Online Versus Offline NMT Quality: An In-depth Analysis on English–German and German–English

Maha Elbayad[1,2] **Michael Ustaszewski**[3] **Emmanuelle Esperança-Rodier**[1]
Francis Brunet Manquat[1] **Jakob Verbeek**[4] **Laurent Besacier**[1]
[1]LIG - Université Grenoble Alpes, France [2]Inria - Grenoble, France
[3] University of Innsbruck, Department of Translation Studies [4] Facebook AI Research

Abstract

We conduct in this work an evaluation study comparing offline and online neural machine translation architectures. Two sequence-to-sequence models: convolutional Pervasive Attention (Elbayad et al., 2018) and attention-based Transformer (Vaswani et al., 2017) are considered. We investigate, for both architectures, the impact of online decoding constraints on the translation quality through a carefully designed human evaluation on English-German and German-English language pairs, the latter being particularly sensitive to latency constraints. The evaluation results allow us to identify the strengths and shortcomings of each model when we shift to the online setup.

1 Introduction

Sequence-to-Sequence models are state-of-the-art in a variety of sequence transduction tasks including machine translation (MT). The most widespread models are composed of an encoder that reads the entire source sequence, while a decoder (often equipped with an attention mechanism) iteratively produces the next target token given the full source and the decoded prefix. Aside from the conventional *offline* use case, recent works adapt sequence-to-sequence models for *online* (also referred to as *simultaneous*) decoding with low-latency constraints (Gu et al., 2017; Dalvi et al., 2018; Ma et al., 2019; Arivazhagan et al., 2019). Online decoding is desirable for applications such as real-time speech-to-speech interpretation. In such scenarios, the decoding process starts before the entire input sequence is available, and online prediction generally comes at the cost of reduced translation quality.

In this work we focus on online neural machine translation (NMT) with deterministic *wait-k* decoding policies (Dalvi et al., 2018; Ma et al., 2019). With such a policy, we first read k tokens from the source then alternate between producing a target token and reading another source token (see Figure 1). We consider two sequence-to-sequence models, a position-based convolutional model and a content-based model with self-attention. We specifically use the recent convolutional Pervasive Attention (Elbayad et al., 2018) and Transformer (Vaswani et al., 2017). We investigate, for both architectures, the impact of online decoding constraints on the translation quality through a carefully designed human evaluation on English→German and German→English language pairs.

Our contributions are twofold: (1) our work, to the best of our knowledge, is the first human evaluation of online *vs*. offline NMT systems. (2) We compare Transformer and Pervasive Attention architectures highlighting the advantages and shortcomings of each when we shift to the online setup. The rest of this paper is organized as follows: we present in §2 related work pertaining to online MT and error analysis of NMT systems. We describe our experimental setup for human evaluation and error analysis in §3. We follow with the evaluation results in §4 and summarize our findings in §5.

2 Related work

2.1 Online NMT

After pioneering works on online statistical MT (SMT) (Fügen et al., 2007; Yarmohammadi et al., 2013; He et al., 2015; Grissom II et al., 2014; Oda et al., 2015), one of the early works with attention-based online translation is Cho and Esipova (2016) using manually designed criteria that dictate whether the

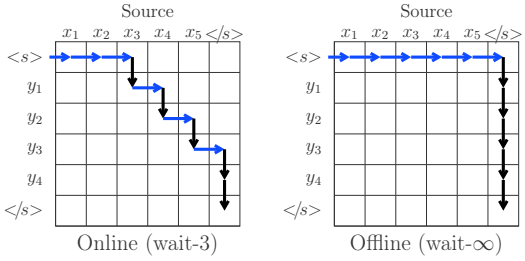


Figure 1: Wait-k decoding as a sequence of reads (horizontal) and writes (vertical) over a source-target grid. After first reading k tokens, the decoder alternates between reads and writes. In Wait-∞, or Wait-until-End (WUE), the entire source is read first.

model should make a read/write operation. Dalvi et al. (2018) proposed a deterministic decoding algorithm that starts with k read operations then alternates between blocks of l write/read operations. This simple approach outperforms the information based criteria of Cho and Esipova (2016), and allows complete control of the translation delay. Ma et al. (2019) trained Transformer models (Vaswani et al., 2017) with a *wait-k* decoding policy that first reads k source tokens then alternate single read-writes. For *dynamic* online decoding, Luo et al. (2017) and Gu et al. (2017) rely on Reinforcement Learning to optimize a read/write policy. To combine the end-to-end training of *wait-k* models with the flexibility of dynamic online decoding, Zheng et al. (2019b) and Zheng et al. (2019a) use Imitation Learning. Recent work on dynamic online translation use monotonic alignments (Raffel et al., 2017) with either a limited or infinite lookback (Chiu and Raffel, 2018; Arivazhagan et al., 2019; Ma et al., 2020). Another adjacent research direction enables revision during online translation to alleviate decoding constraints (Niehues et al., 2016; Zheng et al., 2020; Arivazhagan et al., 2020). In this work, we focus on *wait-k* and greedy decoding strategies, but unlike other *wait-k* models (Ma et al., 2019; Zheng et al., 2019b; Zheng et al., 2019a) we opt for uni-directional encoders which are efficient to train in an online setup (Elbayad et al., 2020).

2.2 Error analysis for NMT

With the advances in NMT (Bahdanau et al., 2015; Vaswani et al., 2017), the quality of translations has improved substantially leading to claims of human parity in high-resource settings (Wu et al., 2016; Hassan et al., 2018). With such improvements, it becomes more and more difficult for automatic evaluation metrics such as BLEU (Papineni et al., 2002) to detect subtle differences. Manual error annotation is a more instructive quality assessment to gain insights into the performance of MT systems, especially in direct comparisons. Human evaluation might lead to conclusions at odd with automatic metrics, as was the case in last year's WMT English-German evaluation (Barrault et al., 2019).

Comparison to SMT and rule-based MT. Bentivogli et al. (2016) studied post-editing of English-German TED talks and found that NMT makes considerably less word order errors than SMT. They also observed that the performance of NMT degrades faster than SMT with increasing sentence length. Toral and Sánchez-Cartagena (2017) reached similar conclusions on news stories in 9 language directions. Isabelle et al. (2017) tested NMT systems with *challenging* linguistic material and highlighted the efficiency of NMT systems at handling subject-verb agreement and syntactic and lexico-syntactic divergences and the struggle of NMT with idiomatic phrases. Castilho et al. (2017a), Castilho et al. (2017b) and Van Brussel et al. (2018) observed that NMT outperforms SMT in terms of fluency, but at the same time it is more prone to accuracy errors. Klubička et al. (2018) made similar observations in an evaluation of English-Croatian, concluding that compared to SMT and rule-based MT, NMT tends to sacrifice completeness of translations in order to increase fluency.

Error typologies for MT. Various error typologies with different levels of granularity have been proposed to evaluate MT systems (Flanagan, 1994; Vilar et al., 2006; Stymne and Ahrenberg, 2012; Lommel et al., 2014b). In their evaluation of SMT outputs, Vilar et al. (2006) defined five error categories: missing words, word order, incorrect words, unknown words and punctuation errors. Bentivogli et al. (2016) followed a simpler classification with three types of errors: morphological, lexical, and word order. Their choice was motivated by the difficulty to disambiguate sub-categories of lexical errors (Popović and Ney,

2011). The evaluation in Klubička et al. (2018) is based on the Multidimensional Quality Metrics (MQM) framework (Lommel et al., 2014b). In their study, they found that mistranslation is the most frequent accuracy error in NMT translations. Van Brussel et al. (2018) observed that mistranslation and omission errors are particularly challenging for NMT users, because contrary to SMT and rule-based MT, these errors are often not indicated by flawed fluency, which makes them more difficult to identify and post-edit.

Error analysis for online MT. Hamon et al. (2009) evaluated a spoken language translation system (ASR+MT) in comparison to a human interpreter, where each segment is judged in terms of adequacy and fluency. Mieno et al. (2015), in search for a unique evaluation metric, examined the usefulness of delay and accuracy in predicting the human judgment of a simultaneous speech translation system. To our knowledge, our work is the first to propose a fine-grained human evaluation of online NMT systems. It focuses on English→German and German→English language pairs, the latter being particularly sensitive to latency constraints. This is in part due to German sentence-final structures (*e.g.* verbs in subordinate clauses) that require long-distance reordering in translation into syntactically divergent languages.

3 Experimental setup

In this work we train Transformer (Vaswani et al., 2017) and Pervasive Attention (Elbayad et al., 2018) models for the tasks of online and offline translation. Following Elbayad et al. (2020), we use unidirectional encoders and train the online MT models with $k_{\text{train}} = 7$, proven to yield better translations across the latency spectrum. We train our models on IWSLT'14 De→En (German→English) and En→De (English→German) datasets (Cettolo et al., 2014). Sentences longer than 175 words and pairs with length-ratio exceeding 1.5 are removed. The training set consists of 160K pairs with 7283 held out for development and the test set has 6750 pairs from TED dev2010+tst2010-2013. All data is tokenized using the standard scripts from the Moses toolkit (Koehn et al., 2007). Unlike existing work experimenting with this dataset, we did not lowercase the bitexts so that we can correctly assess typography errors in German. We segment sequences using byte pair encoding (Sennrich et al., 2016), BPE for short, on the bi-texts resulting in a shared vocabulary of 32K types. We train Pervasive Attention (PA) with 14 layers and 7-wide filters and Transformer (TF) *small* for offline and online translation. We evaluated our *wait-k* models with $k_{\text{eval}} = 3$ achieving a low latency of AL $\in [2.5, 3.5]$ (see Table 1). TF models have 2M more parameters compared to PA (19M to 17M), they are however faster to train (PA is 8 times slower). In test time, the two models decode in comparable speeds. For a fair comparison, both online and offline models are decoded greedily. We will refer these four models with PA-offline, PA-online, TF-offline and TF-online.

3.1 Analysis factors

In this section, we describe the factors we use to analyze the results of automatic and human evaluations.

Source length. Similar to other evaluation studies of NMT systems (Bentivogli et al., 2016; Toral and Sánchez-Cartagena, 2017; Koehn and Knowles, 2017), we look into the length of the source sequence and its effect on the quality of translation.

Lagging difficulty (LD). In the particular context of online translation, source-target alignments are an indicator of how *easy* it is to translate an input.

To measure the lagging difficulty of a pair $(\boldsymbol{x}, \boldsymbol{y})$, we first estimate source-target alignments with `fast-align` (Dyer et al., 2013) and then infer a reference decoding path. he reference decoding path, denoted with $\boldsymbol{z}^{\text{align}}$, is non-decreasing and guarantees that at a given decoding position t, z_t is larger than or equal to all the source positions aligned with t. The lagging difficulty is finally measured as the Average Lagging (AL) (Ma et al., 2019) of the parsed $\boldsymbol{z}^{\text{align}}$ as follows:

$$\text{LD}(\boldsymbol{x}, \boldsymbol{y}) = \frac{1}{\tau} \sum_{t=1}^{\tau} z_t^{\text{align}} - \frac{|\boldsymbol{x}|}{|\boldsymbol{y}|}(t-1), \quad \tau = \arg\min_t \{t \mid z_t = |\boldsymbol{x}|\}. \tag{1}$$

AL measures the lag in tokens behind the ideal simultaneous policy wait-0, and so, LD measures the lag of a *realistic* simultaneous translation that has the aligned context available when decoding. The higher LD, the more challenging it is to constrain the latency of the translation.

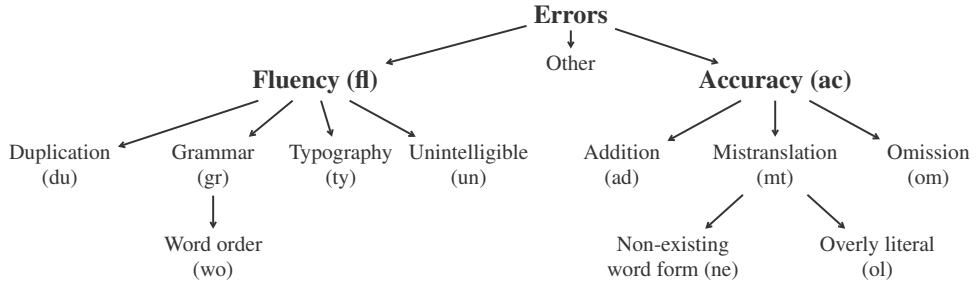


Figure 2: MQM-based error typology used for our manual annotation.

Relying on alignments to assess a pair's difficulty is, however, not ideal; Sridhar et al. (2013) experienced poor accuracies in streaming speech translation when segmenting the input based on alignments and Grissom II et al. (2014) argued that the translator can accurately predict future words from a partial context and consequently beat the alignment-induced latency.

Relative positions. We look into the correlation between the relative positions, source-side and target-side, with the translation's quality. An annotated token $\tilde{y}_t$ of the system's hypothesis $\tilde{\boldsymbol{y}}$ has a target-side relative position $t/|\boldsymbol{y}|$. Similarly, an annotated source token x_j has a relative position $j/|\boldsymbol{x}|$. We argue that with *wait-k* decoding policies, the position of the token might be a contributing factor to the adequacy/fluency of the translation.

3.2 Human evaluation

In addition to the use of automatic evaluation metrics, we conduct an in-depth manual analysis to compare the quality of the output produced by the four systems. From the full test sets, we sample 200 segments in each translation direction. First, we only keep segments whose source sentence lengths fall between the first and third length quartiles. We then remove segments that contain the <unk> token (out-of-vocabulary) and bin the segments by lagging difficulty (see §3.1). We sample from the binned segments to cover all ranges of difficulty and manually remove misaligned segments. Subsequently, the sampled segments were manually error-annotated by a total of four human annotators.

Error typology. For error annotation, a subset of the MQM error typology (Lommel et al., 2014b) was used. A pilot annotation based on 50 translation segments not included into the present test data was carried out to select MQM error types relevant to this study. In this way, the typology was kept to a manageable size to avoid annotators' cognitive overload.[1] The resulting error typology comprises 13 error types, grouped into three major branches: *accuracy*, *fluency* and *other*, as shown in Figure 2. The error type *non-existing word form* was added to the typology to capture target words *invented* with BPE tokenization that do not exist in the target language, such as translating *Pfadfinder* (German for *scout*) into *Badfinder*. The hierarchical nature of the error typology enables annotation and quality analyses on various levels of granularity; annotators were requested to give preference to more specific error types (*i.e.* located deeper in the hierarchy) whenever possible.

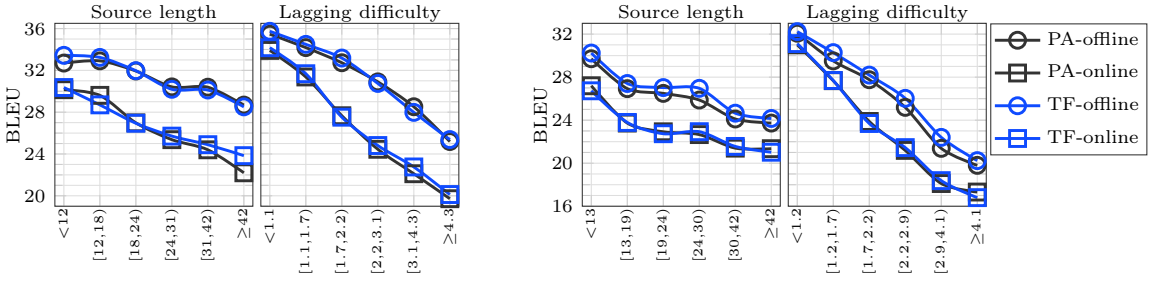
Annotation interface. In order to annotate the segments, we use ACCOLÉ, an online collaborative platform for error annotation (Esperança-Rodier et al., 2019). ACCOLÉ offers a range of services that allow simplified management of corpora and error typologies with the possibility to specify the error typology and search for a particular error type in the annotations. Annotators are tasked to label translation errors by locating the appropriate spans in the target and source segments.

Selection and training of annotators. Per language pair, two annotators with native proficiency in the respective target and near-native to native proficiency in the source language were recruited for the manual annotation task. Following the MQM guidelines and recommendations from NMT evaluation studies (Läubli et al., 2020), the annotators are professional translators. Two of the annotators (one per language pair) also teach in a translation degree programme at university and have thus considerable experience in linguistic translation analysis. To familiarize the annotators with the annotation scheme and

[1] `http://www.qt21.eu/downloads/MQM-usage-guidelines.pdf`, § 2.1

	De→En						En→De					
	PA			TF			PA			TF		
	Offline	Online	Δ%	Offline	Online	Δ%	Offline	Online	Δ%	Offline	Online	Δ%
↑BLEU	**31.24**	26.44	-15	31.13	**26.57**	-15	26.03	**23.04**	-11	**26.60**	22.98	-14
↑METEOR	28.95	**25.97**	-10	**29.25**	25.65	-12	38.81	**35.72**	-8	**39.37**	35.35	-10
↓TER	0.564	**0.621**	+10	**0.555**	0.637	+16	0.626	**0.676**	+8	**0.622**	0.692	+11
↑ROUGE-L	62.89	59.30	-6	**63.15**	**59.51**	-6	57.99	55.41	-4	**58.27**	**55.46**	-5
↑BERTScore	0.937	0.928	-1	**0.939**	**0.930**	-1	0.856	0.848	-1	**0.858**	0.848	-1
AL	21.10	2.59	-88	21.10	3.16	-85	20.71	3.33	-84	20.71	3.49	-83

Table 1: Automatic evaluation of the full test sets. The better scoring system is in bold, underline indicates that the system is better than its competitor with at least 95% statistical significance.



(a) De→En (b) En→De

Figure 3: Bucketed BLEU scores by source length and by lagging difficulty of the full test set.

the use of ACCOLÉ, they were provided with training materials consisting of a written annotation manual, a description of the error typology and a decision tree to guide the selection of appropriate error types. In addition, annotators practiced the annotation procedure on a calibration set of 30 segments representative of the full test data but not included in the 200 segments to be annotated. Subsequently, annotators were given individual feedback and corrective guidance on their annotations. The training materials are made publicly available for reuse.[2] Annotators were remunerated 220 Euros each.

4 Evaluation results

4.1 Automatic evaluation

For each translation direction, En→De and De→En, we assess the quality of our systems by measuring BLEU (Papineni et al., 2002), METEOR (Lavie and Agarwal, 2007), TER (Snover et al., 2006), ROUGE-L (Lin, 2004) and BERTScore (Zhang et al., 2020). We use the default weights and parameters for METEOR[3] and we report the F1 measure combining BERTScore precision and recall.[4] We test for statistical significance with paired bootstrap resampling (Koehn, 2004) using a sample size of 3000 segments. We report the automatic scores evaluated on IWSLT'14 De↔En test set in Table 1 and Figure 3. For bucketed BLEU scores, we bin the test data based on the lagging difficulty of the pair or the source length and measure corpus-level BLEU in each bin.

We observe that (**1**) In offline translation, TF and PA have a comparable performance on De→En with a slight advantage to TF on all metrics except from BLEU. In the En→De direction, TF widens the gap with PA significantly. When binning En→De by lagging difficulty, PA is outperformed by TF in all ranges of difficulty except from the first *easy* bin. (**2**) As to be expected, online decoding leads to a degradation of the translation quality. The degradation is higher for De→En (5 BLEU points) than for En→De (3 BLEU points), arguably because German uses not only verb-initial but also verb-final constructions depending on clause type, thus posing more latency-related challenges for online translation. (**3**) When switching to online translation, the degradation of PA is narrower on average than the degradation of TF allowing for PA to close the gap with TF in both directions. (**4**) Although the translation quality of the systems

[2] https://github.com/elbayadm/OnlineMT-Evaluation

[3] English: meteor-1.5-wo-en-no_norm-0.85_0.2_0.6_0.75-ex_st_sy_pa-1.0_0.6_0.8_0.6,
German: meteor-1.5-wo-de-no_norm-0.95_1.0_0.55_0.55-ex_st_pa-1.0_0.8_0.2

[4] English: roberta-large_L17_no-idf_version=0.3.0(hug_trans=2.4.1),
German: bert-base-multilingual-cased_L9_no-idf_version=0.3.0(hug_trans=2.4.1).

		De→En						En→De					
		PA			TF			PA			TF		
	System	Offline	Online	Δ%	Offline	Online	Δ%	Offline	Online	Δ%	Offline	Online	Δ%
(ac)	Accuracy	2	1	-50	2	3	+50	0	0	+0	0	1	-
(ad)	Addition	**76**	**143**	+88	95	160	+68	30	66	+120	35	97	+177
(mt)	Mistranslation	**433**	587	+36	457	**572**	+25	245	260	+6	**202**	260	+29
(ne)	Non-existing WF	26	17	-35	**14**	**16**	+14	**39**	58	+49	43	**54**	+26
(om)	Omission	**67**	**113**	+69	96	127	+32	**99**	**74**	-25	126	114	-10
(ol)	Overly litteral	78	95	+22	**52**	**81**	+56	150	179	+19	**113**	**125**	+11
(ac+)	Total accuracy	**682**	**956**	+40	716	959	+34	563	**637**	+13	**519**	651	+25
(fl)	Fluency	17	20	+18	**14**	20	+43	26	**21**	-19	**20**	24	+20
(du)	Duplication	11	32	+191	22	144	+555	5	15	+200	13	71	+446
(gr)	Grammar	57	65	+14	**36**	**34**	-6	198	260	+31	**142**	**222**	+56
(ty)	Typography	41	**42**	+2	**33**	59	+79	52	92	+77	**49**	**78**	+59
(un)	Unintelligible	2	3	+50	2	2	+0	**4**	**8**	+100	11	11	+0
(wo)	Word order	**65**	105	+62	66	**78**	+18	46	85	+85	**37**	**74**	+100
(fl+)	Total fluency	193	**267**	+38	**173**	337	+95	331	481	+45	**272**	480	+76
(ac+fl)	Total	**875**	**1223**	+40	889	1296	+46	894	**1118**	+25	**791**	1131	+43

Table 2: Total number of errors (sum of two annotations) per error type for each system. The system (PA or TF) with less errors is put in bold.

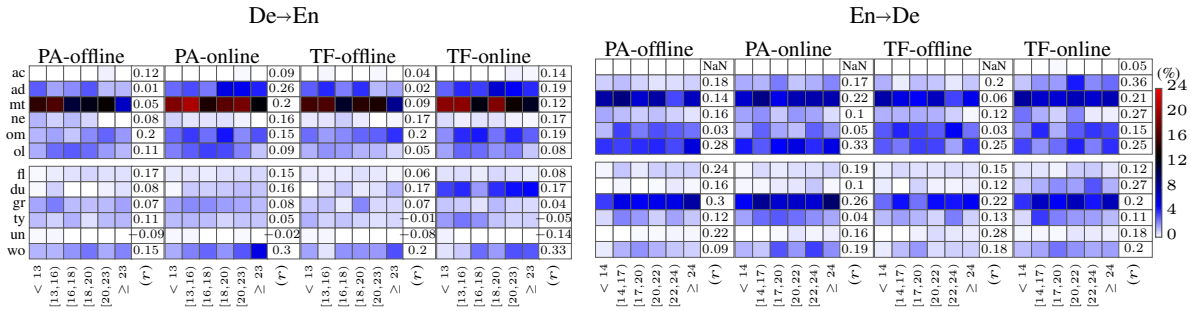
in both directions decreases w.r.t. the length of the source segment, the length is a weaker feature for En→De compared to De→En. Lagging difficulty proves to be a better feature, not only in online translation, but also in offline translation with a steeper decline in BLEU scores as we increase the difficulty.

4.2 Human evaluation

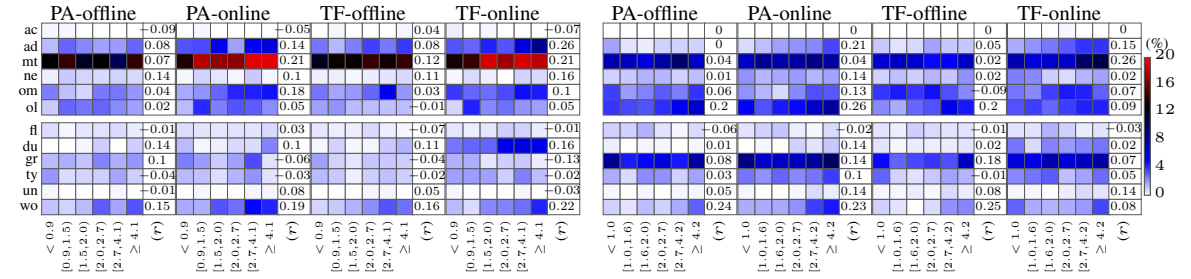
To analyse the annotation data, we rely on the sum count of errors reported by the two annotators. For token-level analysis, we parse the span of each reported error and consider the union of the two annotations to label each output token. To assess the reliability of the error annotation, we measure inter-annotator agreement (IAA) with Cohen's κ (Cohen, 1960) at the token level measuring whether the two annotators agree on the exact error type assigned to each token. We observe an agreement of 0.33 for De→En and 0.40 for En→De which is compatible with other MQM-based evaluation studies (Lommel et al., 2014a; Specia et al., 2017).

For each error type in our typology, we report in Table 2 the count of its occurrences as labeled by two annotators. The frequencies are arranged by task (De→En and En→De), by system (PA and TF) and by decoding setup (offline and online). We observe that **(1)** In alignment with previous works analysing NMT outputs, NMT systems are more prone to accuracy errors than fluency errors (Castilho et al., 2017b; Toral and Sánchez-Cartagena, 2017; Klubička et al., 2018; Van Brussel et al., 2018). **(2)** In accordance with automatic evaluation, the total increase of errors between offline and online (last row of Table 2) is higher for De→En than for En→De and PA is slightly less impacted by the shift from offline to online. This is possibly due to the fact that TF, a context-based model, is more affected by missing context compared to the position-based convolutional PA. **(3)** Unlike automatic evaluation where TF slightly outperforms PA, in three out of the four setups in Table 2, PA has less errors than TF. **(4)** Relative increase of errors between offline and online is larger for fluency than for accuracy, especially for TF. **(5)** Relative increase of errors is particularly high for addition (ad), word order (wo) and duplication (du)[5], the latter being even more problematic for TF. **(6)** In line with other error-annotation studies (Klubička et al., 2018; Van Brussel et al., 2018; Specia et al., 2017), mistranslation (mt) is the largest contributor to accuracy errors. Annotating mistranslations is particularly ambiguous leading to a lower inter-annotator agreement. **(7)** Offline errors with the most consistent gap between PA and TF are duplication and omission in favor of PA and grammar and overly literal in favor of TF. **(8)** More grammar errors are found for En→De compared to De→En, one reason might be that German morphology is richer and more complex than in English, leading to more possibilities for a system to make grammar errors. **(9)** Typography errors are more prevalent in En→De and are highly impacted by latency constraints: most of these typography errors are incorrect

[5]Unwarranted content duplications were marked as duplication errors if fluency was affected or as addition errors otherwise, see annotation guidelines available at `https://github.com/elbayadm/OnlineMT-Evaluation`.



(a) Source length ($|\boldsymbol{x}|$)



(b) Lagging difficulty (LD($\boldsymbol{x}, \boldsymbol{y}$))

Figure 4: Bucketed segment-level count of errors.

punctuation marks (extraneous or missing commas) as well as wrong casing and missing white spaces between compound nouns (producing correct German compounds seems to be especially difficult for online systems). This increase of typography errors suggests that online systems are more literal, as evidenced by the prevalence of incorrect punctuation.

4.3 Fine-grain analysis

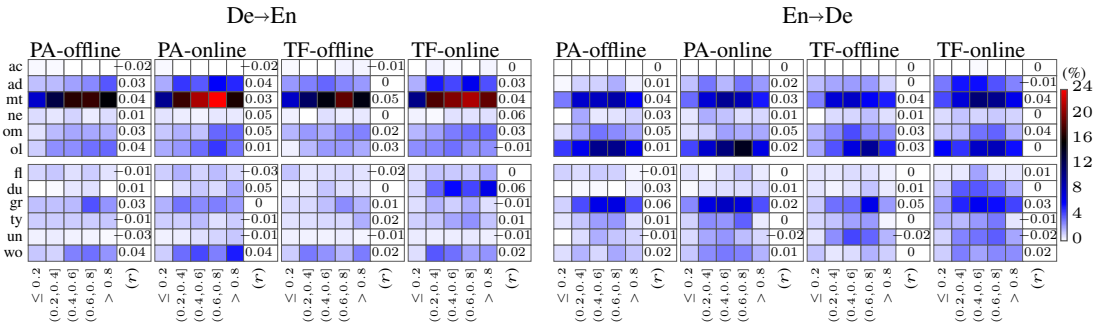
In the following section, we breakdown the annotated set according to the analysis factors (source length ($|\boldsymbol{x}|$), lagging difficulty (LD($\boldsymbol{x}, \boldsymbol{y}$)) and relative positions). For each of these factors, we bin the annotated segments or tokens and report in Figures 4 and 5 normalized counts of errors (divided by the total number of tokens in each bin). An error corresponds to a row of the figure with a heat-map of the normalized counts (read left-to-right with an increasing factor) followed with Pearson's r correlation coefficient.

Source length. In Figure 4a, (**1**) in accordance with the automatic evaluation results of Figure 3, the relative count of errors per length bucket is positively correlated with length. However, for mistranslation (mt) we observe a peak of the relative count of errors in early bins with a decrease of errors for longer segments. This could be attributed to the higher cognitive load in the annotation of longer segments, as it may be easier for annotators to exhaustively label the errors in short segments than in longer ones.[6] The ease of annotating short segments can also be attributed to their fluency; since these segments have fewer fluency errors, labeling mistranslated words is less ambiguous. (**2**) Addition (ad) errors in De→En online systems are considerably more correlated with the source length compared to the offline systems (PA: 0.01→0.26 and TF: 0.02→0.19). This shows that the increase in these errors is mostly located in longer segments. (**3**) Omission (om) and duplication (du) errors, more problematic for TF, have a higher correlation with the source length in En→De (du: 0.12→0.27 and om: 0.03→0.15) but not in De→En.

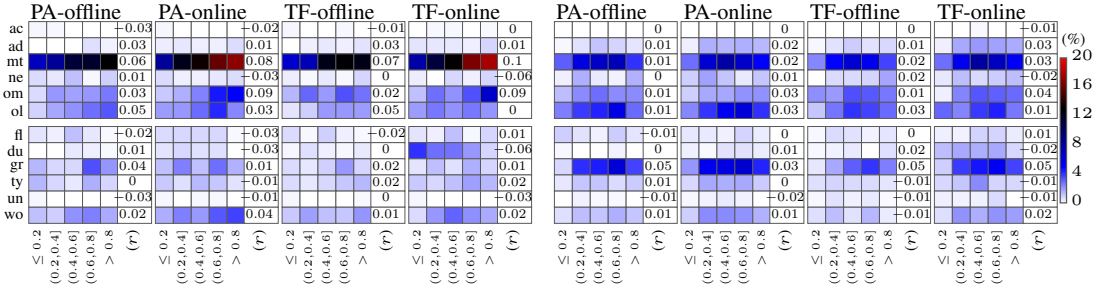
Difficulty. Figure 4b shows that (**1**) even in offline systems, the relative error count is positively correlated with lagging difficulty. This is particularly noticeable for word order (wo) errors. (**2**) For online systems, additions (ad) and omissions (om) are particularly correlated with the lagging difficulty. These errors are the system's solution to deal with missing context. (**3**) Although duplication (du) can also be thought of as a solution to missing context, it is more correlated with the source length than with the difficulty.

Relative position. Unlike length and lagging difficulty, analysis of relative position is based on token-wise labels. In Figure 5, we observe that (**1**) in online De→En systems, most omission (om) and mistranslation (mt) errors concern final source positions and occur near the end of the translation. The high prevalence

[6] IAA scores drop in longer segments which leads us to speculate that it is difficult to annotate longer segments.



(a) Target-side relative position ($t/|\tilde{\boldsymbol{y}}|$)



(b) Source-side relative position ($j/|\boldsymbol{x}|$)

Figure 5: Bucketed token-level count of errors.

Example 1	(a)	Source: Es geht darum, die Nuancen der Sprache zu verstehen. PA-offline: It's about understanding the nuances of language.
	(b)	Source: Es geht darum, die Nuancen der Sprache zu(ol) verstehen(mt, wo). PA-online: It's about the nuances of language to(ol) understand(mt, wo).
	(c)	Source: Es geht darum, die Nuancen der Sprache zu verstehen. TF-offline: It's about understanding the nuances of language.
	(d)	Source: Es geht darum, die Nuancen der Sprache zu verstehen(wo). Tf-online: It's about taking(ad) the nuances of language, understanding(wo) the(du) nuances(du).
Example 2	(e)	Source: Und wenn wir dies für Rohdaten machen können warum nicht auch für Inhalte selbst(ol)? PA-offline: And if we can do that for raw data, why not for content itself(ol)?
	(f)	Source: Und wenn wir dies für Rohdaten machen können(om) warum nicht auch für Inhalte selbst?(ol) PA-online: And if we (om) do this for raw data, why not content itself?(ol)
	(g)	Source: Und wenn wir dies für Rohdaten machen können warum nicht auch für Inhalte selbst?(ol) TF-offline: And if we can do this for raw data, why not for content itself(ol)?
	(h)	Source: Und wenn wir dies für Rohdaten machen können warum nicht auch(om) für Inhalte selbst?(ol) TF-online : And if we could do this for raw data, why not do it (om) for content itself?(ol)

Table 3: Example annotations from De→En. Accuracy errors are in red and fluency in blue.

of these error types in final source positions confirms the well-known difficulty of German sentence-final structures. Note that sentence-final errors on the source side (De) do not necessarily lead to sentence-final errors in the target (En), since the structural differences between the two languages require reordering operations. (**2**) Duplication (du) errors in TF, on the other hand, affect initial source positions near the end of the translation. This means that TF circles back to the beginning of the source to pad the length of the hypothesis. (**3**) Compared to De→En, the errors in En→De systems are well spread across the positions. This is likely due to the fact that decoding English (verb-medial) from German (verb-final) is more exposed to the issue of missing context from final positions. (**4**) Having more mistranslation (mt) errors near the end of the hypotheses in De→En is probably due to the ambiguity of mistranslation errors (the most frequent type). An omission followed by an addition is logically interpreted as a mistranslation (see annotation (d) of Table 3 where *taking* could also be labeled a mistranslation).

4.4 Agreement of the automatic metrics with human annotation

In Table 4 we evaluate the correlation (Pearson's r) between the automatic metrics and the human judgement by proxy of the error count. Unlike in §4.1 where we evaluate corpus-level BLEU, in this

	PA-offline					PA-online					TF-offline					TF-online				
	T	sB	M	B	p	T	sB	M	B	p	T	sB	M	B	p	T	sB	M	B	p
De→En																				
ad	0.26	-0.16	-0.19	-0.28	**-0.42**	**0.38**	-0.28	-0.27	**-0.35**	**-0.40**	**0.43**	-0.28	-0.28	**-0.33**	**-0.46**	**0.39**	**-0.33**	-0.26	**-0.38**	**-0.41**
mt	**0.36**	**-0.41**	**-0.36**	**-0.54**	**-0.45**	**0.41**	**-0.42**	**-0.42**	**-0.51**	**-0.46**	**0.42**	**-0.41**	**-0.44**	**-0.60**	**-0.54**	**0.40**	**-0.41**	**-0.40**	**-0.54**	**-0.50**
ne	0.09	-0.06	-0.07	-0.04	-0.17	0.13	-0.04	-0.10	-0.09	-0.16	0.10	-0.08	-0.07	-0.10	-0.04	**0.30**	-0.23	-0.19	-0.21	-0.21
om	0.10	-0.11	-0.14	-0.21	-0.20	0.06	-0.11	-0.06	-0.24	-0.10	0.07	-0.13	-0.08	-0.10	-0.14	0.06	-0.11	-0.07	-0.21	-0.10
ol	0.12	-0.12	-0.16	-0.07	-0.07	0.12	-0.11	-0.12	-0.15	-0.10	0.14	-0.12	-0.17	-0.15	-0.15	0.26	-0.20	-0.22	-0.23	-0.20
ac+	**0.43**	**-0.43**	**-0.44**	**-0.62**	**-0.58**	**0.49**	**-0.48**	**-0.46**	**-0.63**	**-0.56**	**0.53**	**-0.48**	**-0.49**	**-0.63**	**-0.66**	**0.50**	**-0.48**	**-0.43**	**-0.62**	**-0.58**
fl	-0.01	-0.01	0.05	0.02	0.06	0.03	-0.04	0.02	-0.05	-0.00	0.05	-0.10	-0.05	-0.08	-0.01	-0.00	-0.04	-0.01	0.02	0.00
du	0.09	-0.06	-0.07	-0.04	-0.17	0.13	-0.04	-0.10	-0.09	-0.16	0.10	-0.08	-0.07	-0.10	-0.04	**0.30**	-0.23	-0.19	-0.21	-0.21
gr	0.08	-0.12	-0.08	-0.08	-0.01	0.02	-0.11	0.04	-0.09	-0.10	0.04	-0.11	-0.02	-0.09	-0.05	-0.02	-0.04	0.04	-0.03	-0.10
ty	-0.01	-0.06	0.04	0.02	0.02	0.05	-0.06	-0.04	-0.05	0.09	-0.04	0.02	0.01	-0.01	0.11	0.07	-0.08	-0.15	-0.04	0.04
wo	0.07	-0.05	-0.06	-0.09	-0.10	0.13	-0.10	0.07	-0.17	0.01	0.17	-0.19	-0.17	-0.17	-0.03	0.11	-0.09	-0.10	-0.13	-0.02
fl+	0.11	-0.14	-0.11	-0.11	-0.11	0.20	-0.17	-0.14	-0.23	-0.09	0.17	-0.21	-0.16	-0.20	-0.02	**0.30**	-0.24	-0.24	-0.24	-0.18
ac+fl	**0.41**	**-0.43**	**-0.42**	**-0.58**	**-0.54**	**0.50**	**-0.48**	**-0.45**	**-0.64**	**-0.51**	**0.54**	**-0.51**	**-0.50**	**-0.64**	**-0.59**	**0.54**	**-0.50**	**-0.46**	**-0.61**	**-0.55**
En→De																				
ad	0.06	-0.11	-0.10	-0.14	**-0.23**	0.22	-0.24	-0.25	-0.23	**-0.30**	0.18	-0.09	-0.14	-0.19	**-0.34**	0.22	-0.24	-0.21	-0.25	**-0.48**
mt	0.14	-0.27	-0.21	-0.28	**-0.43**	0.05	-0.24	-0.15	-0.20	**-0.56**	0.08	-0.22	-0.17	-0.18	**-0.39**	0.22	-0.26	-0.26	-0.31	**-0.44**
ne	-0.02	0.03	0.03	-0.01	0.02	0.17	-0.13	0.09	-0.06	-0.14	0.26	-0.21	-0.25	-0.19	-0.27	0.25	-0.18	-0.12	-0.11	-0.28
om	0.07	-0.18	-0.15	-0.20	**-0.34**	0.17	-0.25	-0.24	-0.25	**-0.35**	0.11	-0.21	-0.15	-0.16	-0.22	0.10	-0.25	-0.17	-0.24	**-0.40**
ol	0.17	**-0.31**	-0.22	**-0.27**	-0.25	0.18	-0.28	-0.24	**-0.34**	-0.27	0.12	-0.18	-0.11	-0.15	-0.10	0.11	-0.16	-0.16	-0.26	-0.24
ac+	0.20	**-0.39**	-0.29	**-0.41**	**-0.58**	0.21	**-0.40**	**-0.32**	**-0.41**	**-0.66**	0.20	**-0.35**	-0.26	**-0.32**	**-0.51**	0.25	**-0.36**	**-0.33**	**-0.43**	**-0.66**
fl	0.00	-0.01	-0.02	-0.03	-0.13	0.02	-0.08	-0.06	-0.12	-0.12	0.01	-0.04	-0.00	-0.03	-0.10	-0.07	0.06	0.09	0.08	0.00
du	-0.02	0.03	0.03	-0.01	0.02	0.17	-0.13	0.09	-0.06	-0.14	0.26	-0.21	-0.25	-0.19	-0.27	0.25	-0.18	-0.12	-0.11	-0.28
gr	-0.00	-0.13	-0.03	-0.04	-0.12	0.03	-0.19	-0.11	-0.10	-0.28	0.01	-0.08	-0.03	-0.05	-0.01	0.11	-0.19	-0.10	-0.08	-0.17
ty	0.10	-0.06	-0.04	-0.09	-0.07	-0.01	0.04	0.07	0.03	0.00	0.09	-0.08	-0.06	-0.05	-0.07	-0.00	-0.02	0.03	0.02	-0.06
wo	0.23	-0.24	-0.21	-0.26	-0.13	0.04	-0.14	-0.08	-0.17	-0.18	0.25	-0.18	-0.18	-0.26	-0.09	0.19	-0.21	-0.18	-0.20	-0.20
fl+	0.11	-0.20	-0.11	-0.16	-0.19	0.08	-0.22	-0.13	-0.18	**-0.32**	0.17	-0.19	-0.15	-0.19	-0.16	0.23	-0.27	-0.16	-0.17	**-0.31**
ac+fl	0.21	**-0.39**	-0.28	**-0.39**	**-0.53**	0.18	**-0.39**	-0.29	**-0.38**	**-0.63**	0.24	**-0.35**	-0.27	**-0.33**	**-0.44**	**0.30**	**-0.39**	**-0.31**	**-0.38**	**-0.62**

Table 4: Correlation of the automatic metrics (T: TER, sB: SentBLEU, M: METEOR, B: BERTScore, p: the model's confidence $p(\tilde{\boldsymbol{y}}|\boldsymbol{x})$) with the error counts (Pearson's r). Correlations with $|r| \geq 0.3$ in bold.

section we use sentence-level BLEU denoted with SentBLEU. We highlight the following observations: **(1)** On average, the model's confidence ($p(\tilde{\boldsymbol{y}}|\boldsymbol{x})$) and BERTScore have the highest correlation with the annotated errors in De→En. In En→De BERTScore is less correlated with the human evaluation, this is likely due to the use of a smaller German model (*base* instead of the English *large*). The correlation is higher for accuracy errors. **(2)** With its high correlation with errors, the confidence might be used to efficiently decide when to read and when to write instead of following a deterministic decoding path such as *wait-k*. (Cho and Esipova, 2016; Liu et al., 2020) **(3)** TER (the normalized number of edits required to get from the hypothesis to the reference) is a better indicator of quality than BLEU and METEOR in online systems. It is particularly correlated with addition (ad) and duplication (du) errors frequent online. **(4)** Although omissions (om) in En→De are negatively correlated with confidence and can thus be avoided with a well tuned decoding algorithm, this is not the case for De→En (verb-final to verb-medial). This means that the model omits tokens with a high confidence and is unable to predict that context is missing from the source.

5 Conclusion

We have conducted an evaluation of offline and online NMT systems for spoken language translation. Our aim was to shed light on the strengths and weaknesses of *wait-k* decoding under two different architectures, Transformer and Pervasive Attention. We found that Transformer models are strongly affected by the shift to online decoding with a significant increase in fluency errors, most of which are duplications. PA on the other hand accrues less degradation in online decoding. Our error analysis shows that translation quality in online models can be potentially improved by making read/write decisions based on the model's confidence in order to filter out avoidable additions, mistranslations and duplications. The syntactic asymmetry between German and English remains a challenge for deterministic online decoding. A more detailed analysis of syntactically required long-distance reorderings is left for future work. In this regard, indicators such as lagging difficulty or relative position are more informative for online translation. Another line of future work is to assess to what extent these findings carry over to ASR outputs.

Acknowledgements. We would like to thank the annotators for their valuable contributions. This work partially results from a research collaboration co-financed by the University of Innsbruck's France Focus.

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 *Proc. of ACL*.

Naveen Arivazhagan, Colin Cherry, Wolfgang Macherey, and George Foster. 2020. Re-translation versus streaming for simultaneous translation. In *Proc. of IWSLT*.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *Proc. of ICLR*.

Loïc Barrault, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, Christof Monz, Mathias Müller, Santanu Pal, Matt Post, and Marcos Zampieri. 2019. Findings of the 2019 conference on machine translation (WMT19). In *Proc. of WMT*.

Luisa Bentivogli, Arianna Bisazza, Mauro Cettolo, and Marcello Federico. 2016. Neural versus phrase-based machine translation quality: a case study. In *Proc. of EMNLP*.

Ondřej Bojar, Miloš Ercegovčević, Martin Popel, and Omar Zaidan. 2011. A grain of salt for the WMT manual evaluation. In *Proc. of WMT*.

Sheila Castilho, Joss Moorkens, Federico Gaspari, Iacer Calixto, John Tinsley, and Andy Way. 2017a. Is neural machine translation the new state of the art? *The Prague Bulletin of Mathematical Linguistics*.

Sheila Castilho, Joss Moorkens, Federico Gaspari, Rico Sennrich, Vilelmini Sosoni, Yota Georgakopoulou, Pintu Lohar, Andy Way, Antonio Barone, and Maria Gialama. 2017b. A comparative quality evaluation of pbsmt and nmt using professional translators. *Proceedings of Machine Translation Summit XVI, Nagoya, Japan*.

Mauro Cettolo, Jan Niehues, Sebastian Stüker, Luisa Bentivogli, and Marcello Federico. 2014. Report on the 11th IWSLT evaluation campaign. In *Proc. of IWSLT*.

Chung-Cheng Chiu and Colin Raffel. 2018. Monotonic chunkwise attention. In *Proc. of ICLR*.

Kyunghyun Cho and Masha Esipova. 2016. Can neural machine translation do simultaneous translation? *ArXiv preprint*.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*.

Fahim Dalvi, Nadir Durrani, Hassan Sajjad, and Stephan Vogel. 2018. Incremental decoding and training methods for simultaneous translation in neural machine translation. In *Proc. of NAACL-HLT*.

Chris Dyer, Victor Chahuneau, and Noah A. Smith. 2013. A simple, fast, and effective reparameterization of IBM model 2. In *Proc. of NAACL-HLT*.

Maha Elbayad, Laurent Besacier, and Jakob Verbeek. 2018. Pervasive attention: 2D convolutional neural networks for sequence-to-sequence prediction. In *Proc. of CoNLL*.

Maha Elbayad, Laurent Besacier, and Jakob Verbeek. 2020. Efficient wait-k models for simultaneous machine translation. In *Proc. of INTERSPEECH*.

Emmanuelle Esperança-Rodier, Francis Brunet-Manquat, and Sophia Eady. 2019. ACCOLÉ: A Collaborative Platform of Error Annotation for Aligned Corpora. In *Translating and the computer 41*.

Giovanni Flammia and Victor Zue. 1995. Empirical evaluation of human performance and agreement in parsing discourse constituents in spoken dialogue. In *Fourth European Conference on Speech Communication and Technology*.

Mary Flanagan. 1994. Error classification for mt evaluation. In *Proc. of AMTA*.

Christian Fügen, Alex Waibel, and Muntsin Kolss. 2007. Simultaneous translation of lectures and speeches. *Machine translation*.

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 *Proc. of EMNLP*.

Jiatao Gu, Graham Neubig, Kyunghyun Cho, and Victor OK Li. 2017. Learning to translate in real-time with neural machine translation. In *Proc. of EACL*.

Olivier Hamon, Christian Fügen, Djamel Mostefa, Victoria Arranz, Muntsin Kolss, Alex Waibel, and Khalid Choukri. 2009. End-to-end evaluation in simultaneous translation. In *Proc. of EACL*.

Hany Hassan, Anthony Aue, Chang Chen, Vishal Chowdhary, Jonathan Clark, Christian Federmann, Xuedong Huang, Marcin Junczys-Dowmunt, William Lewis, Mengnan Li, Shujie Liu, Tie-Yan Liu, Renqian Luo, Arul Menezes, Tao Qin, Frank Seide, Xu Tan, Fei Tian, Lijun Wu, Shuangzhi Wu, Yingce Xia, Dongdong Zhang, Zhirui Zhang, and Ming Zhou. 2018. Achieving human parity on automatic chinese to english news translation. *ArXiv preprint*.

He He, Alvin Grissom II, John Morgan, Jordan Boyd-Graber, and Hal Daumé III. 2015. Syntax-based rewriting for simultaneous machine translation. In *Proc. of EMNLP*.

Pierre Isabelle, Colin Cherry, and George Foster. 2017. A challenge set approach to evaluating machine translation. In *Proc. of EMNLP*.

Filip Klubička, Antonio Toral, and Víctor M Sánchez-Cartagena. 2018. Quantitative fine-grained human evaluation of machine translation systems: a case study on english to croatian. *Machine Translation*.

Philipp Koehn and Rebecca Knowles. 2017. Six challenges for neural machine translation. In *Proc. of WNMT*.

Philipp Koehn, Marcello Federico, Wade Shen, Nicola Bertoldi, Ondrej Bojar, Chris Callison-Burch, Brooke Cowan, Chris Dyer, Hieu Hoang, Richard Zens, et al. 2007. Moses: Open source toolkit for statistical machine translation. In *Proc. of ACL*.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proc. of EMNLP*.

Alon Lavie and Abhaya Agarwal. 2007. METEOR: An automatic metric for MT evaluation with high levels of correlation with human judgments. In *Proc. of the Second Workshop on Statistical Machine Translation*, June.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*.

Danni Liu, Gerasimos Spanakis, and Jan Niehues. 2020. Low-latency sequence-to-sequence speech recognition and translation by partial hypothesis selection. *ArXiv preprint*.

Arle Lommel, Aljoscha Burchardt, Maja Popovic, Kim Harris, Eleftherios Avramidis, and Hans Uszkoreit. 2014a. Using a new analytic measure for the annotation and analysis of mt errors on real data. In *Proc. of EAMT*.

Arle Lommel, Hans Uszkoreit, and Aljoscha Burchardt. 2014b. A framework for declaring and describing translation quality metrics. *Revista Tradumàtica*.

Yuping Luo, Chung-Cheng Chiu, Navdeep Jaitly, and Ilya Sutskever. 2017. Learning online alignments with continuous rewards policy gradient. In *Proc. of ICASSP*.

Samuel Läubli, Sheila Castilho, Graham Neubig, Rico Sennrich, Qinlan Shen, and Antonio Toral. 2020. A set of recommendations for assessing human–machine parity in language translation. *Journal of Artificial Intelligence Research*.

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 *Proc. of ACL*.

Xutai Ma, Juan Pino, James Cross, Liezl Puzon, and Jiatao Gu. 2020. Monotonic multihead attention. In *Proc. of ICLR*.

Takashi Mieno, Graham Neubig, Sakriani Sakti, Tomoki Toda, and Satoshi Nakamura. 2015. Speed or accuracy? a study in evaluation of simultaneous speech translation. In *Proc. of INTERSPEECH*.

Jan Niehues, Thai Son Nguyen, Eunah Cho, Thanh-Le Ha, Kevin Kilgour, Markus Müller, Matthias Sperber, Sebastian Stüker, and Alex Waibel. 2016. Dynamic transcription for low-latency speech translation. In *Proc. of INTERSPEECH*.

Yusuke Oda, Graham Neubig, Sakriani Sakti, Tomoki Toda, and Satoshi Nakamura. 2015. Syntax-based simultaneous translation through prediction of unseen syntactic constituents. In *Proc. of ACL*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proc. of ACL*.

Maja Popović and Hermann Ney. 2011. Towards automatic error analysis of machine translation output. *Computational Linguistics*.

Colin Raffel, Minh-Thang Luong, Peter J Liu, Ron J Weiss, and Douglas Eck. 2017. Online and linear-time attention by enforcing monotonic alignments. In *Proc. of ICML*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proc. of ACL*.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proc. of AMTA*.

Lucia Specia, Kim Harris, Aljoscha Burchardt, Marco Turchi, Matteo Negri, and Inguna Skadina. 2017. Translation quality and productivity: A study on rich morphology languages. In *Machine Translation Summit XVI*.

Vivek Kumar Rangarajan Sridhar, John Chen, Srinivas Bangalore, Andrej Ljolje, and Rathinavelu Chengalvarayan. 2013. Segmentation strategies for streaming speech translation. In *Proc. of NAACL-HLT*.

Sara Stymne and Lars Ahrenberg. 2012. On the practice of error analysis for machine translation evaluation. In *Proc. of LREC*.

Antonio Toral and Víctor M. Sánchez-Cartagena. 2017. A multifaceted evaluation of neural versus phrase-based machine translation for 9 language directions. In *Proc. of EACL*.

Laura Van Brussel, Arda Tezcan, and Lieve Macken. 2018. A fine-grained error analysis of nmt, pbmt and rbmt output for english-to-dutch. In *Proc. of LREC*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proc. of NeurIPS*.

David Vilar, Jia Xu, Luis Fernando D'Haro, and Hermann Ney. 2006. Error analysis of statistical machine translation output. In *Proc. of LREC*.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *ArXiv preprint*.

Mahsa Yarmohammadi, Vivek Kumar Rangarajan Sridhar, Srinivas Bangalore, and Baskaran Sankaran. 2013. Incremental segmentation and decoding strategies for simultaneous translation. In *Proc. of NAACL-HLT*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *Proc. of ICLR*.

Baigong Zheng, Renjie Zheng, Mingbo Ma, and Liang Huang. 2019a. Simpler and faster learning of adaptive policies for simultaneous translation. In *Proc. of EMNLP*.

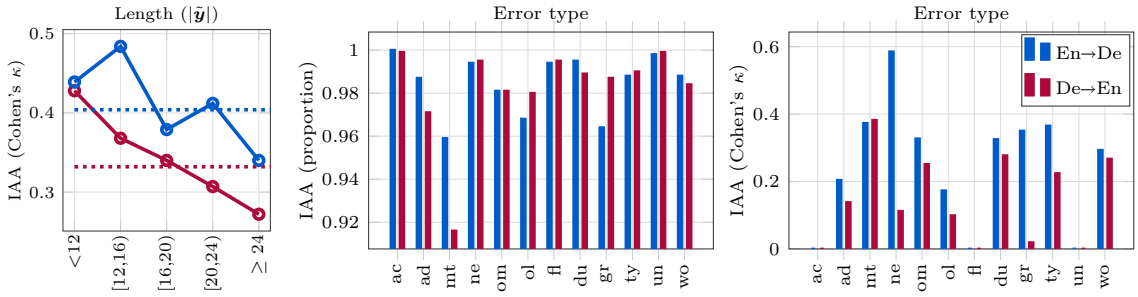
Baigong Zheng, Renjie Zheng, Mingbo Ma, and Liang Huang. 2019b. Simultaneous translation with flexible policy via restricted imitation learning. In *Proc. of ACL*.

Renjie Zheng, Mingbo Ma, Baigong Zheng, Kaibo Liu, and Liang Huang. 2020. Opportunistic decoding with timely correction for simultaneous translation. In *Proc. of ACL*.

A Inter-annotator agreement

Length. To study whether long segments are more challenging to annotate, we measure inter-annotator agreement (IAA) in buckets of source length. Similar to prior studies (Flammia and Zue, 1995; Stymne and Ahrenberg, 2012; Bojar et al., 2011), we found that the length of the sequence has a negative correlation with the agreement possibly because of the increasing cognitive load (see Figure 6a).

Error type. To assess the ambiguity of the error types in our study, we measure binary agreements for each error in the typology. In this setup, we consider the task of annotating each error as a binary classification. Without chance correction (Figure 6b), mistranslation (mt) error is the one with the highest disagreement possibly because of its ambiguity. Agreement on rare errors such as accuracy (ac) and unintelligible (un) is zeroed out after chance correction.



(a) Length - κ (b) Error type - agreement proportion (c) Error type - κ

Figure 6: IAA measured with Cohen's kappa or as agreement proportion without chance correction. The left panel shows the IAA per hypothesis length and the two right panels breakdown the agreement per error type.