to-end subtitling systems (Rousseau et al., 2012; Cettolo et al., 2012; Lison and Tiedemann, 2016; Pryzant et al., 2018; Di Gangi et al., 2019; Karakanta et al., 2020). Other speech translation tasks have been considered, involving an increased level of interactivity, such as multilingual information systems (van den Heuvel et al., 2006), or translation tools for mediated conversations in various contexts, e.g. interactions between patients and doctors (Rayner, 2000; Ji et al., 2023) or military applications (Stallard et al., 2011). For these tasks, translations can happen in turns and the focus is often on the informational adequacy of the translated content.

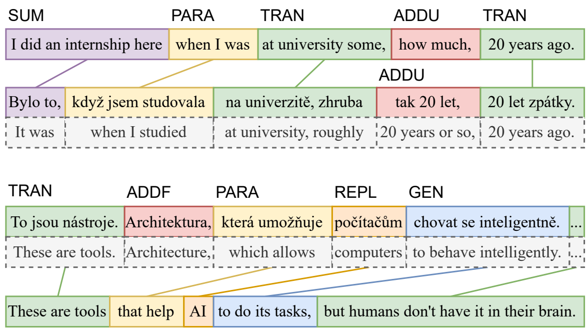

Figure 1: Examples of span-level annotation from our dataset. The first and second rows display transcripts of the original speech and its interpretation. The gray dashed row is the gloss of the Czech part. Span labels are displayed above the corresponding spans, see Table 2 for a description of labels.

In this study, we focus on another type of multilingual task: simultaneous interpreting.[1] This mode of interpretation typically occurs in international conferences, where a presenter's speech is immediately rendered into a foreign language. Simultaneous interpreting has been an active area of research, particularly thanks to resources derived from institutions such as the European Parliament

[1]Defined by Diriker (2015) as: "Broadly speaking, simultaneous interpreting (SI) is the mode of interpreting in which the interpreter renders the speech as it is being delivered by a speaker into another language with a minimal TIME LAG of a few seconds."

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 16339–16356
July 27 - August 1, 2025

(Macháček et al., 2021) and, more recently, ACL conferences (Agarwal et al., 2023).

Building on this research, we introduce *MockConf*, a dataset centered on Czech, comprising simultaneous interpreting data with human-annotated transcriptions at both the span and word levels. The dataset creation process involves several key steps: First, we obtain a faithful transcription of human simultaneous interpretings that were collected from Mock Conferences run as part of the student interpreters curriculum. These data was then manually aligned and annotated at the word and span level using *InterAlign*, a dedicated tool designed to facilitate the annotation at the span and word levels. Some example annotations are shown in Figure 1. Additionally, we propose a new automatic alignment task that aims to reproduce these manual alignments. In our experiments, we establish baselines and discuss the challenges associated with this task.

MockConf, serves multiple purposes. First, it offers valuable opportunities for linguistic analyses (Doi et al., 2024; Wein et al., 2024), some of which we have already explored. Second, span-level annotations are beneficial for the development and evaluation of automatic alignment tools. Alignments can aid in tasks such as detecting MT hallucinations (Pan et al., 2021; Guerreiro et al., 2023; Dale et al., 2023) or MQM evaluation using error span classification (Burchardt, 2013; Kocmi and Federmann, 2023; Li et al., 2025; Lu et al., 2025). *MockConf* can also be useful for educational purposes, e.g., to automatically monitor and analyze the productions of student interpreters, or to evaluate human interpreting (Stewart et al., 2018; Wein et al., 2024; Makinae et al., 2025). Finally, the dataset can contribute to the evaluation of automatic simultaneous interpreting systems (Wang et al., 2023). The *MockConf*[2] dataset with the analysis and baselines, and the *InterAlign*[3] annotation tool are publicly released to the community.

2 *MockConf*: A dataset of simultaneous interpreting

2.1 Recordings and data collection

The dataset was collected from Mock Conferences that took place as part of the interpreting curriculum at a university. During these conferences, a student plays the role of some celebrity and prepares a speech on some predefined topic. Students who are enrolled in Master's level studies listen to the speech and interpret it. The interpreters are familiar with the topic and are provided with a short description of the content. The languages covered are Czech, English, French, German, and Spanish and each direct interpreting is always from or into Czech. There are also *relay interpretings*, which are analogous to pivot translations: talks in foreign language are interpreted into Czech, from which they are further interpreted into other languages. All recordings have been automatically transcribed using WhisperX (Bain et al., 2023), then manually revised by native Czech speakers, with sufficient self-reported proficiency in the respective foreign language. Transcribers were asked to capture exactly what was said, even though utterances might contain disfluencies such as hesitations and false starts, or even translation errors. They also labeled spans containing proper names, which we will further use for anonymization purposes. The full transcription guidelines are in Appendix C.

	Language		Recordings		Token count	
split	src	trg	count	duration	# src	trg
dev	cs	de	2 0	00:21:08	2377	2187
	cs	en	0 6	01:06:56	7876	7001
	cs	es	0 1	00:11:20	1370	988
	cs	fr	1 0	00:20:07	1922	2196
all			3 7	01:59:31	13545	12372
test	cs	de	1 2	00:30:27	3211	2833
	cs	en	0 6	01:00:46	6819	6118
	cs	es	0 3	00:31:22	2873	2810
	cs	fr	3 0	00:29:29	3858	3789
	de	cs	2 0	00:21:14	2299	1840
	en	cs	0 5	01:02:27	9070	6395
	es	cs	0 2	00:19:19	2360	1837
	fr	cs	4 1	00:46:12	7229	4791
all			10 19	05:01:16	37719	30413

Table 1: Main statistics of *MockConf*. We identify languages with ISO-632-2 codes. The values in the "count" cell denote the number of recordings with consent to publish only transcripts or both transcripts and audio, respectively. Tokens are obtained using Moses tokenizer.[4]

Consent to publish We asked each participant for their consent to redistribute their recordings and ended up with around 7 hours of recordings for which we obtained consent from the two participants (speaker and interpreter), which we split into development and test set with a 1:3 ratio. Note that development set is limited to only cs→xx direction

[2] https://github.com/J4VORSKY/MockConf

[3] https://github.com/J4VORSKY/InterAlign

[4] https://pypi.org/project/mosestokenizer/

and does not proportionally represent all annotators. We assume that evaluating on such data might lead to a better generalization. Participants were allowed to choose between: no consent (excluded from the data), partial consent (to publish the transcripts) and full consent (to publish transcripts and also the voice recordings). The duration of recordings for which we can publish only the transcripts amounts to 41:15 and 1:36:29 for dev and test sets. Consent to publish also the audio was given for an additional amount of 1:18:16 and 3:24:47 for dev and test set, respectively. Statistics regarding *MockConf* are in Table 1; more details for each recording pair can be found in Appendix A.1 and in Appendix A.2, where we list the conference main themes. We have also collected an equivalent amount of recordings with consent from only one of the participant students; these are not used in this study and are reserved for the future creation of training data.

2.2 *InterAlign*: Our annotation tool

After transcriptions, a second layer of annotations consists of alignments between the source and target speeches. We perform this alignment for transcripts of complete speeches. Existing tools are designed mainly to align parallel textual corpora of translations, which differ from our transcripts in many ways: for instance, we cannot rely on existing sentence correspondences (Zhao et al., 2024), which is also illustrated in Figure 1. We therefore implemented and used our own annotation tool, *InterAlign*, with the main focus on facilitating the annotation process of interpreting spans and word alignments. We discuss existing tools and their limitations in Appendix B, as well as the implementation and usage details of *InterAlign*.

2.3 Annotation guidelines and process

Span-level annotation The goal of the span-level alignment is to help us monitor and analyze the interpreting process: to separate parts that are adequate and precise *translations* from *reformulations*, where the interpreter needed to compress its translation for the sake of time, and from *errors*. *Reformulations* happen when interpreters are cognitively overloaded or decide that the audience in the target language could be similarly overloaded and adopt strategies such as *generalization*, *summarization*, or *paraphrasing* (Al-Khanji et al., 2000; He et al., 2016). Generally, we define *reformulations* as a less literal version of translations that convey the same meaning in the given context. For errors, we consider the taxonomy of translation departures in simultaneous interpreting designed by (Barik, 1994) consisting of *omissions*, *additions*,[5] and *replacements*. We further sub-categorize additions and omissions as *factual* or *uninformative*. The difference between them is that *factual* omissions (resp. additions) alter the amount of information conveyed, whereas *uninformative* omissions (resp. additions) do not. A similar labeling system is used by Doi et al. (2021); Zhao et al. (2024). The list of span labels is in Table 2.

Label types

category	subcategory	label
Translation	-	TRAN
Reformulation	Paraphrase	PARA
	Summariaztion	SUM
	Generalization	GEN
Addition	Factual	ADDF
	Uninformative	ADDU
Replacement	-	REPL

Table 2: Label types and their subcategories.

Word-level annotation For each span-aligned pair, we also annotate word alignments. We forbid word alignment links between different span pairs. We define word alignment as *sure* if the corresponding pair of words is a context-independent translation and as *possible* if the context is needed or a grammatical dependency is required (Bojar and Prokopová, 2006) to understand the correspondence. An example annotation is in Figure 7 in Appendix B.

For this alignment process, we recruited 5 professional translators, all of them native Czech speakers, who were paid 200 CZK per hour. The total cost of annotating the whole dataset was 25 000 CZK. The annotator guidelines as well as the precise definitions of labels are in Appendix D; the activity of each annotator is in Table 3.

3 *MockConf*: Properties and analysis

3.1 Annotation differences per annotator

Granularity Figure 2 displays the distribution of span lengths across labels and annotators. The data reveal notable differences in annotator styles, particularly in the lengths of the spans they identify.

[5]Additions in the source side correspond to omissions in interpreting and vice versa.

		Development		Test	
id	lang.	count	duration	count	duration
1	de	1	00:09:47	5	00:51:41
2	en	5	00:55:47	3	00:30:13
3	en	1	00:11:09	8	01:33:00
4	es	2	00:22:41	5	00:50:41
5	fr	1	00:20:07	8	01:15:41

Table 3: Summary of annotators's activity on the development and test sets.

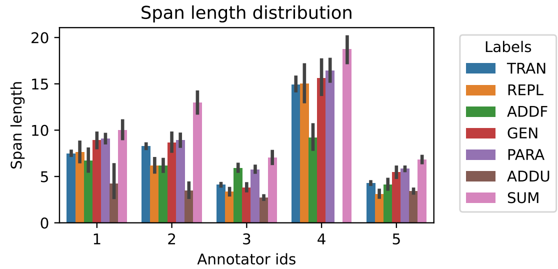

Figure 2: Span length (in tokens) distribution per label and per annotator. The annotators are denoted by their ids which are consistent with Table 3.

Annotator 4 consistently reports longer spans — nearly twice as long as those of other annotators. In contrast, Annotators 3 and 5 tend to annotate much shorter spans. These differences may stem from two potential factors: (1) variability in the annotators' interpretation of the boundary between translation and non-translation, or (2) a lack of adherence to the annotation guidelines.

We believe that the major factor influencing the outputs in Figure 2 is the former. For example, a paraphrase might be labeled as a single span by one annotator, while another might use a more fine-grained approach, resulting in multiple spans. This stems from the fact that, at the token level, distinctions between translations and synonyms / paraphrases can be ambiguous.

Inter-annotator agreement To better understand the differences between annotators, we annotated one recording from the development set twice. The selected recording involves Czech and English and was annotated by two annotators.[6] We computed Cohen's Kappa for segmentation (a binary decision regarding span boundaries) and for label agreement, evaluated at the token level (assigning span labels to individual tokens). Additionally, we assessed whether alignment links match, counting

[6]We chose this language pair because it was the only one with two annotators available.

				Exact match			
segmentation		label		Ann2-Ann3		Ann3-Ann2	
src	tgt	src	tgt	w/	w/o	w/	w/o
0.56	0.57	0.41	0.25	14.85	24.26	19.87	30.46

Table 4: Cohen's Kappa for segmentation and label prediction, and the percentage of links the annotators agree upon with the distinction labels vs. no-labels.

both exact matches (corresponding both to similar span boundaries and labels) or a less strict matches (disregarding labels).

The results presented in Table 4 show the following trends: for segmentation, Cohen's Kappa scores are 0.56 and 0.57 for the source and target sides, indicating moderate agreement (Landis, 1977). For label agreement, the scores are 0.41 and 0.25 for the source and target sides, corresponding to moderate and fair agreement, respectively. The proportions of identical alignment links are 14.85% (with labels) and 24.26% (without labels) when using annotator 3 as the reference. In the reverse direction, these proportions increase to 19.87% and 30.46%. Upon further inspection, we attribute this discrepancy to the fact that annotator 2 produced fewer alignment links. See Appendix A.5 for an example of such disagreement. Overall, these results underscore the difficulty of the task, as alignment link presupposes accurate segmentation, which, as we saw, is not guaranteed due to the task ambiguities.

3.2 Analysis of length differences

Since interpreting typically produces shorter output than the input speech, we analyze this phenomenon from several perspectives: span length, relay (indirect) interpreting, and multi-track interpreting.

Spans Figure 3 (left) displays the distribution of span lengths (in tokens). The distribution seems to be uniform, except for *uninformative additions*. Further inspection of *additions* reveals that they are shorter because they contain only filler words, incomplete words or words such as "very", "much" etc. This figure also suggests that there is clear shortening happening in pairs of segments labeled *summarization*. We thus plot the weighted average (with weights corresponding to the word counts in the source segment) of ratios of the target and source span length. We use a weighted average to make longer segments contribute more since the ratio in short segments can be caused only by the

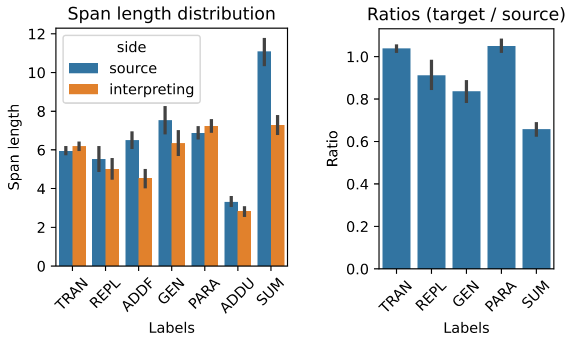

Figure 3: Left: Span length (in tokens) distribution per label for both source and target sides; Right: Weighted average of span length ratios (target / source) per label.

grammatical properties of language (e.g. articles in the English text that are not present in Czech).

Figure 3 (right) displays length ratios for each span label.[7] We see that the ratios for translation and paraphrase are very close to 1, as expected. Another observation is that length ratios for *generalization* and *summarization* are lower than one: 0.9 and 0.6 on average, respectively. This also aligns well with our intuition.

Relay interpreting Our corpus contains 27 direct interpretations and 12 indirect (relay) interpretations. On average, the ratio of source length to interpreting length, measured in characters, is 77.5% for direct interpreting and 97.43% for relay interpreting. This suggests that relay interpreting may be somewhat easier than direct interpreting, as the first interpreter often already simplifies the content. Additionally, we observe a higher proportion of translations and fewer additions in relay interpreting. Further details are in Appendix A.3.

Multi-track interpreting Another interesting feature of our interpreting dataset is the inclusion of multi-track interpreting, where the same speech is interpreted into the same language by two interpreters. We identified 7 such pairs and computed the average length ratio at both the character and token levels. On average, such pairs of interpretations differ by only 2%, but the *maximum* difference reaches 15% for characters and 10% for tokens. Detailed statistics are in Appendix A.4.

3.3 Errors in interpreting

We study the coverage of spans with respect to the distribution of labels to analyze potential errors and discrepancies. In Table 5, we report the number of tokens belonging to each span label for both the source and the interpreting sides. The most frequent span label is *translation*, which makes up for approximately half of all cases. The second is *paraphrase*, accounting for one fifth. These results are in line with our intuition. We also observe that 13.3% of tokens belong to spans where a factual omission is detected. Interestingly, there are also some factual additions in the target speech. We hypothesize it might happen when the interpreter misunderstands some part of the speech, but given the context, it is not suitable to label it as a *replacement*. Some examples are discussed in Section 3.4.

	TRAN	PARA	SUM	ADDF	GEN	ADDU	REPL
source	42.82	17.91	11.89	13.28	4.68	5.45	3.96
target	52.16	22.08	9.07	4.02	4.57	3.91	4.18

Table 5: The percentage of tokens with respective labels in the source and target side.

3.4 Examples

Table 6 presents some examples of annotations. We observe that there are some factual additions in the interpreting. This happens in cases when an interpreter is influenced by the preceding context and repeats information that conflicts with the original speech. For instance, in one talk, the speaker mentioned "camera" in combination with "artificial intelligence". This was later brought up by the interpreter even though it was not mentioned in the corresponding speech segment.

4 Towards automatic alignment

In this section, we showcase the use of *MockConf* as a useful resource to develop and evaluate alignment tools for interpreting. We describe a baseline system computing annotations at the word and span levels, then propose metrics to measure its accuracy and finally highlight its limitations.

4.1 Methodology

We implemented a simple system for automatic alignment similar to the proposal of Zhao et al. (2024), which operates in three steps: (1) coarse alignment, (2) sub-segmentation to identify span-aligned pairs (with word alignment links within them), and (3) assigning labels to the span-level alignment links.

Coarse Alignment The first step is to obtain a high-precision coarse alignment at the span level.

[7]We do not display ADDU and ADDF, as additions lack the counterpart for comparison.

Label	Example (source speech $\rightarrow$ target speech)
TRAN	share the screen with my presentation $\rightarrow$ *share the screen with my presentation*
PARA	No one can predict what will or won't happen $\rightarrow$ *Because many things can happen*
SUM	And what can you do as an expectant mother? $\rightarrow$ *As for mothers*
GEN	gynaecologist $\rightarrow$ *doctor*; abuse $\rightarrow$ *rude behavior*
REPL	36.1 $\rightarrow$ 36,9; 12.4 $\rightarrow$ 12; in 2005 or after , not before 2005 $\rightarrow$ *from 2005 to 2016*
ADDF	towards this artificial intelligence which didn't $\rightarrow$ towards *this camera* and the artificial intelligence didn't
ADDU	*For example; Next; Okay; can be also seen; And obviously*

Table 6: Example alignment links and their labels. For illustration purposes, all texts are translated into English even though they occurred in a different language in the dataset. Parts in *italics* denote spans that were marked with the corresponding line label.

For this, we use BERTAlign (Liu and Zhu, 2023), a sentence alignment tool, configured with the following parameters: `max_align 10`, `top_k 10`, `window 10`, `skip 0.0` and `len_penalty`.[8] We emphasize that this process produces n-m sentence alignments, as interpreting naturally deviates from the traditional 1-1 sentence alignment that is majoritary observed in textual parallel corpora. High precision is prioritized at this stage to ensure the quality of subsequent sub-segmentations. We denote the resulting system for this first step as *BA*.

Sub-segmentation We compute sub-segmentation and word alignments simultaneously. First, we identify all word alignment links using the `itermax` strategy from Jalili Sabet et al. (2020), configured with zero distortion, and the XLM-R model for computing contextual word embeddings (Conneau et al., 2020). Next, we refine the spans by splitting them at points where two punctuations align in the source and target transcripts. This step generates additional span-level alignment links with shorter spans, resulting in system *BA+sub*.

Labeling As previous steps may generate additions (n-0 or 0-m alignments) and translations (n-m alignments), we label additions as ADDU as it is the most frequent subcategory, and translations simply as TRAN. To also predict the other labels, we implemented a very simple classification model in PyTorch[9] which takes the similarity score calculated by the multilingual sentence embedder LaBSE (Feng et al., 2022), taking source and target span length as input features. It passes them through two hidden fully connected layers of size 100 and classifies the output into 5 categories, resulting in the system denoted *BA+sub+lab*. Since we do not have training data yet, we use devset for training where we take 80% of devset for actual training and 20% as held-out data for the evaluation during the training.

[8]Refer to the original work for the parameter description.
[9]`https://pytorch.org/`

4.2 Metrics

The tasks considered in this work combine three difficulties: (a) to find the right spans, both in the source and the target; (b) to identify the correct alignment links between these spans, and with them the correct word alignments; (c) to label the links with their appropriate type. Our evaluation metrics take these three aspects into account.

Segmentation We evaluate the quality of span splits using accuracy, precision, recall, and F_1 of span boundaries, separately for the source and target texts. To also reward segmentation boundaries that are almost correct, we consider less severe metrics such as P_k (Beeferman et al., 1999) and WindowDiff (Pevzner and Hearst, 2002). P_k works by sliding a window of size k over the text, comparing whether pairs of words at the boundaries of the window fall within the same span or not in both the source and target language. WindowDiff focuses on comparing the number of boundaries within a sliding window of size k. In practice, k is set to the half of the average span size in the reference ($k = 3$ in our case). Both metrics report a probability of error, with lower values corresponding to better segmentation. We use the NLTK implementation of these metrics (Loper and Bird, 2002).[10]

Span and word Alignment We compute the proportion of exact matches for span alignment, which we call *Exact match*. We distinguish between matching both span boundaries and labels or

[10]`https://www.nltk.org`

only span boundaries. As this metric is very strict, we also define an approximate span alignment error, which, similarly to the sentence alignment error (Véronis and Langlais, 2000), takes near-miss into account. This is computed as follows: for each pair of segments (s, t) occurring in the reference or hypothesis alignment, we compute a list containing all possible word pairs (u, v) with $u \in s$ and $v \in t$. Taking the union of such lists over the reference and hypothesis alignment yields two lists of word-level links, from which we compute Precision, Recall, and F_1. We refer to this metric as *Relaxed match*. For the word alignment, we report Alignment Error Rate (AER) and F_1 macro-averaged over all recordings. These scores are computed with the implementation of Azadi et al. (2023).

Label match Given the difficulty of obtaining high segmentation quality and exact matches for alignment links, we only evaluate label correctness at the token level: Each token is labeled like the span it belongs to, and we then assess the proportion of correct link labels using accuracy and F_1.

4.3 Baselines

Span alignment baseline For the evaluation of segmentation, span alignment, and labeling, we compare BA to a random baseline, which randomly selects the same number of boundaries in the source (resp. target) sides compared to the reference alignment, and iterates through segments on both sides in parallel from left to right, randomly selecting a link label from the shuffled pool of reference alignment links. This ensures that the number of labels of each type is the same as in the reference. Note that if the label is ADDU or ADDF, the span on only one side is labeled; otherwise the alignment link is created.

Word alignment baseline For word alignments, we use SimAlign (Jalili Sabet et al., 2020) as a baseline applied to the whole set of transcripts. We compute contextual embeddings using a sliding window of size 128 with stride 64. We discard links that connect words further than 50 tokens away, i.e. given source word w_s and target word w_t with their respective positions p_{w_s} and p_{w_t}, we discard links if $|p_{w_s} - p_{w_t}| > 50$.

4.4 Results

We evaluate the random baseline and our systems on three dataset splits: (1) one recording for which a double annotation is available; (2) development set and (3) test set. The alignments for (1) are evaluated separately for each annotator, with the annotator ID is indicated as a subscript. The results are in Table 7 and further detailed below:

Segmentation The first block of Table 7 presents the evaluation of segmentation quality. As intended in the first step, BA demonstrates a very high precision. While sub-segmentation slightly reduces precision, it improves the overall F_1 score. Notably, BA+sub even surpasses annotator 2 in inter-annotator comparisons, as reflected in both the F_1 score and in metrics such as Window Diff and P_k.

Span and Word Alignment The second and third blocks of Table 7 report the quality of relaxed and exact matches for predicted span alignment links, respectively. For relaxed matches, BA+sub performs slightly below the level of inter-annotator agreement. In the case of exact matches (third block), performance varies depending on the comparison with Annotator 1 or Annotator 2. This difference can be attributed to the number of alignment links: Annotator 1 (143 links) aligns more closely with BA (90 links) compared to Annotator 2. The fourth block of Table 7 evaluates the quality of word alignment links, showing that traditional word alignment tools designed for MT struggle due to the longer context in interpreting. Even with a moving window that discards distant links, the baseline approach performs significantly worse than our method.

Label Match The final block of Table 7 reports the quality of per-token annotation labels. While label classification improves upon the default label prediction, the improvement is modest. This suggests that segmenting solely based on punctuations inserted in the transcription phase is insufficient for interpreting, highlighting the need for a more fine-grained solution. We leave this for future work.

5 Related work

Sentence alignment Sentence-aligned corpora are key to modern MT and have been studied since statistical MT emerged (Tiedemann, 2011). Their mostly monotonic, 1-to-1 nature makes alignment computationally efficient, enabling large parallel data repositories like Opus (Tiedemann, 2012).[11]

Word alignment Word alignment annotation has been widely studied, starting with the Bible

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

	sys$_{annotator}$	Segmentation P↑	R↑	F_1↑	Df↓	P_k↓	Relaxed match P↑	R↑	F_1↑	Exact match w/↑	w/o↑	Word align. AER↓	F_1↑	Label match acc↑	F_1↑	#span src	tgt
1 recording	Baseline$_2$	15.12	14.44	14.77	0.50	0.47	0.08	0.06	0.07	0.00	0.00	0.74	0.25	55.35	56.65	145	123
	Baseline$_3$	21.32	15.80	18.15	0.54	0.49	0.05	0.04	0.04	0.00	0.00	0.70	0.28	36.00	29.48		
	BA$_2$	97.37	41.11	57.81	0.23	0.23	0.43	0.99	0.60	10.60	14.57	0.30	**0.71**	**76.20**	67.73	59	53
	BA$_3$	**98.25**	32.18	48.48	0.33	0.30	0.39	**1.00**	0.56	2.97	10.40	0.36	0.65	48.74	34.05		
	BA+sub$_2$	86.67	52.96	**65.75**	**0.21**	**0.20**	0.52	0.82	0.63	15.89	18.54	0.34	0.66	**76.20**	67.73	87	76
	BA+sub$_3$	85.45	40.52	54.97	0.32	0.28	0.45	0.80	0.58	4.46	11.39	0.40	0.61	48.74	34.05		
	BA+sub+lab$_2$	86.67	52.96	**65.75**	**0.21**	**0.20**	0.52	0.82	0.63	15.89	18.54	0.34	0.66	72.43	**70.36**	87	76
	BA+sub+lab$_3$	85.45	40.52	54.97	0.32	0.28	0.45	0.80	0.58	4.95	11.39	0.40	0.61	47.61	37.46		
	Annotator3$_2$	56.61	**72.96**	63.75	0.30	0.25	**0.78**	0.70	**0.74**	**19.87**	**30.46**	**0.28**	**0.71**	57.60	65.39	184	159
	Annotator2$_3$	72.96	56.61	63.75	0.30	0.25	0.70	0.78	**0.74**	14.85	24.26	0.36	0.66	57.60	49.82	145	123
devset	Baseline	17.16	16.07	16.60	0.47	0.43	0.04	0.03	0.03	0.14	0.18	0.70	0.30	36.18	37.93	195	176
	BA	**95.59**	35.33	51.59	0.25	0.23	0.38	**0.97**	0.54	6.83	11.98	**0.32**	**0.69**	**58.52**	44.98	72	64
	BA+sub	79.45	**50.04**	**61.40**	**0.24**	**0.21**	**0.51**	0.71	**0.60**	**9.70**	**16.44**	0.38	0.63	58.48	44.98	125	107
	BA+sub+lab	79.45	**50.04**	**61.40**	**0.24**	**0.21**	**0.51**	0.71	**0.60**	9.61	**16.44**	0.38	0.63	52.25	**47.68**	125	107
testset	Baseline	19.34	17.79	18.53	0.51	0.45	0.05	0.03	0.04	0.14	0.24	0.75	0.27	26.86	27.70	213	185
	BA	**95.05**	28.26	43.56	0.31	0.30	0.28	**0.95**	0.44	4.21	10.39	**0.37**	**0.65**	**41.24**	25.92	62	59
	BA+sub	82.52	**43.43**	**56.91**	**0.28**	**0.25**	**0.44**	0.74	0.55	6.41	**13.80**	0.42	0.59	41.22	25.99	110	104
	BA+sub+lab	82.52	**43.43**	**56.91**	**0.28**	**0.25**	**0.44**	0.74	**0.55**	**6.55**	**13.80**	0.42	0.59	38.16	**31.91**	110	104

Table 7: The evaluation of our system is detailed as follows: w/ and w/o in the Exact Match evaluation represent results with and without labels, respectively. #span represents the average span count for each split. BA refers to the system after applying BERTAlign (the first step), +sub indicates the BA system extended with follow-up sub-segmentation (the second step), and +lab represents the system further enhanced by labeling (the third step). For 1 recording, the subscript indicates the ID of the annotator whose annotation is used for evaluating the alignment.

(Melamed, 1998) and the Canadian Hansards proceedings (Och and Ney, 2000), then expanding to more languages, mostly paired with English: Romanian, Hindi, Inuktitut (Martin et al., 2003), Spanish (Lambert et al., 2005), Czech (Bojar and Prokopová, 2006; Kruijff-Korbayová et al., 2006), and Portuguese (Graça et al., 2008), etc. These alignments are typically "flat", linking words directly. More complex alignments, mapping nodes in parallel parse trees, exist for Japanese, Chinese (Uchimoto et al., 2004), German (Volk et al., 2006), Danish (Buch-Kromann, 2007), and Chinese and Arabic (Gale project) (Li et al., 2010). The Czech-English parallel dependency treebank (Hajič et al., 2012) also provides large-scale automatic annotations. Such annotations capture not only word correspondences but also syntax-level equivalences. Hierarchical span alignments have been manually annotated for French using an iterative divisive procedure (Xu and Yvon, 2016). These works inspired our annotation guidelines (Appendix D). While most word alignments focus on written texts, speech data remains underexplored, except for broadcast news transcripts in the Gale project (Li et al., 2010).

Interpreting Datasets Several simultaneous interpreting corpora exist, including EPIC (Sandrelli and Bendazzoli, 2006), EPIC-Ghent (Defrancq, 2015), and EPTIC (Bernardini et al., 2016), which are small collections of transcribed European Parliament interpretations for analysis. Additional corpora have been published by Temnikova et al. (2017); Pan (2019). The ESIC corpus (Macháček et al., 2021) covers multiple languages and includes transcripts, translations, and simultaneous interpreting transcripts. Other resources, mainly for consecutive interpreting, are documented by Lazaro Gutierrez (2023). However, none of these corpora provide alignments between speeches.

Alignment annotation in interpreting Doi et al. (2021) present a large-scale (around 300 hours) English-Japanese simultaneous interpretation corpus along with the results of its analysis. Part of the dataset is manually annotated (14 TED talks) with categories such as additions, pragmatically uninformative omissions, and factual omissions. They further evaluate the dataset based on latency, quality, and word order. Building on this corpus, Zhao et al. (2024) provide an automatically aligned parallel English-Japanese interpretation dataset. Their approach, similar to ours, involves two steps: coarse alignment followed by fine-grained alignment. Their error analysis addresses unintentional omissions (corresponding to our "additions" in source speech), intentional omissions (summarization), and mistranslations (replacements).

6 Conclusion

In this paper, we have detailed our efforts to collect, prepare and annotate a corpus of simultaneous interpretings, performed by student interpreters in mock conferences. We discussed the guidelines used at each annotation step and reported the results of the first analysis of the resulting corpus. They illustrate how interpreting activities could be studied and monitored with corpus-based techniques; they also highlight the need to develop dedicated tools for their annotation. The resulting corpus and tools will be released to the community. In a final step, we used this new resource to evaluate automatic alignment tools for interpreting corpus: as it seems, this new task, which combines the difficulties of multiple existing annotation processes, poses challenges for our existing alignment tools.

Future work We plan to deepen our preliminary observations at several levels: to better correlate the main speaker's oral production with labels on the interpreting side; to also study how interpreting strategies vary depending on the source and target languages. A lot finally remains to be done to improve our automatic processing tools which do not rely on punctuation as it is a very unreliable alignment indicator in interpreting.

Limitations

We acknowledge that the current dataset is only limited in size and linguistic diversity, which is hardly compensated by the richness of available annotations. We are continuously working on extending this dataset, with the hope of accumulating a sufficiently large set of annotated speeches that could also be used for training (or fine-tuning) a supervised machine learning system and improving the automatic span-level annotations. Regarding the alignment tool, an obvious limitation is the lack of connection with the original speech, which needs to be transcribed by an external tool, then revised, before the alignment takes place. As a first step towards a tighter integration, we could work on providing the annotators with an integrated player, providing them with a way to listen to the original audio tracks and even correct the corresponding transcripts. We additionally emphasize that the Random baseline does not uses Reformulation or Replacement labels and our approach is suboptimal the second phrase where we sub-segment on the punctuation match.

Lower Inter-Annotator Agreement We consider an annotation "correct" when annotators agree. In an ideal scenario, annotators would discuss and align their approaches during annotation. However, we found this setup both time-consuming and impractical. Additionally, defining the distinction between paraphrase and non-paraphrase is inherently challenging. While introducing minimal blocks corresponding to syntactic units might be a potential direction, Leech (2000) has shown that syntax is not a good indicator of units in speech. Currently, we provided feedback on how well annotators adhered to guidelines after they annotated a part of the data. Despite these efforts, some divergence remains, reflecting the complexity of the task.

Ethics Statement

All data contained in the *MockConf* dataset are fully anonymized, e.i. they do not contain any personal information (names) about the speakers. We collected consents from speakers to publish recordings containing their voice and the transcripts of their speech. The participants were informed that their recordings will be used for research purposes.

Acknowledgements

The work has been partially supported by the grant 272323 of the Grant Agency of Charles University, SVV project number 260 821, the grant CZ.02.01.01/00/23_020/0008518 ("Jazykověda, umělá inteligence a jazykové a řečové technologie: od výzkumu k aplikacím"), and by the funds provided by the French-Czech Barrande Fellowship Programme. Part of this work has been done while the first author was visiting the Institute for Intelligent Systems and Robotics (ISIR) at Sorbonne Université in Paris, France. We would like to express our sincere gratitude to prof. PhDr. Ivana Čeňková, CSc., Mgr. Kateřina Ešnerová, and all the students who generously gave their consent to have their interpreting data used in this study.

References

Milind Agarwal, Sweta Agrawal, Antonios Anastasopoulos, Luisa Bentivogli, Ondřej Bojar, Claudia Borg, Marine Carpuat, Roldano Cattoni, Mauro Cettolo, Mingda Chen, William Chen, Khalid Choukri, Alexandra Chronopoulou, Anna Currey, Thierry Declerck, Qianqian Dong, Kevin Duh, Yannick Estève, Marcello Federico, Souhir Gahbiche, Barry

Haddow, Benjamin Hsu, Phu Mon Htut, Hirofumi Inaguma, Dávid Javorský, John Judge, Yasumasa Kano, Tom Ko, Rishu Kumar, Pengwei Li, Xutai Ma, Prashant Mathur, Evgeny Matusov, Paul McNamee, John P. McCrae, Kenton Murray, Maria Nadejde, Satoshi Nakamura, Matteo Negri, Ha Nguyen, Jan Niehues, Xing Niu, Atul Kr. Ojha, John E. Ortega, Proyag Pal, Juan Pino, Lonneke van der Plas, Peter Polák, Elijah Rippeth, Elizabeth Salesky, Jiatong Shi, Matthias Sperber, Sebastian Stüker, Katsuhito Sudoh, Yun Tang, Brian Thompson, Kevin Tran, Marco Turchi, Alex Waibel, Mingxuan Wang, Shinji Watanabe, and Rodolfo Zevallos. 2023. FINDINGS OF THE IWSLT 2023 EVALUATION CAMPAIGN. In *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023)*, pages 1–61, Toronto, Canada (in-person and online). Association for Computational Linguistics.

Lars Ahrenberg, Magnus Merkel, and Michael Petterstedt. 2003. Interactive word alignment for language engineering. In *10th Conference of the European Chapter of the Association for Computational Linguistics*, Budapest, Hungary. Association for Computational Linguistics.

Raja Al-Khanji, Said El-Shiyab, and Riyadh Hussein. 2000. On the use of compensatory strategies in simultaneous interpretation. *Meta*, 45(3):548–557.

Fatemeh Azadi, Heshaam Faili, and Mohammad Javad Dousti. 2023. PMI-align: Word alignment with pointwise mutual information without requiring parallel training data. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 12366–12377, Toronto, Canada. Association for Computational Linguistics.

Max Bain, Jaesung Huh, Tengda Han, and Andrew Zisserman. 2023. WhisperX: Time-accurate speech transcription of long-form audio. In *Proceedings of InterSpeech 2023*.

Henri C. Barik. 1994. A description of various types of omissions, additions and errors of translation encountered in simultaneous interpretation. *Bridging the gap: Empirical research in simultaneous interpretation*, 3:121–137.

A Barreiro, F Raposo, T Luís, and CLUE-Aligner-An Alignment Tool. 2016. CLUE-Aligner: An alignment tool to annotate pairs of paraphrastic and translation units. *Translation Evaluation: From Fragmented Tools and Data Sets to an Integrated Ecosystem*, page 7.

Doug Beeferman, Adam Berger, and John Lafferty. 1999. Statistical models for text segmentation. *Machine Learning*, 34(1-3):177–210.

Silvia Bernardini, Adriano Ferraresi, and Maja Miličević. 2016. From EPIC to EPTIC—Exploring simplification in interpreting and translation from an intermodal perspective. *Target*, 28(1):61–86.

Ondřej Bojar and Magdelena Prokopová. 2006. Czech-English word alignment. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy. European Language Resources Association (ELRA).

Matthias Buch-Kromann. 2007. Computing translation units and quantifying parallelism in parallel dependency treebanks. In *Proceedings of the Linguistic Annotation Workshop*, pages 69–76, Prague, Czech Republic. Association for Computational Linguistics.

Aljoscha Burchardt. 2013. Multidimensional quality metrics: a flexible system for assessing translation quality. In *Proceedings of Translating and the Computer 35*, London, UK. Aslib.

Mauro Cettolo, Christian Girardi, and Marcello Federico. 2012. WIT3: Web inventory of transcribed and translated talks. In *Proceedings of the 16th Annual Conference of the European Association for Machine Translation*, pages 261–268, Trento, Italy. European Association for Machine Translation.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

David Dale, Elena Voita, Loic Barrault, and Marta R. Costa-jussà. 2023. Detecting and mitigating hallucinations in machine translation: Model internal workings alone do well, sentence similarity Even better. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 36–50, Toronto, Canada. Association for Computational Linguistics.

Bart Defrancq. 2015. Corpus-based research into the presumed effects of short EVS. *Interpreting*, 17(1):26–45.

Mattia A. Di Gangi, Roldano Cattoni, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2019. MuST-C: a Multilingual Speech Translation Corpus. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2012–2017, Minneapolis, Minnesota. Association for Computational Linguistics.

Ebru Diriker. 2015. Simultaneous interpreting. In Franz Pöchhacker, editor, *Routledge Encyclopedia of Interpreting Studies*. Routledge, Taylor & Francis Group, London and New York.

Kosuke Doi, Yuka Ko, Mana Makinae, Katsuhito Sudoh, and Satoshi Nakamura. 2024. Word order in

English-Japanese simultaneous interpretation: Analyses and evaluation using chunk-wise monotonic translation. In *Proceedings of the 21st International Conference on Spoken Language Translation (IWSLT 2024)*, pages 254–264, Bangkok, Thailand (in-person and online). Association for Computational Linguistics.

Kosuke Doi, Katsuhito Sudoh, and Satoshi Nakamura. 2021. Large-scale English-Japanese simultaneous interpretation corpus: Construction and analyses with sentence-aligned data. In *Proceedings of the 18th International Conference on Spoken Language Translation (IWSLT 2021)*, pages 226–235, Bangkok, Thailand (online). Association for Computational Linguistics.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891, Dublin, Ireland. Association for Computational Linguistics.

Ulrich Germann. 2008. Yawat: Yet Another Word Alignment Tool. In *Proceedings of the ACL-08: HLT Demo Session*, pages 20–23, Columbus, Ohio. Association for Computational Linguistics.

Timur Gilmanov, Olga Scrivner, and Sandra Kübler. 2014. SWIFT aligner, a multifunctional tool for parallel corpora: Visualization, word alignment, and (morpho)-syntactic cross-language transfer. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 2913–2919, Reykjavik, Iceland. European Language Resources Association (ELRA).

João Graça, Joana Paulo Pardal, Luísa Coheur, and Diamantino Caseiro. 2008. Building a golden collection of parallel multi-language word alignment. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

Nuno M. Guerreiro, Pierre Colombo, Pablo Piantanida, and André Martins. 2023. Optimal transport for unsupervised hallucination detection in neural machine translation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13766–13784, Toronto, Canada. Association for Computational Linguistics.

Jan Hajič, Eva Hajičová, Jarmila Panevová, Petr Sgall, Ondřej Bojar, Silvie Cinková, Eva Fučíková, Marie Mikulová, Petr Pajas, Jan Popelka, Jiří Semecký, Jana Šindlerová, Jan Štěpánek, Josef Toman, Zdeňka Urešová, and Zdeněk Žabokrtský. 2012. Announcing Prague Czech-English Dependency Treebank 2.0. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 3153–3160, Istanbul, Turkey. European Language Resources Association (ELRA).

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

Masoud Jalili Sabet, Philipp Dufter, François Yvon, and Hinrich Schütze. 2020. SimAlign: High quality word alignments without parallel training data using static and contextualized embeddings. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1627–1643, Online. Association for Computational Linguistics.

Meng Ji, Pierrette Bouillon, and Mark Seligman. 2023. *Translation Technology in Accessible Health Communication*. Cambridge University Press.

Alina Karakanta, Matteo Negri, and Marco Turchi. 2020. MuST-cinema: a speech-to-subtitles corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3727–3734, Marseille, France. European Language Resources Association.

Tom Kocmi and Christian Federmann. 2023. GEMBA-MQM: Detecting translation quality error spans with GPT-4. In *Proceedings of the Eighth Conference on Machine Translation*, pages 768–775, Singapore. Association for Computational Linguistics.

Ivana Kruijff-Korbayová, Klára Chvátalová, and Oana Postolache. 2006. Annotation guidelines for Czech-English word alignment. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy. European Language Resources Association (ELRA).

Patrik Lambert, Adrià De Gispert, Rafael Banchs, and José B Mariño. 2005. Guidelines for word alignment evaluation and manual alignment. *Language resources and evaluation*, 39(4):267–285.

JR Landis. 1977. The measurement of observer agreement for categorical data. *Biometrics*.

Raquel Lazaro Gutierrez. 2023. Natural language processing tools and resources for translation and interpreting applications. introduction. In *Proceedings of the First Workshop on NLP Tools and Resources for Translation and Interpreting Applications*, pages 1–3, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Geoffrey Leech. 2000. Grammars of spoken english: New outcomes of corpus-oriented research. *Language Learning*, 50(4):675–724.

Xuansong Li, Stephanie Strassel, Stephen Grimes, Safa Ismael, Xiaoyi Ma, Niyu Ge, Ann Bies, Nianwen Xue, and Mohamed Maamouri. 2010. Parallel aligned treebank corpora at LDC: Methodology, annotation and integration. In *Proceedings of Workshop on Annotation and Exploitation of Parallel Corpora (AEPC), Tartu*.

Yunmeng Li, Jun Suzuki, Makoto Morishita, Kaori Abe, and Kentaro Inui. 2025. MQM-chat: Multi-dimensional quality metrics for chat translation. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 3283–3299, Abu Dhabi, UAE. Association for Computational Linguistics.

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).

Lei Liu and Min Zhu. 2023. Bertalign: Improved word embedding-based sentence alignment for chinese–english parallel corpora of literary texts. *Digital Scholarship in the Humanities*, 38(2):621–634.

Edward Loper and Steven Bird. 2002. Nltk: the natural language toolkit. *CoRR*, cs.CL/0205028.

Qingyu Lu, Liang Ding, Kanjian Zhang, Jinxia Zhang, and Dacheng Tao. 2025. MQM-APE: Toward high-quality error annotation predictors with automatic post-editing in LLM translation evaluators. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 5570–5587, Abu Dhabi, UAE. Association for Computational Linguistics.

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

Mana Makinae, Katsuhito Sudoh, Masaru Yamada, and Satoshi Nakamura. 2025. An automatic quality metric for evaluating simultaneous interpretation.

Joel Martin, Howard Johnson, Benoit Farley, and Anna Maclachlan. 2003. Aligning and using an English-Inuktitut parallel corpus. In *Proceedings of the HLT-NAACL 2003 Workshop on Building and Using Parallel Texts: Data Driven Machine Translation and Beyond*, pages 115–118.

I Dan Melamed. 1998. Annotation style guide for the Blinker project. *arXiv preprint cmp-lg/9805004*.

Franz Josef Och and Hermann Ney. 2000. Improved statistical alignment models. In *Proceedings of the 38th annual meeting of the association for computational linguistics*, pages 440–447.

Gustavo Paetzold, Fernando Alva-Manchego, and Lucia Specia. 2017. MASSAlign: Alignment and annotation of comparable documents. In *Proceedings of the IJCNLP 2017, System Demonstrations*, pages 1–4, Tapei, Taiwan. Association for Computational Linguistics.

Jun Pan. 2019. 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)*, pages 82–88, Varna, Bulgaria. Incoma Ltd., Shoumen, Bulgaria.

Xiao Pan, Mingxuan Wang, Liwei Wu, and Lei Li. 2021. Contrastive learning for many-to-many multilingual neural machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 244–258, Online. Association for Computational Linguistics.

Lev Pevzner and Marti A. Hearst. 2002. A critique and improvement of an evaluation metric for text segmentation. *Computational Linguistics*, 28(1):19–36.

Reid Pryzant, Youngjoo Chung, Dan Jurafsky, and Denny Britz. 2018. JESC: Japanese-English subtitle corpus. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Manny Rayner. 2000. *The spoken language translator*. Cambridge University Press.

Anthony Rousseau, Paul Deléglise, and Yannick Estève. 2012. TED-LIUM: an automatic speech recognition dedicated corpus. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 125–129, Istanbul, Turkey. European Language Resources Association (ELRA).

Annalisa Sandrelli and Claudio Bendazzoli. 2006. 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)*, Genoa, Italy. European Language Resources Association (ELRA).

Maria Skeppstedt, Magnus Ahltorp, Gunnar Eriksson, and Rickard Domeij. 2020. Line-a-line: A tool for annotating word-alignments. In *Proceedings of the 13th Workshop on Building and Using Comparable Corpora*, pages 1–5, Marseille, France. European Language Resources Association.

David Stallard, Rohit Prasad, Prem Natarajan, Fred Choi, Shirin Saleem, Ralf Meermeier, Kriste Krstovski, Shankar Ananthakrishnan, and Jacob Devlin. 2011. The bbn transtalk speech-to-speech translation system. In *Speech and Language Technologies*. IntechOpen.

Steinþór Steingrímsson, Hrafn Loftsson, and Andy Way. 2021. CombAlign: a tool for obtaining high-quality word alignments. In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 64–73, Reykjavik, Iceland (Online). Linköping University Electronic Press, Sweden.

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

Irina Temnikova, Ahmed Abdelali, Samy Hedaya, Stephan Vogel, and Aishah Al Daher. 2017. Interpreting strategies annotation in the WAW corpus. In *Proceedings of the Workshop Human-Informed Translation and Interpreting Technology*, pages 36–43, Varna, Bulgaria. Association for Computational Linguistics, Shoumen, Bulgaria.

Jörg Tiedemann. 2011. *Bitext Alignment*. Number 14 in Synthesis Lectures on Human Language Technologies, Graeme Hirst (ed). Morgan & Claypool Publishers.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218, Istanbul, Turkey. European Language Resources Association (ELRA).

Kiyotaka Uchimoto, Yujie Zhang, Kiyoshi Sudo, Masaki Murata, Satoshi Sekine, and Hitoshi Isahara. 2004. Multilingual aligned parallel treebank corpus reflecting contextual information and its applications. In *Proceedings of the Workshop on Multilingual Linguistic Resources*, pages 57–64, Geneva, Switzerland. COLING.

Henk van den Heuvel, Khalid Choukri, Christian Gollan, Asuncion Moreno, and Djamel Mostefa. 2006. TC-STAR: New language resources for ASR and SLT purposes. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy. European Language Resources Association (ELRA).

Jean Véronis and Philippe Langlais. 2000. *Evaluation of Parallel Text Alignment Systems*, Text Speech and Language Technology Series, chapter X. Kluwer Academic Publishers. Parallel Text Processing.

Martin Volk, Sofia Gustafson-Capková, Joakim Lundborg, Torsten Marek, Yvonne Samuelsson, and Frida Tidström. 2006. XML-based phrase alignment in parallel treebanks. In *Proceedings of the 5th Workshop on NLP and XML (NLPXML-2006): Multi-Dimensional Markup in Natural Language Processing*.

Alex Waibel. 2004. Speech translation: past, present and future. In *Interspeech 2004*, pages 353–356.

Xiaolin Wang, Masao Utiyama, and Eiichiro Sumita. 2023. Japanese-to-English simultaneous dubbing prototype. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 169–178, Toronto, Canada. Association for Computational Linguistics.

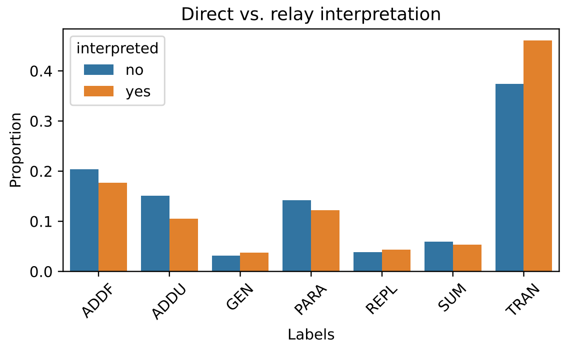

Figure 4: Relative proportion of each span label within each category (source interpreted: yes and no).

Shira Wein, Te I, Colin Cherry, Juraj Juraska, Dirk Padfield, and Wolfgang Macherey. 2024. Barriers to effective evaluation of simultaneous interpretation. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 209–219, St. Julian's, Malta. Association for Computational Linguistics.

Yong Xu and François Yvon. 2016. Novel elicitation and annotation schemes for sentential and subsentential alignments of bitexts. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 628–635, Portorož, Slovenia. European Language Resources Association (ELRA).

Tariq Yousef, Chiara Palladino, Farnoosh Shamsian, and Maryam Foradi. 2022. Translation alignment with ugarit. *Information*, 13:65.

Jinming Zhao, Katsuhito Sudoh, Satoshi Nakamura, Yuka Ko, Kosuke Doi, and Ryo Fukuda. 2024. NAIST-SIC-aligned: An aligned English-Japanese simultaneous interpretation corpus. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 12046–12052, Torino, Italia. ELRA and ICCL.

A Details about the *MockConf* dataset

A.1 Statistics

Detailed statistics of our *MockConf* dataset are presented in Table 8.

A.2 Topics

The topics for each speech of our *MockConf* dataset are presented in Table 9.

A.3 Direct vs. relay interpreting

Figure 4 presents the difference between direct and relay interpreting in terms of distribution of labels. we observe a higher proportion of translations and fewer additions in relay interpreting.

	Language		Interpreting		Annotator		Recording		
split	src	trg	relay	interpreter id	consent	annotator id	src id	trg id	duration
dev	cs	de	no	8	1	es	9	10	00:11:21
			yes	8	1	de	11	12	00:09:47
		en	no	1	3	en 1	5	6	00:12:31
							7	8	00:09:04
							16	17	00:13:52
				6	3	en 2	9	13	00:11:09
			yes	11	3	en 1	3	15	00:10:15
				6	3	en 1	3	4	00:10:05
		es	no	4	3	es	9	14	00:11:20
		fr	yes	3	1	fr	1	2	00:20:07
test	cs	de	no	5	3	de	9	46	00:11:13
			yes	12	1	de	57	58	00:09:27
				5	3	de	11	39	00:09:47
		en	no	11	3	en 2	9	45	00:10:59
				2	3	en 1	7	18	00:09:04
							30	31	00:11:30
			yes	6	3	en 2	11	40	00:09:47
				7	3	en 2	11	41	00:09:47
				9	3	en 1	48	49	00:09:39
		es	yes	13	3	es	48	55	00:09:31
				4	3	es	3	51	00:10:20
							52	53	00:11:31
		fr	no	3	1	fr	5	27	00:12:31
							7	19	00:09:04
							34	35	00:07:54
	de	cs	no	12	1	de	47	48	00:09:37
				8	1	de	54	52	00:11:37
	en	cs	no	10	3	en 2	42	44	00:09:25
				2	3	en 2	22	23	00:09:42
							26	1	00:20:07
							36	37	00:13:48
				9	3	en 2	42	43	00:09:25
	es	cs	no	13	3	es	56	57	00:09:32
				4	3	es	38	11	00:09:47
	fr	cs	no	14	3	fr	50	3	00:09:53
				3	1	fr	20	21	00:09:10
							24	25	00:07:59
							28	29	00:08:32
							32	33	00:10:38

Table 8: Detailed statistics of *MockConf*. Consent values 1 and 3 denote consents to publish only transcripts or both the transcripts and audio, respectively.

A.4 Multi-track interpreting

Table 10 presents the length ratios calculated on characters and tokens for the pairs of interpretations that share the same speech.

A.5 Annotator Disagreement Example

Figure 5 illustrates the difference in annotation granularity that we discuss in Section 3. The first row in Figure 5 is annotated by Annotator 3 and the second row by Annotator 2. We can see that Annotator 3 makes segment splits more often and produces a more fine-grained annotation, whereas Annotator 2 prefers longer segments.

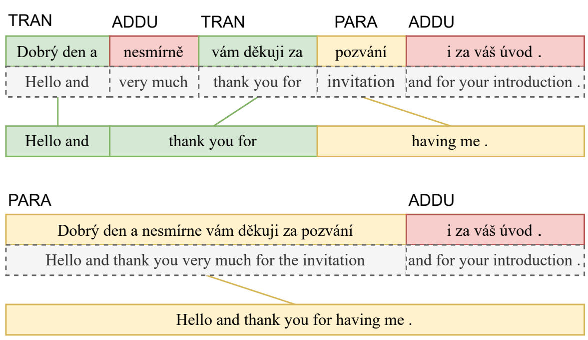

Figure 5: Two alignment annotations (by two different annotators) of the same sentence from the speech and its interpreting.

B Annotation tools and *InterAlign*

Existing tools We considered several existing tools: I*Link provides word-level alignment, com-

doc id	topic
1	Prevention of Traumatic Birth Experiences
3	From Maison des Cultures du Monde: The Scope of Work of This Institution
5	What Are the Benefits of Hypnobirthing
7	The Brain Is Not a Computer
9	A Cultural Anthropologist and Ethnologist Based at the University of Plzeň
11	From Yucatan University: Mayan Script and Its Decipherment
16	Harnessing Modern Technologies to Achieve Sustainable Development Goals
20	Utilization of AI in the Military Field
22	Scottish Inspiration for Prague
24	Shift Moonwalkers - The Future of Walking?
26	Prevention of Traumatic Birth Experiences
28	School Transport: Pedibus
30	Traffic Snake Game: Achieving Sustainable Mobility Through a Game
32	Que Choisir: Activities and Mission of This Association
34	Consumer Rights in the Past and Present and the Goals and Role of the dTest Organization
36	Regulating Ads in the Digital Age: An Impossible Task
38	From Yucatan University: Mayan Script and Its Decipherment
42	On Freelance Business Development: Benefits of Cultural Diversity in the Workplace
47	Team Leader of Charta der Vielfalt (Diversity Charter): Goals of the Charter and Activities of the Association
50	From Maison des Cultures du Monde: The Scope of Work of This Institution
52	Antigypsyism – History of Antigypsyism in Europe, Personal Experiences, Possible Solutions
56	From the Spanish Organization Unión Romaní: Antigypsyism and the Paradox of Tolerance During the Pandemic

Table 9: The topics of the speeches are listed alongside their document IDs in the first column. These IDs correspond to those in Table 8.

Document id		Ratio	
1. doc id	2. doc id	character	token
18	8	0.96	0.96
39	12	0.86	0.90
40	41	0.94	0.95
43	44	0.97	0.93
13	45	1.04	1.01
10	46	1.16	1.11
15	4	0.95	0.96

Table 10: Character and token ratios for multi-track interpreting. The first two columns denote ids of documents that are interpretations of the same speech. More details about the documents are in Table 8 and Table 9.

Alignment annotation tool	Word-level	Phrase-level	Multilevel	Long texts	Modern
I*Link (Ahrenberg et al., 2003)	✓				
YAWAT (Germann, 2008)	✓	✓			✓
Swift Aligner (Gilmanov et al., 2014)	✓				
CLUE-Aligner (Barreiro et al., 2016)	✓	✓	✓		✓
MASSAlign (Paetzold et al., 2017)	✓	✓		✓	
Line-a-line (Skeppstedt et al., 2020)	✓	✓			✓
ManAlign (Steingrímsson et al., 2021)	✓				
Ugarit (Yousef et al., 2022)	✓	✓		✓	✓
InterAlign (ours)	✓	✓	✓	✓	✓

Table 11: Existing alignment annotation tools and their main features.

piling reports and statistics, and automatic proposals for token alignments (Ahrenberg et al., 2003); YAWAT is a web-based tool for word and phrase-level alignments of parallel texts that are pre-segmented to sentences (Germann, 2008); Swift Aligner supports word-level alignment with additional capabilities for annotating dependency syntax and part-of-speech (Gilmanov et al., 2014); CLUE-Aligner is a web alignment tool designed for annotation of word or phrasal units in parallel sentences. (Barreiro et al., 2016); MASSAlign is Python library for the alignment and annotation of monolingual comparable documents at word and sentence levels (Paetzold et al., 2017); Line-a-line is a web-based tool for manual annotation of word-alignments in sentence-aligned parallel corpora (Skeppstedt et al., 2020); AlignMan is a tool for manual word alignment of parallel sentences (Steingrímsson et al., 2021); Ugarit is a public web-based tool for manual annotation of parallel texts for generating word- and phrase-level translation alignment, supporting the alignment between three parallel texts. A compact overview of all these tools is in Table 11.

Our requirements Alignment of interpretings, however, differs from that of text translations, which is usually performed in two stages: first at the sentence, then at the word level. This is because interpretings do not include unambiguous sentence boundaries in their transcripts. Interpreters often also omit, or rephrase long spans, trying to jointly accommodate time and content-preservation constraints, making the resulting transcripts difficult to word-align.

Since we cannot rely on any prior sentence segmentation or sentence alignment between the source and interpreting, a strong requirement for us was to support the annotation of long spans comprising dozens of tokens. This narrowed our list of options down to practically one tool: Ugarit (Yousef et al., 2022). Upon testing, we observed that it could not be used to perform both lexical and phrasal alignments at the same time.

InterAlign We, therefore, implemented a new annotation tool, *InterAlign*, that is primarily designed to be used for aligning transcripts of speech and their interpreting but can be used in any situation when no sentence segmentation and alignment is provided. It supports annotations at both the word- and span-level, can handle long texts, and enables the user to define its own span labels. The tool is implemented in React,[12] a modern web-based framework; it combines many individual features from previous annotation tools.

A screenshot of this tool is in Figure 6. The transcripts are displayed horizontally in two scrollable elements, enabling the alignment of long chunks. Annotation links can be created either by both mouse or keyboard actions. After creating an alignment, the link is added to the list and displayed under the annotation interface. A screenshot of the link list is in Figure 7.

C Transcript revision guidelines

1. Please correct the transcripts to match what is said in the recordings.
    - Do not correct grammar – if the speaker makes grammatical or any other language mistakes (stutters, repeats himself, uses the wrong form of a word or a whole word), the transcript should capture the exact notation of what is said.
    - For example, you can edit the stutter in the word international as: "inter- international" (with space between words).
    - Please record hesitations, interjections, etc. if they are obvious or inaudible. Please mark hesitations with @.
    - Please do not mark smacking and swallowing.
    - Please indicate a longer time delay in the speech with three dots.
2. You can change the segmentation to sentences.
    - Transcripts already contain sentences. It is possible that a different sentence division is suitable, but you are welcome to create your own sentence division (but this is not required).
    - Please edit the sentences so that each one is on a separate line.
3. Label proper names.
    - Recordings can contain the names of cities, organizations - it is important to mark these proper names with the [NAME] tag, for example, the sentence on the left will be the sentence on the right after the arrow: Václav was then in the Czech Republic. → [NAME](Václav) was then in [NAME](Czech Republic).

[12]https://react.dev/

D Annotation Guidelines

D.1 Phrase-level alignments

Segmentation

- Divide the speech and its interpretation into segments that correspond to each other and label them with the following labels.
- Each segment's length should be maximal, meaning adding one more word to either side would change the label of the segment.
- Each word is assigned to exactly one segment.

Labeling Criteria Labels are assigned to the interpretation when you compare it to the source speech. For instance, "summarization" means that some part of the interpreting (the second transcript) is summarized given the original speech (the first transcript). Similarly, "addition" means that some information is added in the interpretation. More precisely, the labels are:

- Translation: Direct translation that holds outside of any additional context.
- Reformulation:
    - "Paraphrase": Equivalent meaning in the context, but not a direct translation.
    - "Summarization": Equivalent meaning but the interpretation is expressed in less words, summarized.

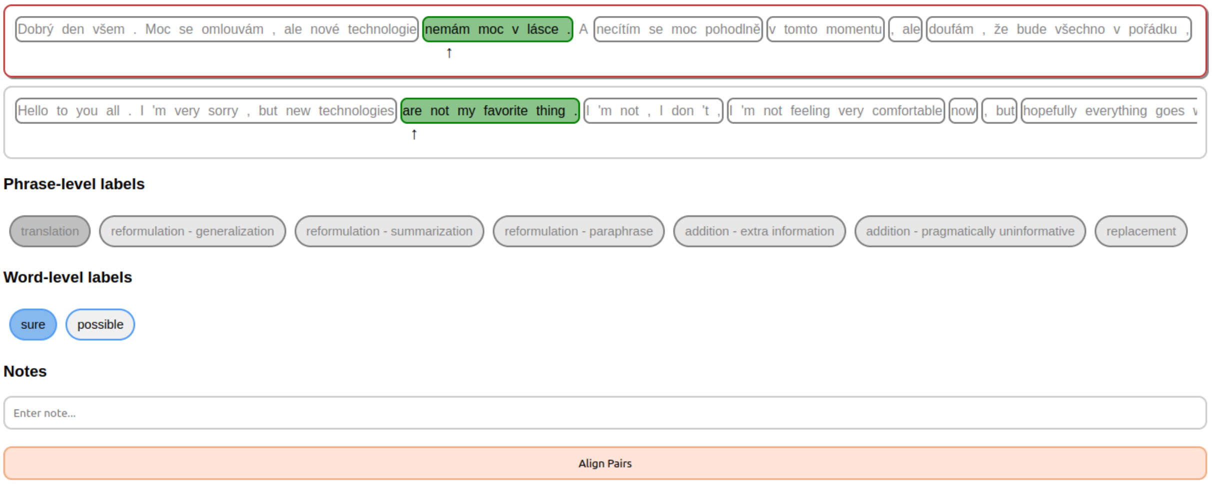

Figure 6: A screenshot of *InterAlign* for aligning transcripts of speech and their interpreting.

Figure 7: List of chunks and word alignment links displayed below the alignment window.

- "Generalization": The meaning is as close as possible, but one side of the aligned pair is less specific. For instance, instead of saying "cats and dogs", it is said "pets". Or instead of a particular name of a village, there is only "some village" mentioned.

- Addition: Used only on one of the sides, to indicate that this span brings additional content not present in the other language. Please distinguish the following sub-classes of "addition":
  - Extra information: the interpreting adds some new information, the meaning of the text is changed;
  - Pragmatically uninformative: the interpreting does not change the meaning, the span repeats something that has already been said or is not related to the topic.

- "Replacement": Obvious error, misunderstanding a number, place, name, etc. (e.g. instead of saying 17, it is said 70. In English it is very similar and it can be clear from the context that 70 is a replacement of 17)

Notes Make notes about any hesitations or uncertainties you may have during the annotation process.

D.2 Priorities of Phrase-level Labels

When considering which label to use for an aligned phrase pair, prefer segmentation and labels in this order:

1. "Translation" (Alignment): If a word in the source span directly corresponds to a translation in the target span out of any additional context, mark it as a translation alignment. Ensure accuracy and precision in aligning words with their translations.

2. "Reformulation": Identify phrases in the source span that convey the same meaning as phrases in the target span but are not direct translations. Use the reformulation label for such alignments with a specific category.

3. "Addition": Highlight cases where phrases are present in one span that do not have a direct counterpart in the other segment. Mark these as addition alignments with a specific category.

D.3 Word-level alignments

Within each pair of aligned segments (so you cannot create word-level alignment between words that belong to different phrase alignments) labeled translation or paraphrase, you will be annotating word-level alignments, distinguishing between "sure" links (direct translations) and "possible" links (including additional contextual information, determiners, cases, etc.).

- Sure Links (Direct Translation): Identify and mark word alignments that represent direct translations without any additional context. These alignments should reflect one-to-one correspondence between words with good translation equivalence.

- Possible Links (Additional Context): Identify and mark word alignments where additional contextual information or linguistic elements (such as determiners, cases, etc.) are present in one language and not in the other. These alignments are not for cross-language counterparts but indicate related, supplementary, or partial information.