You Are What You Train: Effects of Data Composition on Training Context-aware Machine Translation Models

Paweł Mąka and Yusuf Can Semerci and Jan Scholtes and Gerasimos Spanakis
Department of Advanced Computing Sciences
Maastricht University
{pawel.maka, y.semerci, j.scholtes, jerry.spanakis}@maastrichtuniversity.nl

Abstract

Achieving human-level translations requires leveraging context to ensure coherence and handle complex phenomena like pronoun disambiguation. Sparsity of contextually rich examples in the standard training data has been hypothesized as the reason for the difficulty of context utilization. In this work, we systematically validate this claim in both single- and multilingual settings by constructing training datasets with a controlled proportions of contextually relevant examples. We demonstrate a strong association between training data sparsity and model performance confirming sparsity as a key bottleneck. Importantly, we reveal that improvements in one contextual phenomenon do no generalize to others. While we observe some cross-lingual transfer, it is not significantly higher between languages within the same sub-family. Finally, we propose and empirically evaluate two training strategies designed to leverage the available data. These strategies improve context utilization, resulting in accuracy gains of up to 6 and 8 percentage points on the ctxPro evaluation in single- and multilingual settings respectively.[1]

1 Introduction

Context-Aware Machine Translation (MT) models use surrounding sentences (context) to improve translation by maintaining coherence and resolving ambiguities (Agrawal et al., 2018; Bawden et al., 2018; Müller et al., 2018; Voita et al., 2019b). The context can be sentences in the source language and the previously translated sentences in the target language. While many works improved the translation quality of the context-aware MT by applying standard Transformer (Vaswani et al., 2017) model (Sun et al., 2022; Majumde et al., 2022; Gete et al., 2023b; Post and Junczys-Dowmunt, 2024; Alves et al., 2024; Kocmi et al., 2024), specialized architectures (Tu et al., 2017; Bawden et al., 2018; Miculicich et al., 2018; Maruf et al., 2019; Huo et al., 2020; Zheng et al., 2021), and decoder-only LLMs (Alves et al., 2024; Kocmi et al., 2024), the reason why the context utilization is challenging for the models remain an open question.

Figure 1: Composition of the English-to-German training datasets with the Gender phenomenon in Pure IWSLT and IWSLT+OpenSubtitles settings. Annotations are based on ctxPro (Wicks and Post, 2023), and the dashed bars represent the contextually-rich datasets. Note that the horizontal axis starts at 100,000.

The low density of contextually rich (requiring context for correct translation) examples in the training datasets has been suspected as the main reason why MT models have trouble in translating contextual phenomena. For example, Lupo et al. (2022) proposed the two-fold sparsity hypothesis, where the low density of examples in the dataset and the tokens in the examples requiring context increases the difficulty of learning to leverage context. Post and Junczys-Dowmunt (2024) show that sparsity in the evaluation datasets makes it difficult to assess the context utilization of the models. We argue that this also points to the sparsity hypothesis in the training data, as the evaluation datasets are often sampled from the same distribution (the underlying dataset).

[1] https://github.com/Pawel-M/data-composition

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 27414–27437
November 4-9, 2025

In this work, we evaluate how the the proportion of contextually rich examples in the training data of the context-aware MT models affects the overall translation quality measured by BLEU (Papineni et al., 2002) and COMET (Rei et al., 2020), and performance on the examples requiring context (using generative and contrastive evaluations). To this end, we use ctxPro toolset (Wicks and Post, 2023) to extract the relevant examples containing the following phenomena: Gender, Formality, Auxiliary, Inflection, and Animacy. The details of the annotation and phenomena can be found in the original paper (Wicks and Post, 2023) (see Appendix A for short descriptions). We constructed training data by mixing contextually rich and poor examples with varying proportions (Figure 1 illustrates this for Gender in English-to-German). Moreover, we evaluate cross-lingual transfer of context utilization in multilingual models on English-to-X and X-to-English where X is {German, French, Polish, Russian, and Spanish}. Finally, we explore several ways to effectively leverage the available data to obtain models that perform well both generally and in context-sensitive settings. The contributions of this work are:

1. We **empirically validate the sparsity hypothesis**, showing strong relation between the density of the contextual phenomena in the training data and the resulting performance of the context-aware MT models.

2. We **reveal limitations in generalization**, showing that the improvement in one linguistic phenomenon does not transfer to others. We observe limited cross-lingual transfer, not substantially higher between languages in the same subfamily.

3. We propose and empirically evaluate **two training strategies** designed to improve context utilization by leveraging the available data. We show a trade-off between improving context utilization and general translation metrics such as BLEU.

2 Related Work

Through years many dedicated architectures have been proposed for context-aware MT (Miculicich et al., 2018; Voita et al., 2019b,a; Bao et al., 2021; Chen et al., 2022; Feng et al., 2022; Bulatov et al., 2022; Maka et al., 2024) including popular multi-encoder (where a separate encoder is responsible for processing the context sentences; Jean et al., 2017; Miculicich et al., 2018; Maruf et al., 2019; Huo et al., 2020; Zheng et al., 2021), but the standard Transformer model (Vaswani et al., 2017) with the sentences being concatenated (single-encoder; Tiedemann and Scherrer, 2017; Ma et al., 2020; Zhang et al., 2020) exhibited high performance despite its relative simplicity (Majumde et al., 2022; Sun et al., 2022; Gete et al., 2023b; Post and Junczys-Dowmunt, 2023). While decoder-only LLMs have achieved state-of-the-art results in MT (Alves et al., 2024; Kocmi et al., 2024), they require extensive datasets for training, have a large number of parameters, and increased inference time (Pang et al., 2025), which can limit their usefulness in computationally constrained environments. In recent years, research interest in the architectures other than decoder-only has remained relevant (Mohammed and Niculae, 2024; Warner et al., 2024; Alastruey et al., 2024; Azeemi et al., 2025; Marashian et al., 2025). Therefore, we largely focus this paper on encoder-decoder models.

The standard sentence-level metrics (e.g., BLEU (Papineni et al., 2002) do not capture the contextual utilization by the models (Hardmeier, 2012; Wong and Kit, 2012). To address this, several evaluation datasets have been proposed including contrastive (Müller et al., 2018; Bawden et al., 2018; Voita et al., 2019b; Lopes et al., 2020) and generative such as ctxPro (Wicks and Post, 2023) used in this study. Moreover, metrics like CXMI (Fernandes et al., 2021) and PCXMI (Fernandes et al., 2023) can measure how much the model relies on context during translation.

The effects of the training dataset on the final model has also been studied extensively (Kaplan et al., 2020; Hoffmann et al., 2022) in different domains (Alabdulmohsin et al., 2023), including document-level MT (Zhuocheng et al., 2023). The studies mostly concentrated on the scale of the training dataset. We, instead, investigate the composition of the dataset and its effect on the context-aware MT models.

Several works proposed methods increasing contextual capabilities of the models by training the models on annotated data (Jwalapuram et al., 2020; Yin et al., 2021; Gete et al., 2023a; Mąka et al., 2025) but they target only pronoun disambiguation. Fine-tuning in this case can be seen as similar to domain adaptation (Luong and Manning, 2015; Chu et al., 2017) where loss weighting (similar to one of our methods) is an effective strategy (Wang et al., 2017).

3 Effects of Data Composition

We first measured how the presence of contextually rich examples in the training data affects both translation quality and the models' ability to leverage context. To that end, we trained models on datasets whose composition we systematically varied. Specifically, we identified contextual examples (containing relevant phenomena) from the available datasets using ctxPro toolset (Wicks and Post, 2023) and constructed a series of datasets with varying densities of different phenomena. This setup allowed us to assess inter-phenomena as well as cross-lingual effects of the composition of the training datasets. We used three settings: single language pair (English-to-German), and multilingual with encoder-decoder and decoder-only (LLM) models. For the multilingual setting, we used English-to-X and X-to-English language directions, where X is {German, French, Polish, Russian, and Spanish} - a subset of directions covered by the ctxPro. We utilized two Germanic, Romance, and Slavic languages.

3.1 Datasets

We base our research on two document-level translation datasets: IWSLT 2017 English-to-German (Cettolo et al., 2017) and OpenSubtitles 2018 (Lison et al., 2018). For the English-to-German direction, we employ both datasets, and for the multilingual setting, we only use OpenSubtitles. We extract contextual annotations from the training subset of the IWSLT dataset using the ctxPro toolset. The annotated (containing contextually-rich examples) subset forms **IWSLT-dense** dataset, which can be further divided based on the target phenomenon: Gender, Formality, Auxiliary, Inflection, and Animacy. We discard examples containing more than one type of phenomena in any of the sentences. From the remaining examples we form **IWSLT-sparse** dataset of size 123,000, containing no examples annotated with any contextual phenomena. CtxPro released annotations extracted from the OpenSubtitles 2018 dataset divided into *dev*, *devtest*, and *test* subsets. We set aside the *test* subset for the evaluation and used the combined *dev* and *devtest* subsets for training, forming **OS-dense** dataset. The released ctxPro dataset is not exhaustive; therefore, we do not create the sparse version of the OpenSubtitles dataset. Instead, we randomly sample the OpenSubtitles dataset to the desired size (referred to as **OS-random**). It should be noted that OS-random datasets can contain a very limited number of examples from OS-dense datasets (less than 1 per 1000). In Appendix B we present the sizes of the dense component datasets.

To create the training datasets with varying densities of contextually rich examples, we sample and concatenate examples from both dense and sparse datasets to form a training dataset. For English-to-German, we study two settings: *Pure IWSLT* (only IWSLT-sparse and IWSLT-dense datasets) and *IWSLT + OS* (using IWSLT-sparse, IWSLT-dense, and English-to-German OS-rand and OS-dense datasets). These allow us to study two regimes: extremely low sparsity with the first setting, and very dense with the second one. We progressively replace examples from sparse and random datasets with the examples sampled from dense datasets. In the multilingual experiments, we formed the baseline training dataset by sampling 50,000 examples from OS-rand for all language directions we considered. For each phenomenon in a language direction, we formed the enriched datasets by replacing n examples with the examples sampled from the OS-dense dataset corresponding to the phenomenon and language direction. We chose n to be the minimum number of examples (rounded) for a particular phenomenon across language directions maximizing the resulting density of the training datasets while making the results comparable between language directions. We present the illustration of the composition of the datasets in Figure 1 for Gender on English-to-German and further details in Appendix B. To reduce the complexity of the analysis we add only examples containing a single type of phenomenon. Assessing the complex interconnections between phenomena is left for future work.

3.2 Training

For encoder-decoder models, we employed a two-stage training process where first the sentence-level model is trained on more abundant sentence-aligned datasets, followed by the context-aware training on the document-level dataset. Following Mąka et al. (2025), we rely on the publicly available pre-trained sentence-level models, namely *OPUS-MT en-de* (Tiedemann and Thottingal, 2020; Tiedemann et al., 2023) and *No Language Left Behind* (NLLB-200) with 600M parameters (NLLB Team et al., 2022). For LLM-based MT models, we utilize Towerbase 7B model (Alves et al., 2024) which we fine-tune using LoRA (Hu et al., 2022) on

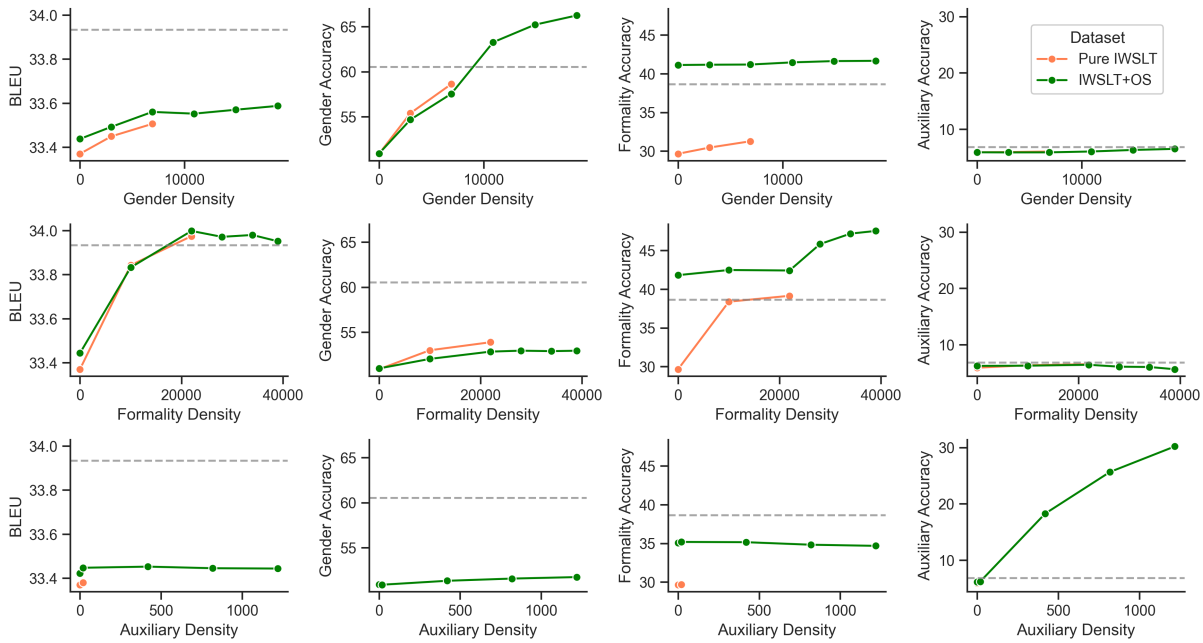

Figure 2: Measured metrics of BLEU on IWSLT 2017 testset, and ctxPro accuracy on Gender, Formality, and Auxiliary phenomena (in columns) of the OpusMT en-de models trained on the datasets with varying amounts of contextually-rich examples of Gender, Formality, and Auxiliary phenomena (in rows). Shows two experimental settings: Pure IWSLT and combined IWSLT+OS.

document-level MT dataset. Because Towerbase models were not pre-trained on Polish language we do not include English-to-Polish and Polish-to-English language pairs in training and evaluation. We concatenate consecutive sentences separated by the special `[SEP]` token in case of encoder-decoder models and `<sep>` string in case of LLMs on both the source and target sides. Similar to Sun et al. (2022), we create examples with all context sizes (number of previous sentences to concatenate) from zero to the maximum context size. We set the maximum context size to three as further increases have shown diminishing returns regarding context utilization (Post and Junczys-Dowmunt, 2023). In Appendix A, we show the number of examples in the ctxPro dataset with antecedent distance inside the context size. During inference, the models receive only the source-side context and generate the target-side context before the current sentence. We obtained the translation of the current sentence by splitting the output on the separator token (for encoder-decoder models) or substring (for LLMs). The training hyper-parameters and additional details can be seen in Appendix C.

3.3 Single Language Pair Results

For the models in the English-to-German experiments, we trained 5 models with different seeds and averaged the results. Apart from the constructed datasets, we also trained a baseline model on the unmodified IWSLT training dataset. To measure the general translation quality, we translated the IWSLT 2017 English-to-German test subset (with BEAM search of 5) and measured BLEU (Papineni et al., 2002). Additionally, we translated test subsets of the ctxPro dataset (based on OpenSubtitles) and measured the accuracy of matching the expected word in the translation (using the scripts provided with the dataset). The results can be seen in Figure 2. Extended results including COMET and ContraPro (Müller et al., 2018) accuracy can be found in Appendix D.

We observed a drop in BLEU for the models trained on the sparse datasets, even for the datasets with mixed OpenSubtitles examples. While the reduction was relatively small (less than 2%), it returned to the baseline value only when Formality IWSLT-dense examples were added to the dataset. This could mean that the examples from the IWSLT dataset annotated with Formality were particularly influential for the model's general translation ability, and mixing in the random examples from OpenSubtitles did not help.

For Gender and Formality, increasing their density in the training dataset improved the ctxPro accuracy for the corresponding phenomenon. No-

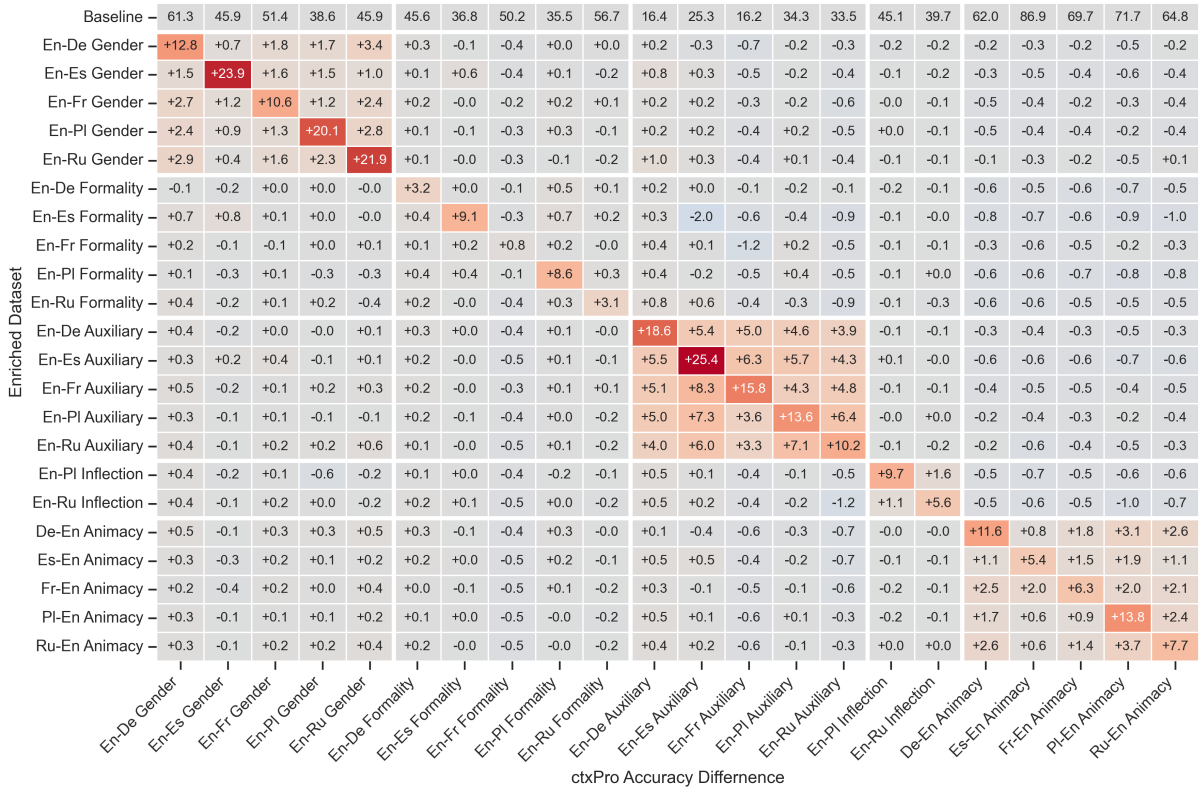

Figure 3: Accuracy on all phenomena for each relevant language direction in ctxPro (in columns) of the NLLB-200 600M models trained on the OpenSubtitles datasets with varying amounts of contextually-rich examples for each phenomenon and language direction (in rows). We show the differences from the baseline model (top row).

tably, Formality in the IWSLT+OS setting only improved when OS-dense examples were added, but exceeded the accuracy of the baseline model even with the most sparse dataset. Adding OS-dense examples improved the accuracy significantly above the baseline (up to 30%). Interestingly, adding dense examples in one phenomenon had minimal effect on the accuracy of the other phenomena, with only a very small increase of Formality for the Gender-enriched dataset and vice versa. Those results show that the generalizability of the models' ability to handle contextual phenomena is very limited. While we argue that experimenting with the publicly-available pre-trained models enhances reproducibility OpusMT was trained on OpenSubtitles dataset on which ctxPro dataset was based. Therefore, we include the results where the weights has been randomly initialized in Appendix D which show the same behavior corroborating our findings.

3.4 Multilingual Results

For the multilingual experiments, we trained models (with a single seed due to the computational cost of training and evaluation) on the composed datasets and measured ctxPro accuracy for all applicable phenomena and language directions included in the experiments. Note that Inflection applies only to English-to-Polish and English-to-Russian, and Animacy only to X-to-English. The results are presented in Figures 3 and 4 for encoder-decoder and decoder-only models respectively. Results in terms of BLEU and COMET on the testsets sampled from OpenSubtitles for each language direction can be seen in Appendix D.

For each model, the highest improvement in accuracy was observed for the phenomenon and language direction that was added to the training dataset (values on the diagonal in the figures). In line with the results on the single language pair, we did not observe any intra-lingual transfer between phenomena. Interestingly, there was some transfer between language directions for the same phenomenon, which was the strongest for Auxiliary, moderate for Gender, Inflection (for encoder-decoder models), and Animacy, and no transfer for Formality. Contrary to our expectations, we did not observe notably stronger transferability between languages in the same linguistic sub-family, with the exception of Auxiliary in encoder-decoder mod-

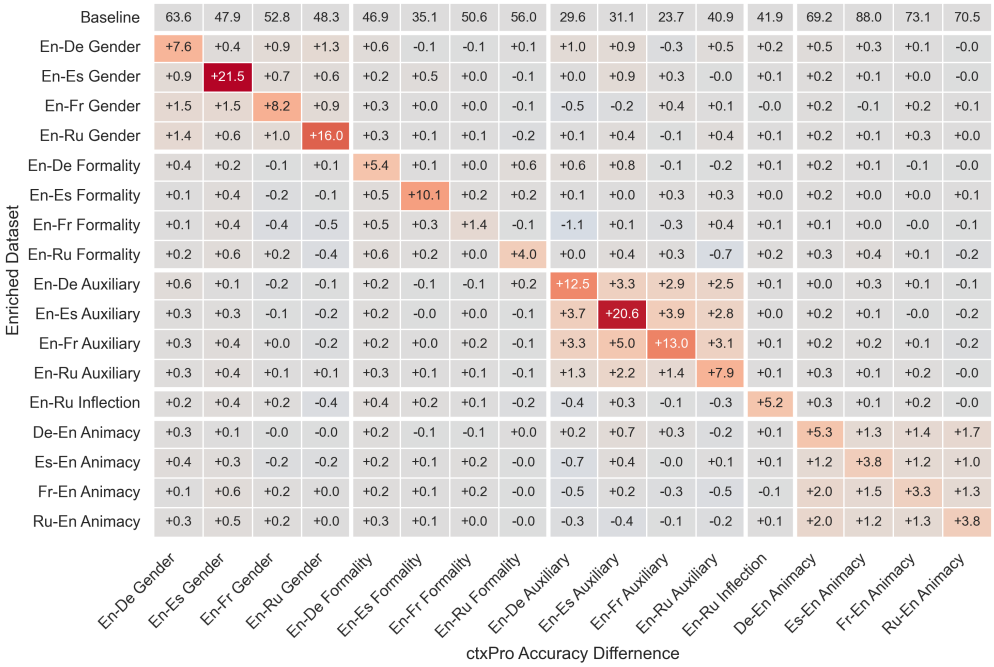

Figure 4: Accuracy on all phenomena for each relevant language direction in ctxPro (in columns) of the Towerbase 7B trained on the OpenSubtitles datasets with varying amounts of contextually-rich examples for each phenomenon and language direction (in rows). We show the differences from the baseline model (top row).

els, where the increase in accuracy is slightly higher inside Romance and Slavic languages than for other languages. Surprisingly, Towerbase did not exhibit higher generalizaility compared to NLLB-200 corroborating the notion that LLMs are a reflection of their training data.

3.5 Discussion

We experimentally confirmed the dataset sparsity hypothesis by showing that the models trained on datasets sparse in contextually rich examples exhibit poor context utilization, and increasing the density leads to large improvements for the tested phenomena. Our experiments showed that the models do not generalize context utilization between phenomena. This finding calls for caution when interpreting the results of evaluations targeting a single phenomenon (Müller et al., 2018; Lopes et al., 2020). While document-level training datasets typically include a representative (for a particular domain) mixture of contextual phenomena, we found that models can develop strong capabilities for some phenomena, while remaining weak on others. Mąka et al. (2025) found attention heads in context-aware MT models responsible for pronoun disambiguation with some cross-lingual behavior, which is in line with the observed transferability between language directions. We hypothesize that the poor transfer between phenomena can be explained by the models developing separate heads for each of them.

4 Methods Exploiting Contextual Data

Inspired by the fact that increased density in contextually-relevant examples of the training dataset leads to improvement in context utilization, we tested several techniques that could leverage the available data more efficiently. We broadly divide them into annotation-based and annotation-free. Annotations can inform the training process but require an external tool (e.g., ctxPro) to mark the relevant examples. A straightforward method is to simply extract the annotated examples from the training dataset and use them to fine-tune the model. Annotation-free methods do not rely on an external tool and have the advantage of generalizability beyond the phenomena covered by any tool. Crucially, the presented methods aim to improve contextual capabilities without the need for any additional data beyond the standard training datasets.

4.1 Token-level Loss Weighting

We adapted the weighting of the loss elements (Wang et al., 2017), which increases the error signal coming from selected examples. Instead of

weighting the whole examples, we apply a token-level approach as phenomena annotations contain an expected word or phrase that requires context for successful translation. We train the models using the weighted negative log-likelihood loss function:

$$\mathcal{L} = -\frac{1}{|D_a|} \sum_{(x_i,y_i,a_i)\sim D_a} \sum_{j=1}^{|y_i|} w(a_{i,j}) log(\hat{y}_{i,j}), \tag{1}$$

where $\hat{y}_{i,j}$ is the probability of the j-th token in i-th example, D_a is the annotated training dataset with examples containing input and output sequences (x_i and y_i respectively), as well as the token-level annotations a_i marking the contextually-dependent tokens, and $w(a_{i,j})$ is defined as:

$$w(a_{i,j}) = \begin{cases} 1+\lambda, & \text{if contextually dependent,} \\ 1, & \text{otherwise,} \end{cases} \tag{2}$$

for each token j in the i-th output sequence, where λ is the hyper-parameter.

4.2 Metric-based Example Selection

A major issue with using annotations is that, according to our experiments on data composition, the model will improve only on the included phenomena. To mitigate this, we propose to utilize the model itself to mark contextually-rich examples. Fernandes et al. (2023) proposed the Point-wise Cross-Mutual Information (PCXMI) metric to measure the context reliance of the translations, which is based on the output probabilities of the context-aware MT model. For a particular example it is calculated as:

$$PCXMI = \sum_{j=1}^{|y|} log \frac{q(y_j|y_{t<j}, x, C)}{q(y_j|y_{t<j}, x)}, \tag{3}$$

where C is the context, and q represents the context-aware MT model (returning token probabilities, noted as $q(y_j|y_{t<j}, x, C)$) that is trained to also be used as a sentence-level model (noted as $q(y_j|y_{t<j}, x)$). We introduce a slightly modified metric that computes the maximum token-level PCXMI for a given example:

$$MaxPCXMI = \max_j \Big(log \frac{q(y_j|y_{t<j}, x, C)}{q(y_j|y_{t<j}, x)} \Big). \tag{4}$$

We motivate it by the fact that an example with even a single token being dependent on context can be considered a contextually-rich example (certainly the case for pronouns), which is better captured by our metric. The proposed method consists of the following steps:

1. **train** the model on context-aware data,
2. **calculate** the metric using the trained model for the examples in the training dataset,
3. **select** top k examples (a hyper-parameter),
4. **fine-tune** the model on the selected subset.

While the method can be seen as similar to curriculum learning (Zhang et al., 2018), we select the examples that the model is already competent at translating using context. Intuitively, this is a positive feedback where the model learns to generalize to the difficult examples by becoming better at what it already knows.

Method	Requires Annotations	Additional Training
Fine-tuning	✓	✓
Adapted D&R	✗	✗
CoWord Dropout	✗	✗
Head-tuning	✓	✓
Weighting	✓	✗
MaxPCXMI	✗	✓

Table 1: Tested methods and whether they require annotated dataset or employ additional fine-tuning.

5 Experiments

We experimentally evaluated Token-level Loss Weighting and Metric-based Example Selection for fine-tuning on encoder-decoder models and compared them to the following baselines (Table 1 summarizes their requirement of annotated dataset and additional training):

- **Fine-tuning** (annotation-based) - simply fine-tuning the model on the annotated data after the context-aware training.
- **CoWord Dropout** (annotation-free; Fernandes et al., 2021) - masking random tokens in the current source sentence to force the model to use context for translation, the probability of masking a token is controlled by the hyper-parameter p.
- **Adapted Divide and Rule** (annotation-free; Lupo et al., 2022) - splitting the current source and target sentences in the middle and appending the first parts to the context. Notably, this method was introduced for the multi-encoder architecture where a separate encoder was used for context sen-

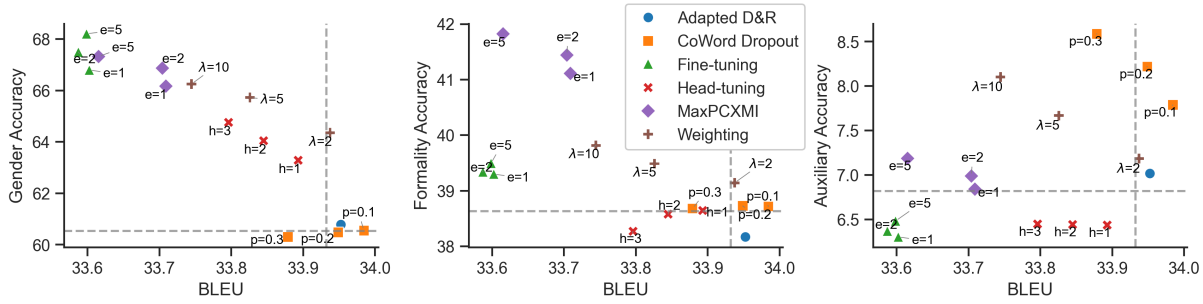

Figure 5: Accuracy of ctxPro English-to-German phenomena against BLEU on the IWSLT 2017 en-de testset of the methods applied to OpusMT en-de model. Labels show: the number of epochs ("e"), CoWord Dropout probability ("p"), number of tuned heads ("h"), and weighting strength ("λ") hyper-parameters.

tences. Contextual parameters were trained only in the second, context-aware phase of training with the rest of the model frozen. We *adapt* it to the single-encoder architectures we use in this study by training the whole model in the context-aware training phase.

- **Head-tuning** (annotation-based; Mąka et al., 2025) - training selected attention heads to attend the context cue, available only for Gender.

We evaluated all methods in the single language pair (English-to-German) setting and annotation-free methods in the encoder-decoder multilingual setting (due to the lack of exhaustive annotations for the dataset; see Table 1). We used the same base sentence-level models: OpusMT en-de and NLLB-200 600M, respectively. For English-to-German, we trained on the full IWSLT 2017 en-de dataset with ctxPro annotations, and for multilingual, we sampled 50,000 examples for each language direction from the OS-rand dataset. We used the same hyper-parameters shared by all methods as in previous experiments (see Appendix C for more details) for both training and fine-tuning with the exception of Head-tuning where we applied the hyper-parameters from the original paper. In the English-to-German setting, we repeated the training 5 times with different seeds and averaged the results. In the multi-lingual setting, we performed a single training run for all encoder-decoder models with the same seed. Fine-tuning used the base model trained with the corresponding seed.

5.1 Single Language Pair Results

We tested several parameters for most methods. For fine-tuning-based models, we trained for $e \in \{1, 2, 5\}$ epochs and utilized only the examples with the maximum context size. For Weighting we set the λ parameter to 2, 5, and 10. In addition to the values of p for CoWord Dropout recommended by the authors (0.1, 0,2), we also included the value of 0.3. For Metric-based example selection, we set k=30,000 based on the number of annotated examples in the dataset, and used the MaxPCXMI metric (in Appendix E we present the comparison to the PCXMI metric). For Head-tuning we selected top $h \in \{1, 2, 3\}$ heads from Mąka et al. (2025). Results in terms of accuracy on the ctxPro dataset and BLEU on the IWSLT testset can be seen in Figure 5. Extended results are presented in Appendix E and calculations of statistical significance of the results can be seen in Appendix F.

It can be seen that with four metrics, the models' performance varies, and improvement in one metric comes at a cost of a reduction in another. In particular, we observe a negative relation between ctxPro accuracies and BLEU for all methods with the increase of the hyper-parameters. This necessitates examining the Pareto front in order to assess the performance of the methods. Metric-based example selection achieved highest improvement in Formality and outperformed the annotation-based selection for fine-tuning in Formality and Auxiliary, and achieved similar results for Gender, with a smaller decrease in BLEU. Head-tuning showed improvement only on Gender but with smaller drop in BLEU. Methods applied during training (Weighting, CoWord Dropout, and Divide and Rule) showed a smaller reduction in BLEU compared to fine-tuning. We attribute this to the smaller discrepancy in the dataset distribution between training and evaluation. Weighting outperformed CoWord Dropout on Gender and Auxiliary. Conversely, CoWord Dropout achieved the highest accuracy on Auxiliary (with Weighting being the second-best) but did not show any improvement for Gender and Formality. Notably, the highest reduc-

Model	BLEU	Gender	Formality	Auxiliary	Inflection	Animacy
Adapted D&R	**-0.05**	-0.06	-0.19	-0.16	-0.03	+0.07
CoWord p=0.1	-0.09	+0.02	-0.16	+0.35	-0.10	+0.16
CoWord p=0.2	-0.11	+0.07	-0.28	+0.65	-0.21	+0.01
CoWord p=0.3	-0.08	+0.01	-0.42	+0.97	-0.29	-0.27
MaxPCXMI e=1	-0.42	+1.13	+0.05	+3.41	+0.44	+1.08
MaxPCXMI e=2	-0.45	+1.42	+0.05	+4.25	+0.57	+1.10
MaxPCXMI e=5	-0.50	**+1.93**	**+0.11**	**+5.80**	**+0.76**	**+1.64**

Table 2: The averaged (over language directions) difference from the baseline in terms of BLEU on OpenSubtitles 2018 testsets and ctxPro phenomena accuracies for the tested methods applied to NLLB-200 600M model. Number of epochs is noted as "e", and CoWord Dropout probability as "p".

tion in BLEU was around 1% compared to the baseline. Lack of improvement exhibited by Adapted Divide and Rule can be attributed to our adaptation implementation, which did not utilize parameter freezing as in the original paper. Among all methods, metric-based example selection achieved the highest average ctxPro accuracy across phenomena, while token-level loss weighting was the most effective among annotation-based approaches, demonstrating that both proposed techniques can substantially improve context utilization.

5.2 Multilingual Results

We trained models based on NLLB-200 600M on all relevant language-directions using annotation-free methods (due to the lack of exhaustive annotations on the OpenSubtitles dataset; see Table 1) to assess their performance in the multilingual setting. For CoWord Dropout, we used the same values of p (0.1, 0.2, and 0.3), and for Metric-based example selection, we set k=10,000 per language direction and the number of epochs equal to 1, 2, and 5. The results aggregated over language directions can be seen in Table 2 and extended results in Appendix E.

Fine-tuning on examples selected by MaxPCXMI outperformed all baselines in terms of ctxPro accuracy across phenomena, with the highest improvement of 5.8, 1.9, and 1.6 percentage points (on average) for Auxiliary, Gender, and Animacy, respectively. Contrary to the English-to-German experiments, no improvement (on average) was observed for Formality. This was caused by a drop of up to 1 percentage point in the English-to-French direction, which offsets small gains in other language directions. These accuracy improvements came at the cost of a greater reduction in BLEU compared to other methods, and both trends—accuracy gains and BLEU drops—intensified with more fine-tuning epochs, mirroring the patterns seen in the single-language-direction experiments. It should be noted that MaxPCXMI was effectively trained for more updates than other methods in this experiment but additional training did not improve their results (as can be seen in Appendix E).

6 Conclusions

This work provided a systematic empirical evaluation of the influence of training data composition, in terms of contextually rich examples, on the context utilization capabilities for MT models. By systematically adapting the proportion of contextually rich examples in the training data, we demonstrated that such data sparsity is the key bottleneck in learning to leverage context efficiently. Crucially, we found that (1) models do not generalize well across different contextual phenomena (e.g. gender or formality) and (2) while there is some cross-lingual transfer, it was not significantly higher between languages in the same linguistic sub-family.

Motivated by these findings, we proposed two methods designed to mitigate the effect of data sparsity in context-aware MT: token-level loss weighting (based on token-level annotations of context-dependent words) and metric-based instance selection (fine-tuning on most contextually important examples). Both methods significantly improved context utilization without the need for extensive architectural changes or additional annotated data. Notably, the metric based method showed strong gains across multiple phenomena and language directions.

In practical terms, data composition and targeted training should be considered as potential solutions to developing strong context-aware MT models. In future work, combine the strengths of weighting and metric-based example selection.

7 Limitations

While we investigate many language directions and three sub-families, all of them come from the Indo-European family. This limitation was imposed by the language directions covered by ctxPro toolset. Additionally, for the single language pair setting, we only tested English-to-German direction. We suspect that the uncovered effects of data composition go beyond the tested language pairs, but this claim has not been tested experimentally.

For encoder-decoder architectures, we only tested the single encoder approach (standard Transformer) and multi-encoder models lay beyond the scope of this study. For decoder-only (LLM) setting, we based our experiments on a single model (Towerbase 7B). Both different model sizes and families could exhibit different behaviors. Furthermore, we tested the proposed methods for enhancing context utilizaiton only on the encoder-decoder models.

Acknowledgments

The research presented in this paper was conducted as part of VOXReality project[2], which was funded by the European Union Horizon Europe program under grant agreement No 101070521. This work used the Dutch national e-infrastructure with the support of the SURF Cooperative using grant no. EINF-12385.

References

Ruchit Agrawal, Marco Turchi, and Matteo Negri. 2018. Contextual handling in neural machine translation: Look behind, ahead and on both sides. In *Proceedings of the 21st Annual Conference of the European Association for Machine Translation*, pages 31–40.

Ibrahim M Alabdulmohsin, Xiaohua Zhai, Alexander Kolesnikov, and Lucas Beyer. 2023. Getting vit in shape: Scaling laws for compute-optimal model design. *Advances in Neural Information Processing Systems*, 36:16406–16425.

Belen Alastruey, Gerard I. Gállego, and Marta R. Costa-jussà. 2024. Unveiling the role of pretraining in direct speech translation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 11259–11265, Miami, Florida, USA. Association for Computational Linguistics.

Duarte M Alves, José Pombal, Nuno M Guerreiro, Pedro H Martins, João Alves, Amin Farajian, Ben Peters, Ricardo Rei, Patrick Fernandes, Sweta Agrawal, and 1 others. 2024. Tower: An open multilingual large language model for translation-related tasks. *arXiv preprint arXiv:2402.17733*.

Abdul Hameed Azeemi, Ihsan Ayyub Qazi, and Agha Ali Raza. 2025. To label or not to label: Hybrid active learning for neural machine translation. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 3071–3082, Abu Dhabi, UAE. Association for Computational Linguistics.

Guangsheng Bao, Yue Zhang, Zhiyang Teng, Boxing Chen, and Weihua Luo. 2021. G-transformer for document-level 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 3442–3455, Online. Association for Computational Linguistics.

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 *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 1–61, Florence, Italy. Association for Computational Linguistics.

Rachel Bawden, Rico Sennrich, Alexandra Birch, and Barry Haddow. 2018. Evaluating discourse phenomena in neural machine translation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1304–1313, New Orleans, Louisiana. Association for Computational Linguistics.

Aydar Bulatov, Yury Kuratov, and Mikhail Burtsev. 2022. Recurrent memory transformer. *Advances in Neural Information Processing Systems*, 35:11079–11091.

Mauro Cettolo, Marcello Federico, Luisa Bentivogli, Jan Niehues, Sebastian Stüker, Katsuhito Sudoh, Koichiro Yoshino, and Christian Federmann. 2017. Overview of the IWSLT 2017 evaluation campaign. In *Proceedings of the 14th International Conference on Spoken Language Translation*, pages 2–14, Tokyo, Japan. International Workshop on Spoken Language Translation.

Linqing Chen, Junhui Li, Zhengxian Gong, Min Zhang, and Guodong Zhou. 2022. One type context is not enough: Global context-aware neural machine translation. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 21(6).

Chenhui Chu, Raj Dabre, and Sadao Kurohashi. 2017. An empirical comparison of domain adaptation methods for neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for*

[2] https://voxreality.eu/

Computational Linguistics (Volume 2: Short Papers), pages 385–391, Vancouver, Canada. Association for Computational Linguistics.

Yukun Feng, Feng Li, Ziang Song, Boyuan Zheng, and Philipp Koehn. 2022. Learn to remember: Transformer with recurrent memory for document-level machine translation. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1409–1420, Seattle, United States. Association for Computational Linguistics.

Patrick Fernandes, Kayo Yin, Emmy Liu, André Martins, and Graham Neubig. 2023. When does translation require context? a data-driven, multilingual exploration. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 606–626, Toronto, Canada. Association for Computational Linguistics.

Patrick Fernandes, Kayo Yin, Graham Neubig, and André F. T. Martins. 2021. Measuring and increasing context usage in context-aware 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 6467–6478, Online. Association for Computational Linguistics.

Harritxu Gete, Thierry Etchegoyhen, and Gorka Labaka. 2023a. Targeted data augmentation improves context-aware neural machine translation. In *Proceedings of Machine Translation Summit XIX, Vol. 1: Research Track*, pages 298–312, Macau SAR, China. Asia-Pacific Association for Machine Translation.

Harritxu Gete, Thierry Etchegoyhen, and Gorka Labaka. 2023b. What works when in context-aware neural machine translation? In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 147–156, Tampere, Finland. European Association for Machine Translation.

Christian Hardmeier. 2012. Discourse in statistical machine translation: A survey and a case study. *Discours-Revue de linguistique, psycholinguistique et informatique*, 11.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, and 3 others. 2022. Training compute-optimal large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Jingjing Huo, Christian Herold, Yingbo Gao, Leonard Dahlmann, Shahram Khadivi, and Hermann Ney. 2020. Diving deep into context-aware neural machine translation. In *Proceedings of the Fifth Conference on Machine Translation*, pages 604–616, Online. Association for Computational Linguistics.

Sebastien Jean, Stanislas Lauly, Orhan Firat, and Kyunghyun Cho. 2017. Does neural machine translation benefit from larger context? *arXiv preprint arXiv:1704.05135*.

Prathyusha Jwalapuram, Shafiq Joty, and Youlin Shen. 2020. Pronoun-targeted fine-tuning for NMT with hybrid losses. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2267–2279, Online. Association for Computational Linguistics.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Benjamin Marie, Christof Monz, Kenton Murray, Masaaki Nagata, Martin Popel, Maja Popović, and 3 others. 2024. Findings of the WMT24 general machine translation shared task: The LLM era is here but MT is not solved yet. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1–46, Miami, Florida, USA. Association for Computational Linguistics.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 388–395, Barcelona, Spain. Association for Computational Linguistics.

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

António Lopes, M. Amin Farajian, Rachel Bawden, Michael Zhang, and André F. T. Martins. 2020. Document-level neural MT: A systematic comparison. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 225–234, Lisboa, Portugal. European Association for Machine Translation.

Minh-Thang Luong and Christopher Manning. 2015. Stanford neural machine translation systems for spoken language domains. In *Proceedings of the 12th International Workshop on Spoken Language Translation: Evaluation Campaign*, pages 76–79, Da Nang, Vietnam.

Lorenzo Lupo, Marco Dinarelli, and Laurent Besacier. 2022. Divide and rule: Effective pre-training for context-aware multi-encoder translation models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4557–4572, Dublin, Ireland. Association for Computational Linguistics.

Shuming Ma, Dongdong Zhang, and Ming Zhou. 2020. A simple and effective unified encoder for document-level machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3505–3511, Online. Association for Computational Linguistics.

Suvodeep Majumde, Stanislas Lauly, Maria Nadejde, Marcello Federico, and Georgiana Dinu. 2022. A baseline revisited: Pushing the limits of multi-segment models for context-aware translation. *arXiv preprint arXiv:2210.10906*.

Paweł Maka, Yusuf Semerci, Jan Scholtes, and Gerasimos Spanakis. 2024. Sequence shortening for context-aware machine translation. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 1874–1894, St. Julian's, Malta. Association for Computational Linguistics.

Paweł Mąka, Yusuf Can Semerci, Jan Scholtes, and Gerasimos Spanakis. 2025. Analyzing the attention heads for pronoun disambiguation in context-aware machine translation models. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 6348–6377, Abu Dhabi, UAE. Association for Computational Linguistics.

Ali Marashian, Enora Rice, Luke Gessler, Alexis Palmer, and Katharina von der Wense. 2025. From priest to doctor: Domain adaptation for low-resource neural machine translation. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 7087–7098, Abu Dhabi, UAE. Association for Computational Linguistics.

Sameen Maruf, André F. T. Martins, and Gholamreza Haffari. 2019. Selective attention for context-aware neural machine translation. 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 3092–3102, Minneapolis, Minnesota. Association for Computational Linguistics.

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

Wafaa Mohammed and Vlad Niculae. 2024. On measuring context utilization in document-level MT systems. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 1633–1643, St. Julian's, Malta. Association for Computational Linguistics.

Mathias Müller, Annette Rios, Elena Voita, and Rico Sennrich. 2018. A large-scale test set for the evaluation of context-aware pronoun translation in neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 61–72, Brussels, Belgium. Association for Computational Linguistics.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia-Gonzalez, Prangthip Hansanti, and 20 others. 2022. No language left behind: Scaling human-centered machine translation.

Jianhui Pang, Fanghua Ye, Derek Fai Wong, Dian Yu, Shuming Shi, Zhaopeng Tu, and Longyue Wang. 2025. Salute the classic: Revisiting challenges of machine translation in the age of large language models. *Transactions of the Association for Computational Linguistics*, 13:73–95.

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

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Matt Post and Marcin Junczys-Dowmunt. 2023. Escaping the sentence-level paradigm in machine translation. *arXiv preprint arXiv:2304.12959*.

Matt Post and Marcin Junczys-Dowmunt. 2024. Evaluation and large-scale training for contextual machine translation. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1125–1139, Miami, Florida, USA. Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Noam M. Shazeer and Mitchell Stern. 2018. Adafactor: Adaptive learning rates with sublinear memory cost. *ArXiv*, abs/1804.04235.

Zewei Sun, Mingxuan Wang, Hao Zhou, Chengqi Zhao, Shujian Huang, Jiajun Chen, and Lei Li. 2022. Rethinking document-level neural machine translation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3537–3548, Dublin, Ireland. Association for Computational Linguistics.

Jörg Tiedemann, Mikko Aulamo, Daria Bakshandaeva, Michele Boggia, Stig-Arne Grönroos, Tommi Nieminen, Alessandro Raganato Yves Scherrer, Raul Vazquez, and Sami Virpioja. 2023. Democratizing neural machine translation with OPUS-MT. *Language Resources and Evaluation*, (58):713–755.

Jörg Tiedemann and Yves Scherrer. 2017. Neural machine translation with extended context. In *Proceedings of the Third Workshop on Discourse in Machine Translation*, pages 82–92, Copenhagen, Denmark. Association for Computational Linguistics.

Jörg Tiedemann and Santhosh Thottingal. 2020. OPUS-MT — Building open translation services for the World. In *Proceedings of the 22nd Annual Conferenec of the European Association for Machine Translation (EAMT)*, Lisbon, Portugal.

Zhaopeng Tu, Yang Liu, Zhengdong Lu, Xiaohua Liu, and Hang Li. 2017. Context gates for neural machine translation. *Transactions of the Association for Computational Linguistics*, 5:87–99.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Elena Voita, Rico Sennrich, and Ivan Titov. 2019a. Context-aware monolingual repair for neural machine translation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 877–886, Hong Kong, China. Association for Computational Linguistics.

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

Rui Wang, Masao Utiyama, Lemao Liu, Kehai Chen, and Eiichiro Sumita. 2017. Instance weighting for neural machine translation domain adaptation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1482–1488, Copenhagen, Denmark. Association for Computational Linguistics.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, and 1 others. 2024. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. *arXiv preprint arXiv:2412.13663*.

Rachel Wicks and Matt Post. 2023. Identifying context-dependent translations for evaluation set production. In *Proceedings of the Eighth Conference on Machine Translation*, pages 452–467, Singapore. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 others. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Billy T. M. Wong and Chunyu Kit. 2012. Extending machine translation evaluation metrics with lexical cohesion to document level. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1060–1068, Jeju Island, Korea. Association for Computational Linguistics.

Kayo Yin, Patrick Fernandes, Danish Pruthi, Aditi Chaudhary, André F. T. Martins, and Graham Neubig. 2021. Do context-aware translation models pay the right attention? 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 788–801, Online. Association for Computational Linguistics.

Pei Zhang, Boxing Chen, Niyu Ge, and Kai Fan. 2020. Long-short term masking transformer: A simple but effective baseline for document-level neural machine translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1081–1087, Online. Association for Computational Linguistics.

Xuan Zhang, Gaurav Kumar, Huda Khayrallah, Kenton Murray, Jeremy Gwinnup, Marianna J Martindale, Paul McNamee, Kevin Duh, and Marine Carpuat. 2018. An empirical exploration of curriculum learning for neural machine translation. *arXiv preprint arXiv:1811.00739*.

Zaixiang Zheng, Xiang Yue, Shujian Huang, Jiajun Chen, and Alexandra Birch. 2021. Towards making the most of context in neural machine translation. In *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, pages 3983–3989.

Zhang Zhuocheng, Shuhao Gu, Min Zhang, and Yang Feng. 2023. Scaling law for document neural machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8290–8303, Singapore. Association for Computational Linguistics.

A Details of ctxPro Dataset

In this section, we provide a short description of the context-dependent phenomena that can be identified by the ctxPro toolset (Wicks and Post, 2023):

- **Gender** (anaphoric pronouns) - translating a pronoun from a non-gendered language to a language with gendered nouns. Available for English-to-X language directions, where X is {German, French, Polish, Russian, and Spanish}.
- **Formality** (anaphoric pronouns) - translating into a language with different second-person pronouns distinguishing intimate from formal relationships between speakers from a language lacking this distinction. Available for English-to-X language directions, where X is {German, French, Polish, Russian, and Spanish}.
- **Animacy** (anaphoric pronouns) - translating into English, a language that distinguishes between animate (she/he) and inanimate (it) pronouns, from a language that does not exhibit this distinction. Available for X-to-English language directions, where X is {German, French, Polish, Russian, and Spanish}.
- **Auxiliary** (verb phrase ellipsis) - translating into a language that require the head of the verb phrase from a language that allows for only the modal or auxiliary to be used. Available for English-to-X language directions, where X is {German, French, Polish, Russian, and Spanish}.
- **Inflection** (verb phrase ellipsis) - translating into a language with noun morphology dependent on the grammatical role from a language where this is not the case. Available for English-to-Polish and English-to-Russian language directions.

In Table 3 we present the number of examples in the ctxPro dataset with a particular antecedent distance. Additionally, we present the proportion of examples that have the antecedent distance larger than three, which is beyond the context size available to our models. Note that for Formality, the antecedent distances are not specified. We refer the reader to the original paper (Wicks and Post, 2023) for more details.

B Composition of the Datasets

In this section, we describe how the constructed datasets were created. Table 4 shows the sizes of the dense component datasets. For the *Pure IWSLT* setting, we start with the IWSLT-sparse (123,000 examples with no annotations) and progressively replace it with the examples sampled from IWSLT-dense. The steps are based on the size of the IWSLT-dense dataset for a particular phenomenon: 3,000 and 6,915 (full size) for Gender, 10,000 and 21,977 (full size) for Formality, and 19 (full size) for Auxiliary. For the *IWSLT + OS* setting, we start with the datasets formed by combining IWSLT-sparse with examples sampled from OS-rand. To maximize the density of the resulting datasets, we set the number of examples sampled from OS-rand to be dependent on the phenomenon and equal to the (rounded) size of the OS-dense datasets: 12,000 for Gender, 17,000 for Formality, and 1,200 for Auxiliary. We start by replacing examples from IWSLT-sparse (we retain the steps from the Pure IWSLT setting). After reaching the maximum density in the IWSLT portion of the dataset, we start replacing OS-rand with OS-dense in the following steps: 4,000, 8,000, and 12,000 for Gender, 6,000, 12,000, and 17,000 for Formality, and 400, 800, and 1,200 for Auxiliary.

Tables 5, 6, and 7 show the composition of the training datasets we used in the experiments for Gender, Formality, and Auxiliary phenomena, respectively. Each example was encoded with the context size ranging from zero to the maximum context size (three in our experiments), increasing the size of the datasets four times.

In the multilingual experiments, we formed the baseline training dataset by sampling 50,000 examples from OpenSubtitles (OS-rand) for each language direction we considered. For each phenomenon in a language direction, we replaced examples with the rich ones: 6,900 for Gender, 10,000 for Formality, 1,200 for Auxiliary, 10,000 for Inflection, and 4,000 for Animacy.

C Details of Context-aware Training

We implemented all experiments in *Huggingface transformers* framework (Wolf et al., 2020). We trained the models in the following categories: single language direction (OpusMT en-de[3]), multilingual (NLLB-200 600M[4]), and LLM-based multilingual (Towerbase 7B[5]). Additionally, we repeated the experiments in the single language direction

[3] https://huggingface.co/Helsinki-NLP/opus-mt-en-de

[4] https://huggingface.co/facebook/nllb-200-distilled-600M

[5] https://huggingface.co/Unbabel/TowerBase-7B-v0.1

Direction	Phenomenon	Antecedent Distance					
		0	**1**	**2**	**3**	**>3**	**% >3**
En↔De	Auxiliary	0	1754	498	256	672	21%
	Gender	7307	13731	4814	2308	3480	11%
	Animacy	5309	9493	3115	1362	1956	9%
En↔Es	Auxiliary	0	4922	1051	323	664	10%
	Gender	4126	6979	2702	1317	2392	14%
	Animacy	2852	4102	1550	750	1291	12%
En↔Fr	Auxiliary	0	5263	1327	474	1258	15%
	Gender	11236	18037	6294	2921	4887	11%
	Animacy	5468	8350	2873	1317	1992	10%
En↔It	Auxiliary	0	3590	1018	344	972	16%
	Gender	6117	7128	2630	1365	2173	11%
	Animacy	3277	3708	1367	707	1057	10%
En↔Pl	Auxiliary	0	5437	1180	391	1077	13%
	Gender	17186	25242	8201	3906	5992	10%
	inflection	0	12905	5094	3235	8766	29%
	Animacy	3455	5565	1784	855	1245	10%
En↔Ru	Auxiliary	0	6056	1467	402	742	9%
	Gender	8227	14283	4873	2243	3322	10%
	inflection	0	15042	4659	2746	7553	25%
	Animacy	5460	9760	3422	1565	2323	10%

Table 3: Number of examples in ctxPro dataset with certain values for antecedent distance for used language directions and phenomena. Antecedent distances larger than 3 were combined and we also show the proportion of those examples in the dataset. Note that for Formality, the antecedent distance is not specified.

Dataset	Language	Gender	Formality	Auxiliary	Inflection	Animacy
IWSLT-dense	En→De	6,915	21,977	19	-	-
OS-dense	En↔De	12,326	16,064	1,230	-	8,334
	En↔Es	6,936	20,374	2,768	-	4,211
	En↔Fr	16,804	10,858	3,314	-	7,904
	En↔Pl	23,683	41,806	3,184	10,897	5,112
	En↔Ru	8,141	14,211	3,443	10,971	4,237

Table 4: Sizes of the dense component datasets divided into phenomena (columns).

setting using randomly initialized OpusMT model for which we performed sentence-level pre-training on the mixture of IWSLT 2017 en-de training subset and randomly sampled 2.5M sentences from WMT 2019 en-de (Barrault et al., 2019) training subset. We trained the models with Adafactor optimizer (Shazeer and Stern, 2018) on a single GPU (NVIDIA GeForce RTX 3090 24GB for Opus MT en-de and NVIDIA H100 80GB for NLLB-200 600M and Towerbase 7B). We used LoRA (Hu et al., 2022) to fine-tune Towerbase models. OpusMT en-de contain 163M parameters, NLLB-200 600M contain 615M parameters, and Towerbase 7B contain 6,770M parameters (32M trainable parameters through LoRA). The inputs during training and prompt used for Towerbase models can be seen in Listings 1 and 2 respectively. We calculated loss during training only based on the target language parts of the examples corresponding to the generations of the model.

Listing 1: Input template used for training Towerbase models. The number of sentences in context is the same for source and target sides but can vary from example to example. Sentences are separated by the "<sep>" string.

```
[src_lang]: [src_ctx] <sep> [src] \n
[tgt_lang]: [tgt_ctx] <sep> [tgt]
```

Setting	IWSLT-sparse	IWSLT-dense	OS-rand	OS-dense	Total
Pure IWSLT	123,000	0	0	0	123,000
	120,000	3,000	0	0	123,000
	116,085	6,915	0	0	123,000
IWSLT+OS	123,000	0	12,000	0	135,000
	120,000	3,000	12,000	0	135,000
	116,085	6,915	12,000	0	135,000
	116,085	6,915	8,000	4,000	135,000
	116,085	6,915	4,000	8,000	135,000
	116,085	6,915	0	12,000	135,000

Table 5: Number of examples from datasets that were used to compose training datasets (in rows) for the **Gender** phenomenon in the single language direction (English-to-German) setting.

Setting	IWSLT-sparse	IWSLT-dense	OS-rand	OS-dense	Total
Pure IWSLT	123,000	0	0	0	123,000
	113,000	10,000	0	0	123,000
	101,023	21,977	0	0	123,000
IWSLT+OS	123,000	0	17,000	0	140,000
	113,000	10,000	17,000	0	140,000
	101,023	21,977	17,000	0	140,000
	101,023	21,977	11,000	6,000	140,000
	101,023	21,977	5,000	12,000	140,000
	101,023	21,977	0	17,000	140,000

Table 6: Number of examples from datasets that were used to compose training datasets (in rows) for the **Formality** phenomenon in the single language direction (English-to-German) setting.

Listing 2: Prompt template used for generation with Towerbase models. The number of context sentences can vary. Sentences are separated by the "<sep>" string.

```
[src_lang]: [src_ctx] <sep> [src] \n
[tgt_lang]:
```

The hyper-parameters are presented in Table 8. We tuned the hyper-parameters (learning rate, batch size, number of epochs) during the preliminary experiments on OpusMT en-de model with context size of one trained on IWSLT 2017 English-to-German dataset. Hyper-parametes for sentence-level pre-training were tuned on WMT 2019 en-de evaluation subset, and on randomly sampled subset of OpenSubtitles en-de dataset for the fine-tuning of Towerbase 7B model.

D Extended Data Composition Results

In this section, we present the extended results of the data composition experiments. For single language pair setting, we measured COMET (Rei et al., 2020) (based on `Unbabel/wmt22-comet-da`) on the IWSLT 2017 en-de testset and evaluated the models on the ContraPro (Müller et al., 2018) contrastive evaluation. The results for the Pure IWSLT and IWSLT+OS settings can be found in Tables 9 and 10, respectively. The results for English-to-German language direction with models randomly initialized can be seen in Figure 6.

For the multilingual setting, we additionally measured BLEU (we used the sacreBLEU library (Post, 2018) using the default parameters) and COMET on the testsets formed by sampling 20,000 examples from OpenSubtitles 2018 for each language direction. The results for models based on NLLB-200 600M can be seen in Tables 11 and 12 for BLEU and COMET, respectively. The results for models based on Towerbase 7B can be seen in Tables 13 (BLEU) and 14 (COMET).

E Extended Fine-tuning Results

For the English-to-German experiment, apart from BLEU and ctxPro accuracy, we also measured COMET (Rei et al., 2020) (based on `Unbabel/wmt22-comet-da`) on the IWSLT 2017 en-de testset and the accuracy on the ContraPro contrastive evaluation. The results (including

Setting	IWSLT-sparse	IWSLT-dense	OS-rand	OS-dense	Total
Pure IWSLT	123,000	0	0	0	123,000
	122,981	19	0	0	123,000
IWSLT+OS	123,000	0	1,200	0	124,200
	122,981	19	1,200	0	124,200
	122,981	19	800	400	124,200
	122,981	19	400	800	124,200
	122,981	19	0	1,200	124,200

Table 7: Number of examples from datasets that were used to compose training datasets (in rows) for the **Auxiliary** phenomenon in the single language direction (English-to-German) setting.

Hyper-parameter	Sentence-level Pre-training	OpusMT Fine-tuning	NLLB-200 Fine-tuning	Towerbase Fine-tuning
Optimizer	Adafactor	Adafactor	Adafactor	Adafactor
Learning Rate	5e-5	1e-5	1e-5	1e-5
LR Scheduler	Linear	Inverse Sqrt	Inverse Sqrt	Inverse Sqrt
LR Warmup Ratio	0.0	0.1	0.1	0.1
Weight Decay	0.01	0.01	0.01	0.01
Batch Size	32	32[a]	32	16
Gradient Accumulation Steps	16	16[a]	16	4
Num Epoch	30	10	10	3
Precission	fp16	fp16	fp16	bf16
Seeds	1,2,3,4,5	1,2,3,4,5	1	1
Max Length	512	512	1024	2048
Max Context Size	-	3	3	3
Beam size	5	5	5	1[b]
Lora alpha	-	-	-	32
Lora r	-	-	-	16

Table 8: The hyper-parameters of training and fine-tuning.
[a] For the cases where the CUDA Out Of Memory error occurred, we reduced the batch size to 16 and increased the Gradient Accumulation Steps to 32, keeping the same effective size of the batch.
[b] For Towerbase models, we use the greedy decoding strategy.

Dataset	Count	COMET	ContraPro
Sparse	0	0.8415	69.23
Gender	3,000	0.8417	74.70
	6,915	0.8417	78.45
Formality	10,000	0.8429	69.55
	21,977	0.8430	70.02
Auxiliary	19	0.8413	69.14

Table 9: Performance in terms of COMET on IWSLT 2017 en-de testset and ContraPro accuracy for the models based on OpusMT en-de in the **Pure IWSLT** setting trained on datasets with different numbers of examples annotated with different phenomena.

BLEU and ctxPro accuracies) can be seen in Table 15.

Next, we present the results of Metric-based selection of examples for fine-tuning for two metrics: PCXMI (Fernandes et al., 2023) and MaxPCXMI (ours). We fine-tuned the models for 1, 2, and 5 epochs and repeated the experiment 5 times with different seeds (using the base context-aware model trained with the corresponding seed). The averaged results can be seen in Figure 7. Selecting examples based on MaxPCXMI outperforms PCXMI in Gender and Formality at a lower reduction in BLEU. PCXMI achieves a better increase in Auxiliary but reduces BLEU even below the level of the annotation-based method.

The un-aggregated results of the trained models for each language direction in the multilingual experiment can be seen in Figure 8 (including models trained for one more epoch) and Tables 16 and 17 for ctxPro accuracies, BLEU and COMET, respec-

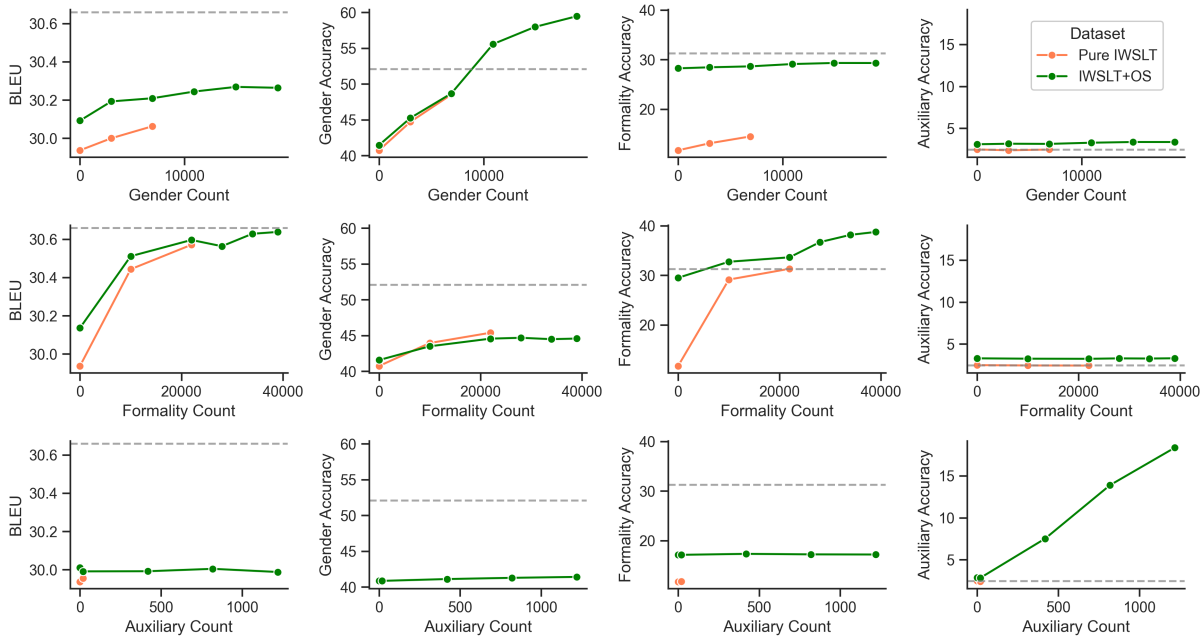

Figure 6: Measured metrics of BLEU on IWSLT 2017 testset, and ctxPro accuracy on Gender, Formality, and Auxiliary phenomena (in columns) of the randomly initialized models trained on the datasets with varying amounts of contextually-rich examples of Gender, Formality, and Auxiliary phenomena (in rows). Shows two experimental settings: Pure IWSLT and combined IWSLT+OS.

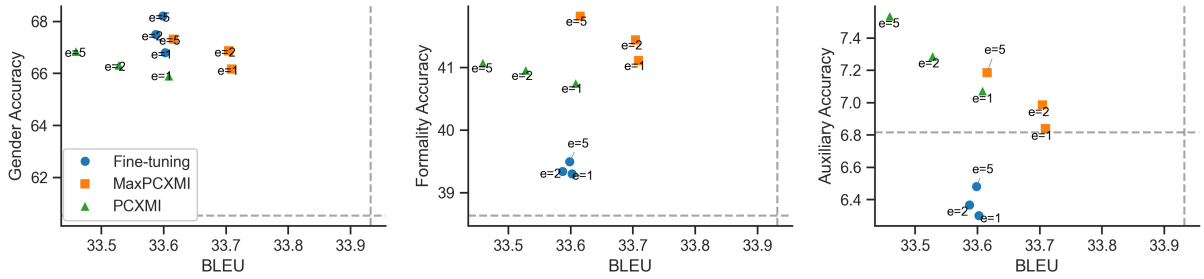

Figure 7: Accuracy of ctxPro English-to-German phenomena (Gender, Formality, and Auxiliary) against BLEU on the IWSLT 2017 en-de testset of the fine-tuned models with Metric-based (PCXMI and MaxPCXMI) and annotation-based (for comparison) selection of examples. Models are based on OpusMT en-de. Labels show the number of epochs ("e").

tively.

F Statistical Significance

In this section, we calculate the statistical significance of the fine-tuning results on the single language pair setting. In particular, we employ the paired bootstrap resampling method (Koehn, 2004) to calculate whether the differences in obtained results between tested methods are statistically significant. We use sacreBLEU (Post, 2018) implementation extended to other metrics. To include the runs with all seeds, we concatenate the predictions (as well as references) for all runs of a particular model. The results in terms of p-values of the paired bootstrapping of the results are presented in Tables 18 to 22 for: BLEU on the IWSLT 2017 en-de testset, COMET on the IWSLT 2017 en-de testset, ctxPro Gender accuracy, ctxPro Formality accuracy, and ctxPro Auxiliary accuracy, respectively.

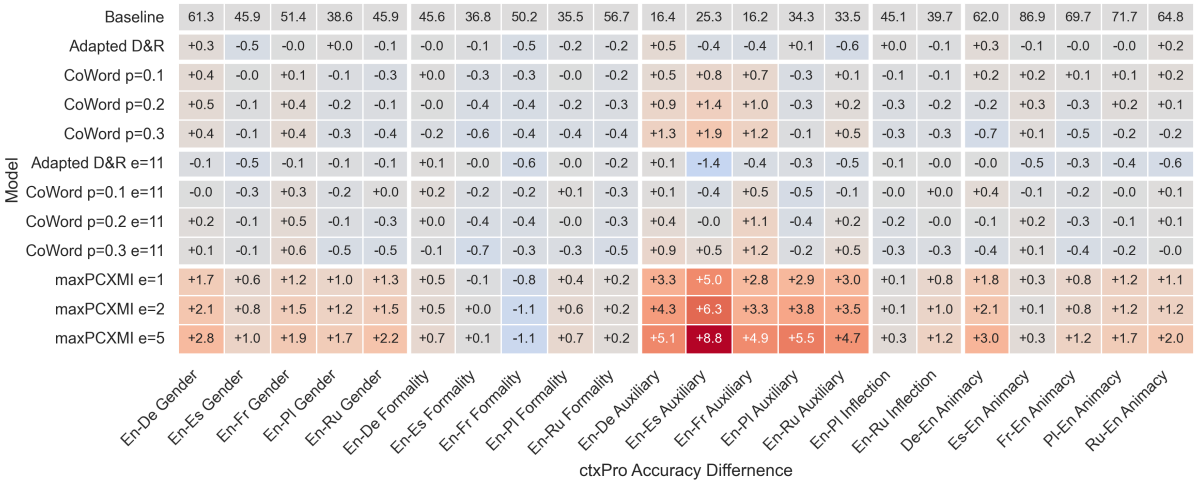

Figure 8: Measured ctxPro accuracy on all phenomena for each of the relevant language directions (in columns) of tested methods (in rows) applied to NLLB-200 600M model.

Dataset	Count	COMET	ContraPro
Gender	0	0.8417	70.28
	3,000	0.8417	75.03
	6,915	0.8420	78.52
	10,915	0.8419	83.58
	14,915	0.8418	84.77
	18,915	0.8420	85.24
Formality	0	0.8416	70.15
	10,000	0.8426	70.59
	21,977	0.8428	71.12
	27,977	0.8429	71.04
	33,977	0.8429	70.85
	38,977	0.8430	71.03
Auxiliary	0	0.8414	69.47
	19	0.8415	69.39
	419	0.8415	69.60
	819	0.8415	69.75
	1,219	0.8416	69.79

Table 10: Performance in terms of COMET on IWSLT 2017 en-de testset and ContraPro accuracy for the models based on OpusMT en-de in the **IWSLT+OS** setting trained on datasets with different numbers of examples annotated with different phenomena.

Model	En-De	En-Es	En-Fr	En-Pl	En-Ru	De-En	Es-En	Fr-En	Pl-En	Ru-En
Baseline	26.50	37.68	29.39	21.98	24.49	32.04	41.99	32.84	29.42	31.35
Gender										
En-De	26.66	37.61	29.27	21.85	24.46	31.98	42.03	32.87	29.54	31.32
En-Es	26.88	37.60	29.33	22.12	24.52	32.03	41.96	32.86	29.46	31.36
En-Fr	26.75	37.53	29.16	22.05	24.41	32.01	41.97	32.87	29.50	31.33
En-Pl	26.80	37.57	29.21	21.54	24.48	32.05	42.00	32.86	29.53	31.34
En-Ru	26.78	37.60	29.56	21.91	24.45	32.01	42.04	32.81	29.52	31.41
Formality										
En-De	26.61	37.27	29.29	21.75	24.44	31.98	42.05	32.85	29.52	31.31
En-Es	26.58	37.29	29.43	21.65	24.57	32.01	42.04	32.84	29.49	31.39
En-Fr	26.70	37.63	29.67	21.89	24.48	32.02	41.99	32.92	29.52	31.37
En-Pl	26.62	37.38	29.44	21.83	24.35	32.03	42.00	32.88	29.44	31.23
En-Ru	26.88	37.53	29.36	22.05	24.22	32.04	42.03	32.91	29.50	31.39
Auxiliary										
En-De	26.86	37.57	29.26	21.77	24.48	32.01	42.08	32.91	29.51	31.42
En-Es	26.88	37.44	29.38	22.01	24.46	32.09	41.98	32.85	29.46	31.40
En-Fr	26.94	37.53	29.56	21.97	24.42	32.01	41.99	32.83	29.51	31.28
En-Pl	26.65	37.69	29.33	21.70	24.47	32.04	42.05	32.82	29.47	31.26
En-Ru	26.73	37.50	29.35	22.03	24.55	32.08	41.95	32.84	29.51	31.36
Inflection										
En-Pl	26.95	37.58	29.41	21.68	24.59	32.07	41.98	32.87	29.49	31.40
En-Ru	26.80	37.63	29.31	21.90	24.43	32.06	42.04	32.85	29.51	31.30
Animacy										
De-En	26.80	37.43	29.32	21.84	24.65	32.05	42.05	32.84	29.48	31.41
Es-En	26.83	37.59	29.39	22.20	24.50	32.02	41.97	32.81	29.51	31.27
Fr-En	26.93	37.70	29.23	21.85	24.55	32.04	42.02	32.88	29.46	31.27
Pl-En	26.71	37.55	29.35	21.89	24.46	32.09	42.01	32.88	29.44	31.35
Ru-En	26.83	37.51	29.35	21.73	24.48	32.00	41.95	32.86	29.48	31.35

Table 11: BLEU scores for the models based on NLLB-200 600M trained on datasets with different densities of annotated examples in the multilingual setting on the test subsets of the OpenSubtitles 2018 datasets for all relevant language pairs.

Model	En-De	En-Es	En-Fr	En-Pl	En-Ru	De-En	Es-En	Fr-En	Pl-En	Ru-En
Baseline	0.8023	0.8459	0.8005	0.8171	0.8321	0.8182	0.8522	0.8192	0.8009	0.8086
Gender										
En-De	0.8025	0.8456	0.8001	0.8171	0.8325	0.8182	0.8522	0.8189	0.8011	0.8085
En-Es	0.8025	0.8462	0.8004	0.8172	0.8326	0.8181	0.8521	0.8193	0.8011	0.8086
En-Fr	0.8023	0.8456	0.8000	0.8172	0.8322	0.8182	0.8521	0.8192	0.8011	0.8085
En-Pl	0.8025	0.8458	0.8004	0.8176	0.8324	0.8182	0.8522	0.8193	0.8011	0.8084
En-Ru	0.8021	0.8456	0.7999	0.8168	0.8321	0.8182	0.8523	0.8189	0.8009	0.8086
Formality										
En-De	0.8023	0.8456	0.8002	0.8168	0.8324	0.8181	0.8522	0.8190	0.8010	0.8084
En-Es	0.8026	0.8455	0.8003	0.8171	0.8325	0.8182	0.8523	0.8191	0.8011	0.8087
En-Fr	0.8024	0.8458	0.8008	0.8173	0.8321	0.8183	0.8522	0.8192	0.8011	0.8087
En-Pl	0.8024	0.8456	0.8005	0.8176	0.8325	0.8185	0.8523	0.8192	0.8009	0.8085
En-Ru	0.8022	0.8456	0.8001	0.8171	0.8318	0.8183	0.8524	0.8190	0.8009	0.8085
Auxiliary										
En-De	0.8023	0.8458	0.8001	0.8171	0.8321	0.8185	0.8524	0.8190	0.8011	0.8085
En-Es	0.8024	0.8458	0.8006	0.8174	0.8327	0.8185	0.8522	0.8191	0.8011	0.8085
En-Fr	0.8025	0.8455	0.7999	0.8165	0.8322	0.8181	0.8521	0.8189	0.8010	0.8085
En-Pl	0.8026	0.8458	0.8001	0.8170	0.8321	0.8183	0.8522	0.8191	0.8009	0.8083
En-Ru	0.8024	0.8457	0.8001	0.8169	0.8326	0.8183	0.8520	0.8190	0.8009	0.8085
Inflection										
En-Pl	0.8025	0.8458	0.8004	0.8162	0.8323	0.8184	0.8522	0.8191	0.8010	0.8087
En-Ru	0.8021	0.8457	0.7999	0.8168	0.8309	0.8184	0.8523	0.8190	0.8010	0.8084
Animacy										
De-En	0.8026	0.8458	0.8003	0.8174	0.8324	0.8184	0.8524	0.8188	0.8010	0.8085
Es-En	0.8025	0.8459	0.8005	0.8171	0.8328	0.8184	0.8522	0.8191	0.8009	0.8086
Fr-En	0.8021	0.8458	0.8000	0.8168	0.8325	0.8181	0.8523	0.8191	0.8008	0.8083
Pl-En	0.8021	0.8456	0.8004	0.8171	0.8322	0.8183	0.8522	0.8192	0.8008	0.8085
Ru-En	0.8022	0.8455	0.8003	0.8172	0.8321	0.8182	0.8521	0.8189	0.8008	0.8083

Table 12: COMET scores for the models based on NLLB-200 600M trained on datasets with different densities of annotated examples in the multilingual setting on the test subsets of the OpenSubtitles 2018 datasets for all relevant language pairs.

Model	En-De	En-Es	En-Fr	En-Ru	De-En	Es-En	Fr-En	Ru-En
Baseline	25.93	32.78	29.45	21.56	31.06	42.23	33.84	28.40
Gender								
En-De	25.81	32.83	29.09	20.81	31.72	41.94	32.74	28.13
En-Es	25.37	34.02	28.95	21.22	31.12	42.61	33.53	28.18
En-Fr	25.60	34.00	28.64	21.86	31.37	42.22	33.77	28.02
En-Ru	24.74	32.82	28.92	22.02	30.94	42.55	33.86	27.28
Formality								
En-De	25.61	33.84	29.00	20.45	31.50	42.71	33.47	28.35
En-Es	25.87	33.61	28.87	22.05	31.40	41.37	33.96	29.01
En-Fr	25.46	33.63	29.35	22.10	30.86	41.40	33.78	27.55
En-Ru	25.43	32.55	29.45	22.65	30.84	42.76	33.81	27.80
Auxiliary								
En-De	26.10	31.95	28.85	21.11	31.48	41.89	33.29	28.93
En-Es	25.66	32.33	29.03	21.44	31.50	41.95	33.64	27.94
En-Fr	25.75	33.30	29.19	21.91	31.12	42.26	33.83	28.76
En-Ru	25.60	33.71	28.96	22.24	31.77	41.57	33.89	28.30
Inflection								
En-Ru	25.52	32.71	28.72	21.56	30.73	42.55	33.80	28.11
Animacy								
De-En	25.07	34.34	28.87	21.31	30.88	41.92	33.42	28.66
Es-En	25.94	32.01	29.03	21.58	30.78	43.05	33.88	27.84
Fr-En	25.32	32.97	29.22	22.39	31.40	41.72	33.15	28.60
Ru-En	25.48	33.04	29.15	22.23	30.26	41.85	33.48	29.34

Table 13: BLEU scores for the models based on Towerbase 7B trained on datasets with different densities of annotated examples in the multilingual setting on the test subsets of the OpenSubtitles 2018 datasets for all relevant language pairs.

Model	En-De	En-Es	En-Fr	En-Ru	De-En	Es-En	Fr-En	Ru-En
Baseline	0.8003	0.8407	0.7979	0.8336	0.8186	0.8547	0.8236	0.8106
Gender								
En-De	0.8013	0.8410	0.7981	0.8342	0.8193	0.8548	0.8239	0.8114
En-Es	0.8000	0.8412	0.7979	0.8340	0.8193	0.8546	0.8235	0.8113
En-Fr	0.8003	0.8412	0.7983	0.8340	0.8193	0.8547	0.8237	0.8110
En-Ru	0.8005	0.8409	0.7982	0.8348	0.8187	0.8540	0.8236	0.8108
Formality								
En-De	0.8013	0.8409	0.7979	0.8344	0.8191	0.8547	0.8241	0.8114
En-Es	0.8004	0.8404	0.7978	0.8341	0.8192	0.8544	0.8235	0.8116
En-Fr	0.8005	0.8412	0.7988	0.8340	0.8188	0.8546	0.8238	0.8109
En-Ru	0.8007	0.8414	0.7979	0.8345	0.8185	0.8542	0.8236	0.8107
Auxiliary								
En-De	0.8005	0.8405	0.7974	0.8339	0.8189	0.8545	0.8239	0.8112
En-Es	0.8005	0.8405	0.7978	0.8338	0.8188	0.8547	0.8236	0.8114
En-Fr	0.8006	0.8406	0.7979	0.8338	0.8189	0.8546	0.8236	0.8109
En-Ru	0.8004	0.8408	0.7982	0.8339	0.8190	0.8541	0.8236	0.8107
Inflection								
En-Ru	0.8007	0.8409	0.7982	0.8335	0.8189	0.8546	0.8236	0.8108
Animacy								
De-En	0.8002	0.8405	0.7978	0.8342	0.8190	0.8548	0.8231	0.8112
Es-En	0.8007	0.8405	0.7977	0.8340	0.8192	0.8550	0.8237	0.8110
Fr-En	0.8003	0.8407	0.7981	0.8337	0.8189	0.8544	0.8236	0.8109
Ru-En	0.8004	0.8409	0.7978	0.8337	0.8191	0.8545	0.8237	0.8111

Table 14: COMET scores for the models based on Towerbase 7B trained on datasets with different densities of annotated examples in the multilingual setting on the test subsets of the OpenSubtitles 2018 datasets for all relevant language pairs.

Model	BLEU	COMET	Gender	Formality	Auxiliary	ContraPro
Baseline	33.93	0.8431	60.52%	38.63%	6.81%	78.88%
Fine-tuning e=1	33.60	0.8416	66.79%	39.30%	6.30%	83.02%
Fine-tuning e=2	33.59	0.8416	67.49%	39.34%	6.37%	83.78%
Fine-tuning e=5	33.60	0.8415	68.20%	39.49%	6.48%	84.50%
Head-tuning h=1	33.89	0.8428	63.28%	38.64%	6.43%	82.61%
Head-tuning h=2	33.85	0.8427	64.04%	38.58%	6.44%	83.40%
Head-tuning h=3	33.80	0.8425	64.75%	38.27%	6.45%	84.36%
Weighting λ=2	33.94	0.8430	64.35%	39.14%	7.18%	83.10%
Weighting λ=5	33.83	0.8430	65.72%	39.48%	7.67%	84.63%
Weighting λ=10	33.74	0.8426	66.24%	39.81%	8.10%	85.11%
Adapted D&R None	33.95	0.8429	60.77%	38.17%	7.01%	78.66%
CoWord p=0.1	33.98	0.8435	60.54%	38.72%	7.79%	78.65%
CoWord p=0.2	33.95	0.8436	60.47%	38.72%	8.22%	78.52%
CoWord p=0.3	33.88	0.8433	60.29%	38.68%	8.59%	78.39%
MaxPCXMI e=1	33.71	0.8420	66.16%	41.11%	6.84%	82.95%
MaxPCXMI e=2	33.70	0.8418	66.86%	41.44%	6.99%	83.79%
MaxPCXMI e=5	33.62	0.8414	67.31%	41.82%	7.18%	84.39%

Table 15: Performance in terms of BLEU and COMET on IWSLT 2017 en-de testset and ctxPro and ContraPro accuracy for the different methods applied to OpusMT en-de model. Number of epochs is noted as "e", and CoWord Dropout probability as "p", number of tuned heads as "h", and weighting strength as "λ".

Model	En-De	En-Es	En-Fr	En-Pl	En-Ru	De-En	Es-En	Fr-En	Pl-En	Ru-En
Baseline	26.50	37.68	29.39	21.98	24.49	32.04	41.99	32.84	29.42	31.35
Adapted D&R	26.50	37.00	29.48	22.00	24.44	32.05	42.01	32.88	29.50	31.30
CoWord p=0.1	26.72	37.48	28.86	21.89	24.27	32.10	41.97	32.77	29.41	31.31
CoWord p=0.2	26.45	37.31	29.27	22.01	24.25	32.05	41.88	32.75	29.35	31.30
CoWord p=0.3	26.58	37.61	29.48	21.95	24.15	32.11	41.82	32.68	29.28	31.22
MaxPCXMI e=1	26.00	37.04	28.71	21.23	24.02	31.89	41.78	32.73	29.35	30.76
MaxPCXMI e=2	26.04	37.02	28.59	21.34	23.90	31.81	41.81	32.71	29.31	30.68
MaxPCXMI e=5	26.09	36.93	28.74	21.29	23.85	31.78	41.65	32.62	29.22	30.46

Table 16: BLEU scores for the methods applied to NLLB-200 600M model in the multilingual setting on the test subsets of the OpenSubtitles 2018 datasets for all relevant language pairs.

Model	En-De	En-Es	En-Fr	En-Pl	En-Ru	De-En	Es-En	Fr-En	Pl-En	Ru-En
Baseline	0.8023	0.8459	0.8005	0.8171	0.8321	0.8182	0.8522	0.8192	0.8009	0.8086
Adapted D&R	0.8026	0.8456	0.8000	0.8175	0.8322	0.8183	0.8522	0.8191	0.8011	0.8085
CoWord p=0.1	0.8023	0.8454	0.7994	0.8167	0.8317	0.8182	0.8521	0.8188	0.8006	0.8086
CoWord p=0.2	0.8015	0.8453	0.7994	0.8166	0.8316	0.8178	0.8518	0.8187	0.8002	0.8083
CoWord p=0.3	0.8014	0.8453	0.7990	0.8164	0.8313	0.8176	0.8516	0.8183	0.7996	0.8083
MaxPCXMI e=1	0.7990	0.8433	0.7963	0.8125	0.8296	0.8155	0.8501	0.8170	0.7988	0.8057
MaxPCXMI e=2	0.7987	0.8431	0.7958	0.8123	0.8296	0.8150	0.8499	0.8167	0.7982	0.8053
MaxPCXMI e=5	0.7974	0.8427	0.7947	0.8109	0.8285	0.8137	0.8490	0.8158	0.7970	0.8043

Table 17: COMET scores for the methods applied to NLLB-200 600M model in the multilingual setting on the test subsets of the OpenSubtitles 2018 datasets for all relevant language pairs.

Table 18: The p-values of the paired bootstrapping of the results in terms of BLEU on the IWSLT2017 English-to-German testset for each pair of the models based on OpusMT en-de. Values <0.05 are in bold.

Model	Baseline	CoWord p=0.1	CoWord p=0.2	CoWord p=0.3	Adapted D&R	Fine-tuning e=1	Fine-tuning e=2	Fine-tuning e=5	MaxPCXMI e=1	MaxPCXMI e=2	MaxPCXMI e=5	Weighting λ=2	Weighting λ=5	Weighting λ=10
Baseline	-	**0.045**	0.243	0.103	0.192	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**	0.319	**0.001**	**0.001**
CoWord p=0.1	**0.045**	-	0.085	**0.007**	0.124	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**	0.069	**0.001**	**0.001**
CoWord p=0.2	0.243	0.085	-	**0.018**	0.372	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**	0.281	**0.001**	**0.001**
CoWord p=0.3	0.103	**0.007**	**0.018**	-	**0.049**	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**	0.080	0.107	**0.002**
Adapted D&R	0.192	0.124	0.372	**0.049**	-	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**	0.243	**0.002**	**0.001**
Fine-tuning e=1	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**	-	0.164	0.340	**0.001**	**0.001**	0.244	**0.001**	**0.001**	**0.001**
Fine-tuning e=2	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**	0.164	-	0.217	**0.001**	**0.002**	0.138	**0.001**	**0.001**	**0.001**
Fine-tuning e=5	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**	0.340	0.217	-	**0.002**	**0.002**	0.224	**0.001**	**0.001**	**0.001**
MaxPCXMI e=1	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**	**0.002**	-	0.321	**0.001**	**0.001**	**0.005**	0.161
MaxPCXMI e=2	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**	**0.002**	**0.002**	0.321	-	**0.001**	**0.001**	**0.002**	0.143
MaxPCXMI e=5	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**	0.243	0.138	0.224	**0.001**	**0.001**	-	**0.001**	**0.001**	**0.003**
Weighting λ=2	0.319	0.069	0.281	0.080	0.243	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**	-	**0.001**	**0.001**
Weighting λ=5	**0.001**	**0.001**	**0.001**	0.107	**0.002**	**0.001**	**0.001**	**0.001**	**0.005**	**0.002**	**0.001**	**0.001**	-	**0.001**
Weighting λ=10	**0.001**	**0.001**	**0.001**	**0.002**	**0.001**	**0.001**	**0.001**	**0.001**	0.161	0.143	**0.003**	**0.001**	**0.001**	-

Table 19: The p-values of the paired bootstrapping of the results in terms of COMET on the IWSLT2017 English-to-German testset for each pair of the models based on OpusMT en-de. Values <0.05 are in bold.

Model	Baseline	CoWord p=0.1	CoWord p=0.2	CoWord p=0.3	Adapted D&R	Fine-tuning e=1	Fine-tuning e=2	Fine-tuning e=5	MaxPCXMI e=1	MaxPCXMI e=2	MaxPCXMI e=5	Weighting λ=2	Weighting λ=5	Weighting λ=10
Baseline	-	**0.003**	**0.006**	0.122	**0.050**	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**	0.138	0.089	**0.002**
CoWord p=0.1	**0.003**	-	0.291	0.065	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**
CoWord p=0.2	**0.006**	0.291	-	**0.009**	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**	**0.002**	**0.001**	**0.001**
CoWord p=0.3	0.122	0.065	**0.009**	-	**0.015**	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**	0.066	**0.043**	**0.001**
Adapted D&R	**0.050**	**0.001**	**0.001**	**0.015**	-	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**	0.127	0.191	0.068
Fine-tuning e=1	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**	-	0.317	0.141	**0.002**	**0.043**	0.087	**0.001**	**0.001**	**0.001**
Fine-tuning e=2	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**	0.317	-	0.103	**0.007**	0.064	0.072	**0.001**	**0.001**	**0.001**
Fine-tuning e=5	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**	0.141	0.103	-	**0.002**	**0.011**	0.238	**0.001**	**0.001**	**0.001**
MaxPCXMI e=1	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**	**0.002**	**0.007**	**0.002**	-	**0.038**	**0.001**	**0.001**	**0.001**	**0.003**
MaxPCXMI e=2	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**	**0.043**	0.064	**0.011**	**0.038**	-	**0.001**	**0.001**	**0.001**	**0.001**
MaxPCXMI e=5	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**	0.087	0.072	0.238	**0.001**	**0.001**	-	**0.001**	**0.001**	**0.001**
Weighting λ=2	0.138	**0.001**	**0.002**	0.066	0.127	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**	-	0.170	**0.001**
Weighting λ=5	0.089	**0.001**	**0.001**	**0.043**	0.191	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**	0.170	-	**0.002**
Weighting λ=10	**0.002**	**0.001**	**0.001**	**0.001**	0.068	**0.001**	**0.001**	**0.001**	**0.003**	**0.001**	**0.001**	**0.001**	**0.002**	-

Table 20: The p-values of the paired bootstrapping of the results in terms of ctxPro accuracy of the Gender phenomenon for each pair of the models based on OpusMT en-de. Values <0.05 are in bold.

Model	Baseline	CoWord p=0.1	CoWord p=0.2	CoWord p=0.3	Adapted D&R	Fine-tuning e=1	Fine-tuning e=2	Fine-tuning e=5	MaxPCXMI e=1	MaxPCXMI e=2	MaxPCXMI e=5	Weighting λ=2	Weighting λ=5	Weighting λ=10
Baseline	-	0.159	**0.014**	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**
CoWord p=0.1	0.159	-	**0.011**	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**
CoWord p=0.2	**0.014**	**0.011**	-	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**
CoWord p=0.3	**0.001**	**0.001**	**0.001**	-	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**
Adapted D&R	**0.001**	**0.001**	**0.001**	**0.001**	-	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**
Fine-tuning e=1	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**	-	**0.001**	**0.001**	**0.001**	0.058	**0.001**	**0.001**	**0.001**	**0.001**
Fine-tuning e=2	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**	-	**0.001**	**0.001**	**0.001**	**0.002**	**0.001**	**0.001**	**0.001**
Fine-tuning e=5	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**	-	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**
MaxPCXMI e=1	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**	-	**0.001**	**0.001**	**0.001**	**0.001**	0.166
MaxPCXMI e=2	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**	0.058	**0.001**	**0.001**	**0.001**	-	**0.001**	**0.001**	**0.001**	**0.001**
MaxPCXMI e=5	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**	**0.002**	**0.001**	**0.001**	**0.001**	-	**0.001**	**0.001**	**0.001**
Weighting λ=2	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**	-	**0.001**	**0.001**
Weighting λ=5	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**	-	**0.001**
Weighting λ=10	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**	0.166	**0.001**	**0.001**	**0.001**	**0.001**	-

Table 21: The p-values of the paired bootstrapping of the results in terms of ctxPro accuracy of the Formality phenomenon for each pair of the models based on OpusMT en-de. Values <0.05 are in bold.

Model	Baseline	CoWord p=0.1	CoWord p=0.2	CoWord p=0.3	Adapted D&R	Fine-tuning e=1	Fine-tuning e=2	Fine-tuning e=5	MaxPCXMI e=1	MaxPCXMI e=2	MaxPCXMI e=5	Weighting λ=2	Weighting λ=5	Weighting λ=10
Baseline	-	**0.002**	**0.002**	**0.027**	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**
CoWord p=0.1	**0.002**	-	0.269	0.110	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**
CoWord p=0.2	**0.002**	0.269	-	**0.019**	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**
CoWord p=0.3	**0.027**	0.110	**0.019**	-	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**
Adapted D&R	**0.001**	**0.001**	**0.001**	**0.001**	-	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**
Fine-tuning e=1	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**	-	0.057	**0.001**	**0.001**	**0.001**	**0.001**	**0.031**	**0.017**	**0.001**
Fine-tuning e=2	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**	0.057	-	**0.001**	**0.001**	**0.001**	**0.001**	**0.015**	**0.045**	**0.001**
Fine-tuning e=5	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**	-	**0.001**	**0.001**	**0.001**	**0.001**	0.356	**0.001**
MaxPCXMI e=1	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**	-	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**
MaxPCXMI e=2	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**	-	**0.001**	**0.001**	**0.001**	**0.001**
MaxPCXMI e=5	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**	-	**0.001**	**0.001**	**0.001**
Weighting λ=2	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**	**0.031**	**0.015**	**0.001**	**0.001**	**0.001**	**0.001**	-	**0.001**	**0.001**
Weighting λ=5	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**	**0.017**	**0.045**	0.356	**0.001**	**0.001**	**0.001**	**0.001**	-	**0.001**
Weighting λ=10	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**	-

Table 22: The p-values of the paired bootstrapping of the results in terms of ctxPro accuracy of the Auxiliary phenomenon for each pair of the models based on OpusMT en-de. Values <0.05 are in bold.

Model	Baseline	CoWord p=0.1	CoWord p=0.2	CoWord p=0.3	Adapted D&R	Fine-tuning e=1	Fine-tuning e=2	Fine-tuning e=5	MaxPCXMI e=1	MaxPCXMI e=2	MaxPCXMI e=5	Weighting λ=2	Weighting λ=5	Weighting λ=10
Baseline	-	**0.001**	**0.001**	**0.001**	**0.031**	**0.001**	**0.001**	**0.010**	0.364	0.108	**0.012**	**0.002**	**0.001**	**0.001**
CoWord p=0.1	**0.001**	-	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**	0.135	**0.016**
CoWord p=0.2	**0.001**	**0.001**	-	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**	0.155
CoWord p=0.3	**0.001**	**0.001**	**0.001**	-	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**	**0.002**
Adapted D&R	**0.031**	**0.001**	**0.001**	**0.001**	-	**0.001**	**0.001**	**0.001**	0.103	0.352	0.109	**0.046**	**0.001**	**0.001**
Fine-tuning e=1	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**	-	0.097	**0.006**	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**
Fine-tuning e=2	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**	0.097	-	**0.029**	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**
Fine-tuning e=5	**0.010**	**0.001**	**0.001**	**0.001**	**0.001**	**0.006**	**0.029**	-	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**
MaxPCXMI e=1	0.364	**0.001**	**0.001**	**0.001**	0.103	**0.001**	**0.001**	**0.001**	-	**0.005**	**0.001**	**0.017**	**0.001**	**0.001**
MaxPCXMI e=2	0.108	**0.001**	**0.001**	**0.001**	0.352	**0.001**	**0.001**	**0.001**	**0.005**	-	**0.002**	0.092	**0.001**	**0.001**
MaxPCXMI e=5	**0.012**	**0.001**	**0.001**	**0.001**	0.109	**0.001**	**0.001**	**0.001**	**0.001**	**0.002**	-	0.421	**0.003**	**0.001**
Weighting λ=2	**0.002**	**0.001**	**0.001**	**0.001**	**0.046**	**0.001**	**0.001**	**0.001**	**0.017**	0.092	0.421	-	**0.001**	**0.001**
Weighting λ=5	**0.001**	0.135	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**	**0.003**	**0.001**	-	**0.001**
Weighting λ=10	**0.001**	**0.016**	0.155	**0.002**	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**	-