PBML

The Prague Bulletin of Mathematical Linguistics

NUMBER 122 DECEMBER 2024

EDITORIAL BOARD

Published twice a year by Charles University (Prague, Czech Republic)

Editorial office and subscription inquiries:
ÚFAL MFF UK, Malostranské náměstí 25, 118 00, Prague 1, Czech Republic
E-mail: `pbml@ufal.mff.cuni.cz`

ISSN 0032-6585
eISSN 1804-0462

PBML

The Prague Bulletin of Mathematical Linguistics

NUMBER 122 DECEMBER 2024

CONTENTS

Articles

PBML

The Prague Bulletin of Mathematical Linguistics

NUMBER 122 DECEMBER 2024

EDITORIAL

Editorial

Everything that has a beginning has an end.
Titus Maccius Plautus,
a Roman playwright of the Old Latin period

The Prague Bulletin of Mathematical Linguistics (PBML) had its beginning in 1964, one year before the start of the series of regular international meetings called COLING and the regular conferences of ACL. In the history of Czechoslovak computational linguistics (at that time called by its Praguian followers “algebraic”, inspired by the logician Y. Bar Hillel, but soon after changing this name to a more internationally common “mathematical”, and later, when computers entered the scene more vigorously, “computational”), this may be considered a milestone of the field (CL in the sequel). The start of PBML was accompanied (on September 18–22, 1964) by an international meeting (“colloquium”) of researchers involved in the newly arising branch of science, which may be claimed to be the first such meeting in our geographical area. No wonder that not many of the participants of the colloquium are among us any more (to name just a few who have already left: Helmut Schnelle, Manfred Bierwisch, Hans Karlgren, Bernard Vauquois, Ferenc Kiefer, Ferenc Papp, Laszlo Kalmár). Also the initiator and for a long time the editor-in-chief of the Bulletin and the main organizer of the 1964 colloquium Petr Sgall, who was the founder of the field of CL in Czechoslovakia, died on May 28, 2019.

Since those who introduced the study and work in CL in Czechoslovakia were linguists, it is no wonder that the first small CL group started its activities at the Faculty of Arts of Charles University (CU) around 1959. However, as they had good personal links to colleagues from the field of logic and mathematics, a still smaller parallel group of Sgall’s students found their place at the Faculty of Mathematics and Physics of CU. The two groups led by Petr Sgall joined together at the beginning of 1968 in the so-called Laboratory of Algebraic Linguistics under the roof of the Faculty of Arts. Due to the Soviet invasion in August of the same year and to the fact that none of the members of this Lab expressed their support for the new political situation, the very existence of the Lab was endangered, Petr Sgall was supposed to leave the University and the other members (at that time they were 15 in total) were also supposed to leave the academic area. However, thanks to our colleagues and friends at the Faculty of Mathematics and Physics (FMP) and also to the fact that the political situation there was not so strict as that at the Faculty of Arts, the members of the Lab found their shelter at that Faculty, not as a group or department but as individual faculty members, facing, of course, restrictions as for their teaching, international contacts and academic degrees. The situation gradually became more favorable; at the beginning of the eighties, there was a possibility to create an unofficial group at the Dept. of Applied Mathematics, and after the "velvet revolution" in 1989, an independent institute was established at FMP called the Institute of Formal and Applied Linguistics (ÚFAL) chaired by Petr Sgall. The chairs of the Institute of course changed during the last 34 years, but the Institute is still there, and the Prague Bulletin of Mathematical Linguistics as well. It should be added in this connection that the existence of the Bulletin was also severely endangered after the Soviet invasion, but it was thanks to our colleague and friend, the logician Professor Karel Berka, a member of the Communist Party but a very devoted friend of ours, who was willing to take over the responsibility and to chair the Editorial Board and thus saved PBML's existence. The role of PBML was more than crucial. Not only that it was almost the only publication forum in the country where the results of our research in CL could be published and thus disseminated internationally, but also, and most importantly, the journal was a very valuable material for exchange at the time when no financial resources were available for buying books or journals published abroad or to pay for mimeographed research papers offered by such activities as those of the Indiana University Linguistics Club at Bloomington University, Indiana, etc. In this way, our CL community has not lost awareness of the most recent developments of the field and kept international contacts alive.

Though the original intention of the first PBML editor-in-chief Petr Sgall and the Editorial Board members (a general linguist Pavel Novák, a quantitative linguistics specialist Marie Těšitelová and a formal logician Pavel Berka) was to provide a forum for the newly arising local CL community to introduce their research results, PBML had soon acquired status of an internationally well-known CL forum for both theoretically- and application-minded researchers from all corners of the world. This was also reflected by the reorganization of the editorial board into a body of internationally recognized researchers in 1997, thus ensuring a high scientific quality of the PBML's contents.

However, with the expansion of the field in the recent years (even before the deep learning revolution), there are many more opportunities to publish computational linguistics research worldwide, many of which are also recognized by national authorities as "worthwhile" when considering grant applications or personal promotions. Despite some effort, PBML was unable to reach that status, apparently reflected by the low number of submissions, which in turn made it harder to submit applications to worldwide indexes. While we **deeply** regret it, we are now closing the operation, hoping that researchers can publish under similar conditions (Open Access, free of fees) in other more recognized venues, such as TACL or CL.

With many sincere thanks to our Editorial Board members and my assistants, Dr. Jana Hamrlová and Dr. Martin Popel of ÚFAL (Computer Science School, Charles University), I wish our prospective contributors good luck in pursuing science in the field of Computational Linguistics and success in publishing its results.

Jan Hajič
Editor-in-chief
PBML

PBML

The Prague Bulletin of Mathematical Linguistics

NUMBER 122 DECEMBER 2024 9–42

Improving Fuzzy Match Augmented Neural Machine Translation in Specialised Domains through Synthetic Data

Arda Tezcan, Alina Skidanova, Thomas Moerman

Language and Translation Technology Team (LT3), Ghent University, Belgium

Abstract

Previous studies have demonstrated the effectiveness of fuzzy match (FM) augmentation in improving the performance of Neural Machine Translation (NMT) models. However, this approach exhibits limitations when applied to scenarios where limited parallel datasets are available for NMT training. This study investigates the effectiveness of leveraging additional monolingual data to improve FM-augmented NMT performance by generating synthetic parallel datasets in domain-specific scenarios. To this end, we adopt a simple strategy for combining two data augmentation methods for NMT, namely back-translation and Neural Fuzzy Repair (NFR). Experiments conducted on three language directions, namely English→Ukrainian, English→French and French→English, two domains and various dataset sizes show that this simple approach yields significant and substantial improvements in estimated translation quality.

1. Introduction

In recent years, the field of neural machine translation (NMT) has undergone rapid advancements, first with the emergence of the (encoder-decoder) transformer models (Vaswani et al., 2017), and more recently with the (decoder-only) large language models (LLMs), exemplified by BLOOM (Scao et al., 2022), Mistral (Jiang et al., 2023), and Llama 3 (Dubey et al., 2024). Despite the growing enthusiasm for utilising LLMs for MT and the additional capabilities they possess over specialised NMT models, such as instruction following (Ouyang et al., 2022; Wei et al., 2022), their adoption does not guarantee superior performance in translation tasks, especially in specialised domains (Kocmi et al., 2023; Jiao et al., 2023; Peng et al., 2023; Son and Kim, 2023).

Corresponding author: arda.tezcan@ugent.be

Cite as: Arda Tezcan, Alina Skidanova, Thomas Moerman. Improving Fuzzy Match Augmented Neural Machine Translation in Specialised Domains through Synthetic Data. The Prague Bulletin of Mathematical Linguistics No. 122, 2024, pp. 9–42. doi: 10.14712/00326585.030.

In domain-specific scenarios, specialised NMT systems, as well as LLMs, have demonstrated a capacity to leverage translations of similar sentences retrieved from the training data or external databases (also referred to as 'fuzzy matches'; FMs) effectively, resulting in remarkable gains in translation quality (Bulté and Tezcan, 2019; Xu et al., 2020; Khandelwal et al., 2021; Moslem et al., 2023a). Despite the differences in the way FMs are utilised by existing approaches, the fundamental concept unifying all of them lies in their capacity to steer the MT output towards translations of retrieved FMs.

In the context of specialised NMT models, previous studies showed that FM-augmented NMT models attain their maximum potential in high-resource, domain-specific scenarios characterised by the availability of large bilingual datasets, which enhance the likelihood of retrieving FMs with higher similarity levels (Bulté and Tezcan, 2019; Tezcan and Bulté, 2022; Xu et al., 2023; Reheman et al., 2023). To address this limitation, some efforts have been undertaken to leverage additional monolingual data in the target language for directly retrieving similar translations through employing multilingual sentence embeddings, resulting in further improvements in translation quality (Cai et al., 2021; Tamura et al., 2023). More related to our work, in the context of general-domain scenarios, Pham et al. (2020) and Xu et al. (2021) proposed a simple yet novel approach for leveraging additional monolingual data in the target language for FM augmentation where synthetic source sentences are generated through back-translation in the first place. However, this approach showed mixed results regarding its impact on translation performance. As both of these studies acknowledged, the challenge of effectively utilising this approach in general domain scenarios is finding highly similar translations for a given input (high FMs).

Following up on previous work, we consider FM augmentation through the generation of synthetic source sentences more suitable for domain-specific scenarios, which are focused on specific subject areas characterised by high levels of repetitiveness in vocabulary, structure, and style. To this end, in this study, we adopt previously proposed methods for in-domain scenarios by combining two data augmentation techniques for NMT: (i) back-translation, a commonly used technique for generating synthetic data in the source language from monolingual data in the target language (Sennrich et al., 2016), and (ii) 'Neural Fuzzy Repair' (NFR), which integrates FMs into NMT through concatenating source sentences with translations of retrieved FMs (Bulté and Tezcan, 2019).

Our experimental results, spanning three language pairs and two specialised domains, demonstrate that combining the two data augmentation approaches yields significant improvements in estimated translation quality in all tested settings. Additionally, we present insights into the effectiveness of this approach by employing reduced sizes of bilingual and additional monolingual datasets and contrast it with state-of-the-art LLMs, as well as NMT systems trained under an alternative scenario, where high-quality translations for the additional monolingual data are available.

2. Related Research

Within the domain of FM-augmented NMT, various approaches have been implemented in the past, resulting in enhanced translation performance. Some examples include integrating FMs to the transformer-based NMT architectures through modifying the decoding process (Cao and Xiong, 2018; Gu et al., 2018; Khandelwal et al., 2021; Reheman et al., 2023), adding a lexical memory to the NMT architecture (Feng et al., 2017), attaching rewards for matched translation pieces from FMs into the NMT output layer (Zhang et al., 2018), introducing additional attention layers to capture relevant information from translation memories (TMs) (He et al., 2021), or modifying the whole architecture, enabling it to edit FMs to obtain a final translation (Gu et al., 2019; Bouthors et al., 2023).

Whereas most of the approaches that utilise FMs for NMT require modifications to the NMT architectures or decoding algorithms, FMs have also been successfully integrated into NMT through data augmentation techniques (Bulté and Tezcan, 2019; Xu et al., 2020; Tezcan et al., 2021). These studies vary in their approaches to measuring FM similarity, employing N-best FMs, or combining FMs with different characteristics. Nonetheless, a shared characteristic among these studies is the reliance on seeking FMs through source text similarity and augmenting source sentences during training and inference times with the translations of retrieved FMs in the target language.

Previous studies have shown that the quality of retrieved samples plays a crucial role in the effectiveness of FM augmentation. Specifically, the translation quality of FM-augmented models improves as the similarity of the retrieved FMs to the input sentence increases, with optimal results observed in high-resource scenarios (Bulté and Tezcan, 2019; Tezcan and Bulté, 2022; Xu et al., 2023; Reheman et al., 2023). As a result, in the domain of European legislation—a dataset also employed in this study—Bulté and Tezcan (2019) found that the NFR approach starts to become effective with at least 300K sentence pairs as training data.

In order to extend the capabilities of FM-augmented NMT systems beyond their reliance on parallel sentences for FM retrieval, some studies proposed leveraging additional monolingual data in the target language. For example, Cai et al. (2021) used sentence encoders to measure the similarity between sentences from the source and target languages and conditioned the translation model on both the retrieved FMs from the target language and the input from the source language to generate translations. On the other hand, Tamura et al. (2023) expanded the usefulness of the NFR approach by leveraging additional monolingual data. To this end, they trained NFR models as proposed by Bulté and Tezcan (2019) but extended the pool of sentences for FM retrieval and augmentation with the additional monolingual data during inference by measuring the similarity of source sentences with sentences in the target language through multilingual sentence embeddings. Both studies, which used additional monolingual data for FM augmentation, reported significant improvements in translation performance. In another relevant study, multilingual sentence embeddings have also been used to support translators by retrieving FMs from additional monolingual data in a TM-based computer-assisted translation environment (Esplà-Gomis et al., 2022).

Apart from these studies, a substantial body of literature exists on leveraging monolingual data in the target language to improve the translation quality of NMT systems. Some effective approaches include incorporating target-side language models into the NMT decoding step (Gulcehre et al., 2015) or for re-ranking the NMT output (Jean et al., 2015), as well as leveraging additional monolingual data through *back-translation*. In this process, a reverse NMT model is trained on existing parallel data to translate monolingual target-language data into the source language, thereby generating additional synthetic parallel training data. (Sennrich et al., 2016; Fadaee et al., 2017; Edunov et al., 2018). Notably, Xu et al. (2019) have further demonstrated that the positive effect of the synthetic training data generated through back-translation on NMT performance gradually waned with increasing sizes due to the noisy nature of source sentences. While other researchers have drawn similar conclusions regarding the performance decrease with increasing ratios of high-quality to synthetic training dataset sizes, the optimal ratios varied considerably across different studies, typically ranging from 1:1 to 1:5 (Sennrich et al., 2016; Edunov et al., 2018; Ng et al., 2019).

Our work stems from the technique explored by Pham et al. (2020) and Xu et al. (2021), which leverages additional monolingual data in the target language for FM augmentation with a different strategy. Both studies extended the pool of source sentences for FM retrieval and augmentation with the synthetically generated source sentences via back-translation, as well as using the same sentence pairs as extra training data, applying this approach in general domain settings. In one set of experiments, for the English→French language direction, this approach yielded limited gains in translation performance compared to FM-augmented NMT systems using only the available parallel datasets, consisting of approx. 4.5M sentence pairs from different domains, resulting in +0.6 and +1.8 BLEU scores in the news and Wikipedia domains, respectively (Pham et al., 2020). Notably, to achieve the improvements from additionally employing the synthetically generated data for FM augmentation, approx. 83.5M (news) and 6.5M (Wikipedia) monolingual sentences in the target language were required. Furthermore, in the same study, this approach was outperformed by a standard NMT system trained only on the original parallel sentences (without FM augmentation) in the news domain (-0.6 BLEU). Overall, despite having access to extensive monolingual datasets, FM-augmentation with backtranslated sentences in these general domain experiments resulted in limited effectiveness.

In a second set of experiments, this approach was applied to the news domain and for domain adaptation, using 10M additional sentences in the target languages[1] (Xu et al., 2021). When applied to the news domain, this approach generally resulted in lower BLEU scores compared to simply using synthetically generated datasets via back-translating target sentences, as additional training data (-2.6 BLEU and -0.1 BLEU on the WMT'19 translation test set; and -3.6 BLEU and +0.2 BLEU on the WMT'20 translation test set for the French→German and German→French language directions, respectively). Similar to the findings of Pham et al. (2020), Xu et al. (2021) demonstrated that FM-augmentation with back-translated sentences for NMT was generally ineffective for improving MT performance in general domain settings[2].

In the same study, this approach was also used for domain adaptation for the German→French language direction, where the bilingual and monolingual news data was used for training and FM augmentation for translating a test set extracted from the European Central Bank (ECB) corpus. The experiments showed that using back-translated sentences (paired with hand-crafted target sentences) both as extra training data and additionally for FM augmentation was detrimental to translation performance, yielding -0.8 BLEU scores for both types of systems compared to a standard NMT system trained on the available parallel datasets (without FM augmentation). Further experiments on increasing the minimum similarity threshold up to $\lambda \geq 0.85$ for FM retrieval also resulted in mixed outcomes in both experiments.

In addition to the aforementioned studies that focus on improving specialised NMT models, FMs and previously seen translations in the same domain as the input have recently been utilised to improve the translation quality of LLMs, by integrating them into the prompts used for generating translations, a method referred to as in-context learning (Mu et al., 2023; Moslem et al., 2023a), and into the fine-tuning process (Alves et al., 2023; Moslem et al., 2023b).

3. Methodology

3.1. Neural Fuzzy Repair

For implementing NFR, we followed the work of Tezcan et al. (2021)[3]: for a given bilingual dataset, consisting of source/target sentence pairs S, T, we augmented each source sentence $s_i \in S$ with the translations $\{t_1, \ldots, t_n\} \in T$ of the most similar

[1] While the authors stated that they used all available parallel data for the WMT'21 translation task with the exception of the ParaCrawl data, the exact size of the parallel data used for training was not explicitly provided.

[2] Neither study included statistical significance analyses for these reported differences in estimated translation performance.

[3] `https://github.com/lt3/nfr`

source sentence in the same dataset[4] $\{s_1, \ldots, s_n\} \in S$ (i.e., fuzzy match; FM), where $s_i \notin \{s_1, \ldots, s_n\}$, given that the FM score is sufficiently high (i.e., above the given threshold): $\lambda \geq 0.5$. To this end, we measured FM score $FM(s_i, s_j)$ between two source sentences s_i and s_j as the cosine similarity between their sentence embeddings e_i and e_j:

$$FM(s_i, s_j) = \frac{e_i \cdot e_j}{\|e_i\| \times \|e_j\|} \tag{1}$$

where $\|e\|$ is the magnitude of vector e.

To generate sentence embeddings, we used sent2vec (Pagliardini et al., 2018), and for efficient retrieval of FMs, we built a FAISS index (Johnson et al., 2021). The hyperparameters used for generating sentence embeddings and building the FAISS index are provided in Appendices A.1 and A.2, respectively. Prior to retrieving FMs, all sentences were segmented into sub-words using SentencePiece (Kudo and Richardson, 2018), using the XLM-RoBERTa (base) tokenizer[5]. An example of the FM retrieval and data augmentation process is provided in Table 1.

S	Debt, breakdown by **residual** maturity
$score$	0.9812
FM_S	Debt, breakdown by **initial** maturity
FM_T	Dette, ventilation par échéance **initiale**
S'	Debt, breakdown by **residual** maturity $< sep >$ Dette, ventilation par échéance **initiale**
T	Dette, ventilation par échéance **résiduelle**

Table 1. An example of FM retrieval and source augmentation (S') for a given source sentence (S) for the EN→FR language direction, with the translation 'T'. 'FM_S' and 'FM_T' refer to the source and target sides of the retrieved FM, respectively. The sentence similarity score is indicated as '$score$'.

The NFR model is trained with an off-the-shelf NMT toolkit, namely the OpenNMT-py toolkit (Klein et al., 2017), using the combined dataset, which consists of the original and the augmented source/target sentence pairs S, T and S', T, respectively. Combining the original parallel data with the source-augmented parallel data allows the NFR model to handle both the augmented and non-augmented source sentences as input (Bulté and Tezcan, 2019). Each source sentence is augmented at inference time using the same FM retrieval method described above. Following previous work,

[4] We used "@@@" as a separator between the source sentence and the translation of the retrieved FM.

[5] https://huggingface.co/docs/transformers/v4.22.2/en/model_doc/xlm-roberta#overview

we used a minimum similarity threshold of $\lambda \geq 0.5$ for FM retrieval (Tezcan et al., 2021). If no FMs are found with a match score above this threshold, the original, non-augmented source sentence is used as input to the FM-augmented NMT model. While different minimum similarity thresholds have been tested in previous studies (Pham et al., 2020; Xu et al., 2020; Tezcan et al., 2021), we keep this value fixed in this study.

3.2. Combining Neural Fuzzy Repair and back-translation

In the context of NMT, NFR and back-translation can be considered complementary data augmentation techniques. While NFR aims to use existing training data more efficiently by steering the MT output for a given input towards the translation of the most similar sentence found in the same data, back-translation aims to generate additional parallel data for training, where the source side consists of synthetically generated sentences. Therefore, following the work of Bulté and Tezcan (2019), Pham et al. (2020), and Xu et al. (2021), by combining these two data augmentation approaches, we expect to leverage gains in translation quality in domain-specific scenarios from two angles: first, by expanding the pool of source sentences for retrieving FMs with high similarity, even if they are partially synthetic; and subsequently, by generating additional, synthetic training data, which is augmented with the translations of retrieved FMs, consisting of non-synthetic, high-quality texts from the target language.

The methodology used in this study, which combines back-translation and NFR, is illustrated in Figure 1. First, we train a back-translation model using the original parallel data to translate in the reverse language direction. Next, we utilise this back-translation model to translate the additional monolingual sentences in the target language to the source language. Finally, we combine the resulting synthetic parallel data with the original training data and implement NFR training and inference, as described in the previous section.

4. Experimental Setup

This section outlines the datasets (Section 4.1), the implementation details of the different NMT systems (Section 4.2), as well as the evaluation methodology used in this study (Section 4.3).

4.1. Data

The first set of experiments was conducted for the English→Ukrainian (EN→UK) language direction, using bilingual and monolingual datasets in the legal domain, collected from the European Language Resource Coordination (ELRC-SHARE) Reposi-

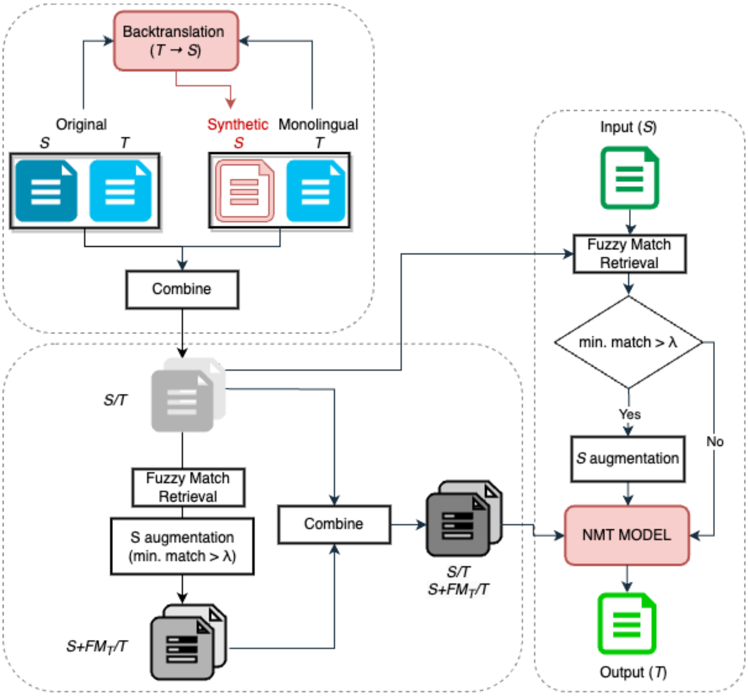


Figure 1. Overview of the methodology used in this study, which utilises the 'Original' source (S) and target (T) sentences to build a back-translation model and translates the additional monolingual data in the target language 'Monolingual T' to source language 'Synthetic S' (top-left), then implements the NFR training (bottom-left), and finally the NFR inference (bottom-right) steps.

tory[6]. The bilingual dataset consisted of the translations of the EU acts into Ukrainian[7] (EN→UK) and the translations of the Ukrainian laws into English, collected from the official web-portal of the Parliament of Ukraine[8,9] (UK→EN). The monolingual Ukrainian data consisted of a random subset of sentences collected from the documents of the official web portal of the Ukrainian parliament[10] and the Legal Ukrainian

[6] `https://elrc-share.eu/`, CC-BY-4.0 license

[7] EU acts in Ukrainian

[8] Abstracts of Ukrainian Laws in English

[9] Ukrainian Laws in English

[10] Legal documents of the Parliament of Ukraine

Crawling Corpus[11], which is built from web documents collected from legislation websites, and governmental sites. The number of sentences in each dataset is provided in Table 2.

Dataset	Language(s)	No. sentences
EU acts	EN–UK	129941
Ukrainian laws	EN–UK	177270
Ukrainian Parliament	UK	665000
Legal Ukrainian Crawling	UK	1000000

Table 2. An overview of the datasets used for the EN→UK experiments.

Prior to training NMT engines, both the additional monolingual and the original bilingual datasets underwent an automatic cleaning process[12]. This entailed removing empty segments, duplicate segments, segments copied from source to target, HTML codes, segments containing more than 100 tokens, and normalising punctuation marks. In addition, sentences that consisted of Russian were removed from the target side of the bilingual data, as well as from the monolingual data[13]. After the automatic cleaning process, the bilingual dataset was randomly partitioned into training, validation and test sets. The randomly selected test sets, consisting of 2000 sentence pairs per language direction, were also manually reviewed to eliminate noisy sentence pairs, such as unaligned sentences, partial translations, sentences consisting only of dates or alphanumeric codes, and sentence pairs with fewer than three tokens on either the source or target side. The additional monolingual data was further utilised for generating synthetic, bilingual training data (EN→UK) through back-translation. The number of segments after cleaning and partitioning the original data is provided in Table 3.

	Train	Validation	Test
Bilingual (EN–UK)	286417	2000	1899
Monolingual (UK)	1461320	–	–

Table 3. The number of sentences used as training, validation and test sets for the EN→UK experiments.

[11] The Legal Ukrainian Crawling Corpus

[12] `filter.py` script from `https://github.com/ymoslem/MT-Preparation`.

[13] `https://github.com/pemistahl/lingua-py`

The second set of experiments was conducted for the English↔French (EN↔FR) language directions. To this end, we used the TM of the European Commission's translation service[14] (DGT-TM) (Steinberger et al., 2012), which consists of texts regarding European legislation, comprising the treaties, regulations and directives adopted by the European Union. The DGT-TM was cleaned using the same steps as described above, and a random subset was collected, consisting of bilingual and monolingual datasets of similar sizes to the EN–UK data, with the monolingual data containing five times more sentences than the bilingual data. Unlike the EN–UK dataset, where the collected monolingual sentences in the target language did not have any translations in the source language, monolingual datasets in EN and FR were extracted from the parallel dataset for EN–FR. Similar to the EN–UK dataset, these extracted monolingual datasets were utilised for generating synthetic bilingual training data through back-translation. The number of sentences in the different partitions of the EN–FR dataset is provided in Table 4.

	Train	**Validation**	**Test**
Bilingual (EN–FR)	300000	2000	1609
Monolingual (FR)	1499436	–	–
Monolingual (EN)	1499436	–	–

Table 4. The number of sentences used as training, validation and test sets for the EN↔FR experiments.

The resulting bilingual datasets, consisting of approximately 300K sentence pairs, is a meaningful starting point to test our hypotheses, as it enables us to evaluate our methodology in a scenario where the NFR approach achieved comparable results to a baseline system trained solely on the original bilingual training data, as previously observed with the DGT datasets (Bulté and Tezcan, 2019). To assess the effectiveness of this approach under varying data conditions, particularly when data resources are scarcer, we conducted additional experiments by gradually reducing the number of sentences in both the bilingual (down to 33% of the original amount) and the monolingual (down to 20% of the original amount) datasets. The resulting datasets used for training the MT systems in this study are available on HuggingFace[15].

4.2. NMT Systems

We trained six types of baseline systems. Among them, four were aimed at assessing the effectiveness of the proposed approach in comparison to existing alternatives

[14] https://opus.nlpl.eu/DGT/corpus/version/DGT

[15] https://huggingface.co/collections/LT3/nfr-bt-nmt-66bcf9db6f39f76a39456df5

in the literature: (i) NMT systems using only the original bilingual data for training (*BASE*); (ii) *NFR* (Tezcan et al., 2021), as described in Section 3.1; (iii) *NFR$_{mono}$*, an adaptation of the NFR approach, which further utilises the additional monolingual data for retrieving FMs during inference (Tamura et al., 2023) using multilingual sentence embeddings generated by LaBSE (Feng et al., 2022); and (iv) *BT*, NMT systems that are trained using a combination of original and synthetic bilingual data, where the synthetic source sentences are generated through back-translation (Sennrich et al., 2016). An overview of the training set sizes used for training these NMT systems is provided in Appendix A.3.

Furthermore, to better understand the limitations of using synthetically generated source sentences compared to an alternative scenario, where large, high-quality parallel datasets are available instead, we trained two additional baseline systems for the EN↔FR language directions: (v) a baseline NMT system, *BASE_HQ*, which utilises high-quality translations for the additional monolingual data in the target language for training, without any additional FM-augmentation (i.e. approx. 1.8M high-quality sentence pairs in total); and (vi) an NFR variant of this system, *NFR_HQ*, which utilises the same high-quality parallel data for training and FM-augmentation. Given that high-quality translations for the monolingual data in the target language were only available for EN↔FR language directions (see Section 4.1), these baseline systems are not available for the EN→UK language direction.

All the systems were trained using the Transformer architecture (Vaswani et al., 2017) and the OpenNMT-py toolkit[16] (Klein et al., 2017). Prior to training, all sentences were segmented into sub-words using SentencePiece, as described in Section 3.1. The resulting vocabulary sizes per system that was trained using all available bilingual and monolingual datasets, per language direction are provided in Appendix A.3. All systems were trained with early stopping with 10 validation rounds in terms of accuracy and perplexity. All training runs were initialised using the same seed. For the systems that did not utilise NFR, the maximum source and target lengths were defined as 200 tokens. The same settings have been used to train NMT systems for back-translation, using the reverse language direction in each case. Maximum source length was doubled to 400 tokens for the systems that utilised NFR, which were trained with augmented source sentences. Other details regarding the hyper-parameters used for training the NMT systems are provided in Appendix A.4.

Finally, using the same test sets, we evaluated four state-of-the-art LLMs for the MT task (Kocmi et al., 2024), namely, GPT4o[17], Llama3.1-Instruct (8b[18] and 70b[19]

[16] `https://github.com/OpenNMT/OpenNMT-py`, v. 3.5.1.

[17] `https://openai.com/index/gpt-4o-system-card/`, translations generated on 27th of October, 2024.

[18] `https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct`

[19] `https://huggingface.co/meta-llama/Llama-3.1-70B-Instruct`

models) (Dubey et al., 2024) and TowerInstruct-Mistral-7b[20] (Alves et al., 2024). Although this study primarily focuses on evaluating the effectiveness of back-translated sentences in improving FM-augmented NMT systems trained from scratch, the comparison offers additional insights into the relative performance of FM-augmented NMT systems alongside state-of-the-art LLMs for MT. From the selected LLMs, while GPT4o and Llama can be regarded as general-purpose multilingual LLMs, TowerInstruct has been further specialized for the MT task through continued pretraining on Mistral using translation data and fine-tuning on translation-relevant instructions. All four LLMs were evaluated on the three language directions, except TowerInstruct for EN→UK, as this language pair is not officially supported.

4.3. Evaluation Methodology

We make use of automated evaluation metrics SacreBLEU[21] (Post, 2018), chrF (Popović, 2015), and COMET[22] (Rei et al., 2020) to assess the quality of the (detokenised) MT output. To verify whether differences between the automated quality metric scores of the different MT systems are statistically significant, we used bootstrap resampling tests (Koehn, 2004). Both the automated evaluations and bootstrap resampling tests have been performed using the MATEO toolkit[23] (Vanroy et al., 2023) with the default settings for each metric.

5. Results

In this section, we first compare the translation performance of the proposed system (*BT+NFR*) with the baseline NMT systems using all the bilingual and monolingual datasets available for training, as well as the LLMs (Section 5.1). Subsequently, we also analyse the effectiveness of this approach using the reduced datasets (Section 5.2).

5.1. System performance with full datasets

Table 5 provides the automated evaluation results for the translations generated by the different MT systems on the corresponding test sets per language direction.

Firstly, looking at the four baseline systems that were trained using the original datasets (upper section), we see that *BT* leads to consistent improvements for all language directions and all metrics over the *BASE* system, confirming the usefulness of back-translation in scenarios where additional monolingual datasets are available in

[20] `https://huggingface.co/Unbabel/TowerInstruct-Mistral-7B-v0.2`

[21] `https://github.com/mjpost/sacrebleu`, v. 2.4.1. (SacreBLEU and chrF)

[22] `https://huggingface.co/Unbabel/wmt22-comet-da`

[23] `https://mateo.ivdnt.org/`

System	**EN→UK**			**EN→FR**			**FR→EN**		
	BLEU	chrF	COMET	BLEU	chrF	COMET	BLEU	chrF	COMET
BASE	54.85	75.95	91.23	51.50	71.21	84.22	53.95	71.61	85.01
BT (*Sennrich et al., 2016*)	56.21	76.93	92.18	54.88	73.34	85.27	56.99	74.23	86.72
NFR (*Tezcan et al., 2021*)	57.73	77.52	91.78	52.67	71.82	84.50	54.43	71.94	85.26
NFR_{mono} (*Tamura et al., 2023*)	60.39	78.89	92.03	52.39	71.68	84.45	55.54	72.62	85.44
BT+NFR (*This work*)	**66.95**	**82.39**	**92.78**	**61.91**	**77.54**	**87.13**	**64.69**	**78.65**	**87.79**
BASE_HQ	–	–	–	59.13	76.28	87.40	61.56	77.04	88.03
NFR_HQ	–	–	–	64.58	79.32	88.36	67.75	80.53	88.61
GPT4o	41.66	67.95	92.35	43.47	67.93	86.75	44.58	68.82	86.87
TowerInstruct-Mistral-7b	–	–	–	39.99	64.66	83.06	42.11	52.14	82.32
Llama3.1-Instruct-8b	18.12	52.77	85.27	26.44	57.96	80.39	27.74	57.94	81.19
Llama3.1-Instruct-70b	27.81	60.14	88.70	44.90	68.57	86.37	49.37	70.49	86.56

Table 5. Results of the automatic evaluations performed on systems using all available datasets.

the target language. While the NFR approach leads to consistent improvements over *BASE* for all metrics for the EN→UK language direction, it only leads to marginal gains for the EN↔FR language directions. The performance of the NFR approach for the DGT datasets is in line with previous research, which showed that the NFR approach did not yield notable improvements with similar training set sizes (Bulté and Tezcan, 2019). While *BT* outperforms *NFR* for the EN↔FR language directions, an opposite observation can be made for the EN→UK language direction, with the exception of COMET scores. Considering all the baseline systems, NFR_{mono} outperforms *NFR* and *BASE* with respect to all metrics for EN→UK, while yielding mixed results for EN↔FR.

Secondly, when we compare the results of *BT+NFR* with the best-performing baseline system per metric, per language direction (upper section), a clear trend emerges: *BT+NFR* consistently outperforms all baseline systems for all language directions and metrics, with improvements of +6.56, +7.03, +7.70 BLEU points over the best baseline system for EN→UK, EN→FR, and FR→EN, respectively. For all language directions, the improvements achieved by *BT+NFR* over all baseline systems are measured to be statistically significant, with $p < 0.001$. Compared to the baseline systems that only utilise the original bilingual datasets (*BASE*), *BT+NFR* yields improvements of up to +12.10 BLEU points (EN→UK).

Thirdly, by comparing the performance of *BT+NFR* to two systems that can be trained in an alternative scenario, where high-quality translations for the monolingual sentences are available in the source language (middle section), we can make two important observations. On the one hand, employing high-quality bilingual datasets alongside FM-augmentation (*NFR_HQ*) results in optimal translation performance for both language directions and across all metrics. While these results highlight the constraints associated with employing synthetically generated source texts alongside the NFR approach, they also offer a clear indication of the upper boundary that can be achieved in terms of translation quality when training NMT models from scratch using these datasets. Consequently, using 1.5M synthetically generated source sentences, instead of high-quality translations, results in decreased MT performance, with reductions of -2.67 BLEU, -1.78 chrF, and -1.23 COMET scores for EN→FR, and -3.06 BLEU, -1.88 chrF, and -0.82 COMET scores for FR→EN. These differences are measured to be statistically significant, with $p < 0.001$.

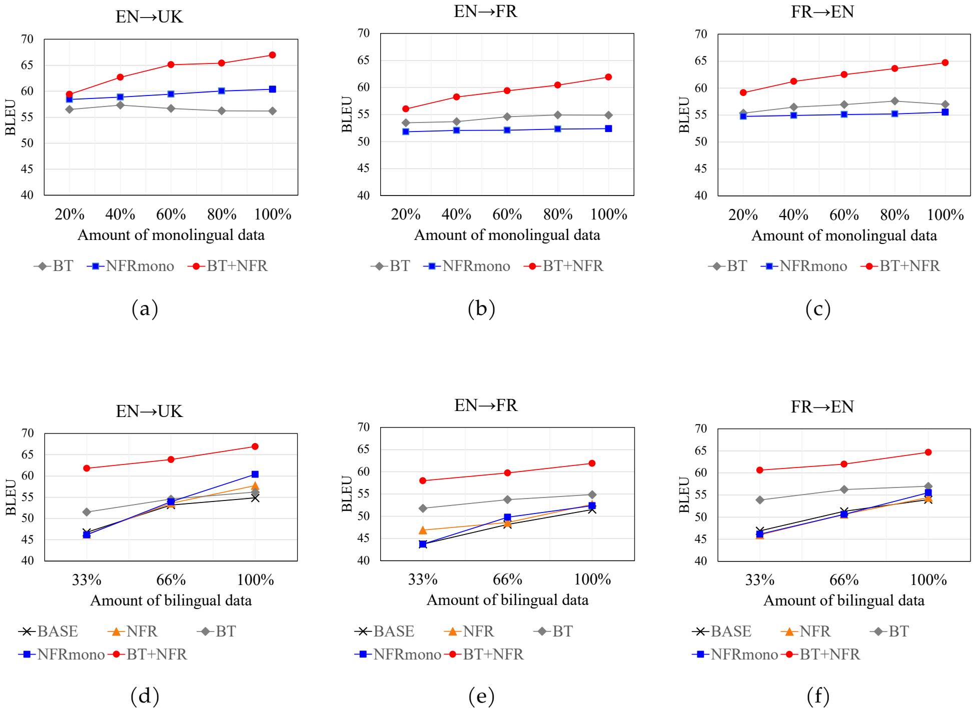
On the other hand, *BT+NFR* surpasses *BASE_HQ* in BLEU and chrF scores for both language directions but does not yield higher COMET scores. Given the disagreement among the three metrics, we can argue that utilising back-translation and FM-augmentation with limited high-quality bilingual data alongside additional monolingual data in the target language produces results comparable to those of a conventional (i.e., non-augmented) NMT system trained on a large, high-quality dataset. In this particular scenario, by using only a monolingual dataset of five times the size of the bilingual data, this combined approach achieved MT performance comparable to that of a conventional NMT system requiring the same amount of additional bilingual data.

Finally, in the lower section, we present the MT performance of four LLMs. Comparative analysis shows that *GPT4o* achieves the best results across all metrics for EN→UK, while it performs similarly to *Llama3.1-Instruct-70b* for EN↔FR, given the mixed rankings each model attains per metric. The larger Llama model also shows clear improvements over the smaller model (70b vs. 8b). Additionally, despite having a similar parameter count, *TowerInstruct-Mistral-7b* performs noticeably better than *Llama3.1-Instruct-8b*.

BT+NFR outperforms all four LLMs in every setting, with improvements that are more pronounced in BLEU and chrF scores than in COMET across all language directions. For each language direction and metric, the improvements over the best-performing LLM are statistically significant, with $p < 0.001$ for BLEU and chrF, and $p < 0.005$ for COMET scores. Notably, *BASE*, which is trained with approx. 300K sentence pairs per language direction, also outperforms all LLMs in terms of BLEU and chrF scores, while being outperformed only by *GPT4o* and *Llama3.1-Instruct-70b* in terms of COMET scores.

5.2. System performance with reduced datasets

In Figure 2, we provide the BLEU scores for different systems for all language directions in two specific lower-resource scenarios, where: (i) the original bilingual datasets are combined with gradually decreasing sizes of monolingual datasets in the target language (upper section), and (ii) the original additional monolingual datasets in the target language are combined with gradually decreasing sizes of bilingual datasets (lower section). The BLEU, chrF and COMET scores for each system are further provided in Appendix A.7.



(a) (b) (c)

(d) (e) (f)

Figure 2. Results of BLEU evaluations performed on systems using gradually decreasing sizes of (i) additional monolingual data in the target language (a, b and c), and (ii) the original bilingual data (d, e and f).

In the upper section of Figure 2, we provide the results for the three systems that employ the extra monolingual data in the target language, namely *BT*, *NFR*$_{mono}$ and *BT+NFR*, as the *BASE* and *NFR* systems are not affected from this adaptation. In this figure, we observe that *BT+NFR* outperforms both baseline systems in all data settings, where the size of the additional monolingual data available in the target language is gradually decreased from 100% down to 20% (approx. 300K sentences) for each language direction, resulting in a 1:1 ratio between the high-quality to synthetic training dataset sizes. Moreover, the performance of *BT+NFR* experiences continuous improvement as larger monolingual datasets become available. While *NFR*$_{mono}$ also shows a performance increase with larger datasets, though, at a much slower pace, this pattern is not observed for *BT*, for which the optimal performance begins to diminish starting with ratios of high-quality to synthetic bilingual data sizes ranging from 1:2 to 1:4 (%40 to %80).

In the lower section, alongside the systems mentioned earlier, we provide the BLEU scores for *BASE* and *NFR*, taking into account the adjustments made to the sizes of the bilingual datasets. These results show that *BT+NFR* system consistently outperforms all baseline systems across all settings by a substantial margin, even when the bilingual dataset size is reduced down to 33%, approx. 100K sentence pairs per language direction. It is worth noting that such clear improvements in MT performance are observed even when the reduction in training set sizes also adversely affects the MT performance of the back-translation systems employed within the proposed approach. Please refer to Appendix A.6 for an overview of the MT performances of the back-translation systems trained in this study. The trends observed for the BLEU scores in both analyses are also reflected in chrF and COMET scores for all systems (see Appendix A.7).

For all configurations with reduced monolingual and bilingual data sets, for all language directions and evaluation metrics (see Appendix A.7), the improvements achieved by *BT+NFR* over the best-performing alternative are measured to be statistically significant (with $p < 0.001$, except for one experiment[24]).

6. Discussion

6.1. Comparison of NMT systems with LLMs

In Table 5, we provided the results of the automated evaluations of the translation quality of the NMT systems built from scratch and three LLMs. While this comparison aims to provide additional perspective on the translation performance of the proposed method, these results should be cautiously interpreted due to three factors: (i) the NMT systems in this study use relatively small training datasets of similar sizes per language direction aimed at investigating the effectiveness of FM-augmentation in low(er)-resource settings, with potential for improvements in MT performance if larger datasets were used; (ii) the LLMs we used in our experiments were not fine-tuned with domain-specific translation data, nor did they leverage FMs through in-context learning, both of which could enhance their MT performance (Moslem et al., 2023a; Alves et al., 2023); and (iii) there is a possibility that the test sets used in this study may (fully or partially) be included in the LLMs' training data, potentially resulting in data leakage and inflated translation performance. While fully preventing data leakage in the LLMs used in this study is challenging, future research could aim for a more balanced comparison between these two types of MT approaches.

Despite challenges in achieving a fully fair comparison and mitigating potential data leakage, these experiments demonstrate that, in these specific settings, the NMT

[24] For EN→UK, when 20% of the monolingual data set is used, the improvements *BT+NFR* yields over *NFR_mono* are observed to be significant with with $p < 0.005$ for all evaluation metrics.

systems trained from scratch with the proposed data augmentation approach outperform state-of-the-art LLMs in translation quality across all automated metrics.

While all three metrics confirm that *BT+NFR* achieves superior translation quality compared to the tested LLMs, the improvements in BLEU and chrF scores are significantly greater than those indicated by COMET scores across all comparisons. The notably higher BLEU and chrF scores for *BT+NFR* suggest that this system generates translations that more closely align with the word order and vocabulary of reference translations than the evaluated LLMs, as these metrics reward translations with overlapping word and character n-grams with the reference translations (Papineni et al., 2002; Popović, 2015). COMET, on the other hand, evaluates the translation quality of a given MT output based on its semantic similarity to the reference, without explicitly measuring word and character n-gram overlaps (Rei et al., 2020). The discrepancy between the improvements achieved by *BT+NFR* against the best-performing LLM with respect to BLEU and chrF (large improvements), in comparison to COMET scores (smaller improvements) suggests that while these LLMs do not produce translations that closely match the reference in vocabulary or word order, they maintain a higher accuracy in conveying the correct meaning. Considering the different dimensions of translation quality highlighted by these metrics, a manual assessment by human evaluators with domain expertise and knowledge of field-specific translation guidelines is crucial to accurately capture and evaluate these nuanced aspects.

6.2. FM similarity

In this study, we argue that FM augmentation using synthetically generated source sentences is most beneficial for domain-specific scenarios, where the chances of finding high FMs for a given input would be considered high due to the repetitive nature of such domains. While the results obtained in different experiments demonstrate the clear benefit of this approach in terms of MT performance, to have a better understanding of the measured level of similarity in the datasets we used in our experiments (i.e. cosine similarity between sent2vec embeddings), we analysed the mean, median, and standard deviation values of the similarity scores of the retrieved FMs for the sentences in the test sets, for all language directions. These statistics, which were analysed for *NFR* and *BT+NFR* (see Table 5) are provided in Tables 6 and 7, respectively. For *BT+NFR*, we also calculated the percentage of FMs retrieved from the additional synthetic source sentences generated via back-translation, which accounts for approximately 1.5 million additional sentences in each language direction.

Given the overall high mean and median FM similarity scores, as well as low standard deviation values, for all language directions, these statistics support the hypothesis that specialised domains are better suited for FM augmentation, whether or not additional synthetic datasets are used. This aligns with earlier research showing a strong positive correlation between FM scores and MT performance in FM-augmented MT systems (Bulté and Tezcan, 2019; Xu et al., 2023; Reheman et al., 2023), with the largest

NFR	**EN→UK**	**EN→FR**	**FR→EN**
Mean	0.8705	0.8554	0.8394
Median	0.8635	0.8384	0.8211
St. Dev.	0.0826	0.0853	0.0883

Table 6. FM similarity statistics for the NFR systems.

BT+NFR	**EN→UK**	**EN→FR**	**FR→EN**
Mean	0.8109	0.8030	0.7923
Median	0.8142	0.7990	0.7867
St. Dev.	0.1219	0.1235	0.1252
FMs from BT	63.73%	86.16%	86.23%

Table 7. FM similarity statistics for the BT+NFR systems, as well as the percentage of FMs retrieved from the additional synthetically generated source sentences via back-translation (FMs from BT).

improvements in MT performance occurring when FM scores exceed 0.8 (Tezcan and Bulté, 2022). These results also demonstrate that when back-translated target sentences into source are added to the pool for FM retrieval, in the test sets, a large portion of the FMs are retrieved from these additional sentences for all language directions despite these sentences being synthetically generated.

On the other hand, the FM scores calculated for the two types of systems indicate a noticeable difference. Considering the higher MT performance achieved by *BT+NFR*, it could be expected that, in the test set, the mean and median FM scores for *BT+NFR* would be higher than for *NFR*. However, our measurements indicate the opposite, showing lower mean and median FM scores for *BT+NFR*. We observed that the disparity between the statistics for both types of systems arises from the differences in the datasets used for creating the corresponding sent2vec models (approx. 300K sentences for *NFR* vs. approx. 1.5M additional sentences for *BT+NFR*). This difference in dataset size leads to distinct vector representations, resulting in different FM scores, even for the same FMs retrieved from the original training data in both systems. As a result, direct comparisons across systems that use different datasets become challenging. While the FM scores still provide a good indication of the level of textual similarity in these specialised domains, this observation should be taken into account in future studies, especially when different minimum similarity thresholds are defined for FM retrieval using sent2vec models with different datasets. It should be highlighted that in all FM-augmented systems used in this study, every sentence in the test set was augmented with an FM, using the minimum similarity threshold of $\lambda \geq 0.5$, as described in Section 3.1.

6.3. Impact of FM retrieval on computational costs

We make a final observation regarding the overhead introduced by FM retrieval in the translation process. As FM retrieval in FM-augmented systems is an additional processing step compared to a standard NMT system, it increases the total time required for generating translations. In the scenario where the full datasets are used for FM retrieval and FM augmentation (see Table 5) – the slowest scenario in our experiments – we observed the total time required for generating an output per sentence using the FM-augmented NMT systems (*BT+NFR*) when FM retrieval and inference are combined, to be approximately 1.5 times the inference time of the standard NMT systems[25] (*BASE_HQ*) across all language directions (approx. 0.078 seconds vs. 0.053 seconds, respectively). It should also be noted that while creating the sent2vec model, FAISS indexing and FM retrieval/augmentation on the source side of the training data also incur additional computational costs (approx. 775 seconds, 62 seconds and 314 minutes in our experiments, respectively, when full data sets are used), these steps are performed only once for NMT training and any translations to be generated via FM augmentation.

7. Conclusion

In this study, we adopted a simple yet effective approach for improving FM-augmented NMT in domain-specific scenarios where limited bilingual datasets are accompanied by additional monolingual data in the target language. Following earlier work, the adopted strategy combines two data augmentation techniques for NMT, namely back-translation and neural fuzzy repair (NFR), without modifying the underlying NMT architecture. Our results show that this approach outperforms NMT systems that employ (i) additional back-translated data for training, (ii) FM-augmentation via NFR, and (iii) a variant of NFR, which utilises additional monolingual data for FM retrieval at inference, yielding substantial improvements in estimated translation quality across two domains and three language directions. These results demonstrate that, unlike previous studies that focus on general-domain scenarios, combining FM augmentation with back-translation is a highly effective strategy for improving NMT systems in specialised domains. Additionally, this approach extends the applicability of FM augmentation to scenarios where bilingual datasets are limited, but additional monolingual datasets in the target language are available. In the specific dataset configurations used in this study, by leveraging monolingual datasets five times the size of the original bilingual datasets, this method effectively matched the performance of traditional NMT systems that would typically rely on the same amount of additional bilingual datasets. Moreover, despite the challenges in ensuring a fair comparison, this straightforward data augmentation method allowed us to develop NMT systems

[25] We used an NVIDIA A100-SXM4-80GB GPU for our experiments.

that outperformed state-of-the-art LLMs across all metrics and language directions, affirming the effectiveness of training NMT systems from scratch for specialized domains.

Our analysis of employing smaller sizes of additional monolingual data reveals a positive correlation between the size of such additional data and MT performance, indicating the potential for further improving such FM-augmented NMT systems through access to larger monolingual datasets. However, this hypothesis needs to be confirmed in future studies. Furthermore, the results demonstrate that combining back-translation with FM augmentation remains an effective method for enhancing NMT performance even in scenarios with smaller bilingual datasets, despite the reduction in high-quality training data and the decline in back-translation performance.

The findings of this study raise interesting research questions for future exploration. These include investigating whether (i) similar improvements can be observed by using the same approach through in-context learning methods when using LLMs for MT; (ii) LLMs can effectively replace back-translation models in lower resource scenarios; and (iii) whether LLMs can be further used to improve the performance this approach by generating additional monolingual sentences in the target language.

Limitations

The experiments were only conducted in two domains, albeit for three language pairs. Additional experiments would be required to confirm these results for other language directions and domains. Moreover, we relied on automated MT evaluation metrics only and did not conduct any experiments involving human evaluation of MT quality.

Acknowledgements

The computational resources (Stevin Supercomputer Infrastructure) and services used in this work were provided by the VSC (Flemish Supercomputer Center), which is funded by Ghent University, FWO and the Flemish Government department EWI.

Bibliography

Alves, Duarte, Nuno Guerreiro, João Alves, José Pombal, Ricardo Rei, José de Souza, Pierre Colombo, and Andre Martins. Steering Large Language Models for Machine Translation with Finetuning and In-Context Learning. In Bouamor, Houda, Juan Pino, and Kalika Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 11127–11148, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.744. URL `https://aclanthology.org/2023.findings-emnlp.744`.

Alves, Duarte M., José P. Pombal, Nuno M. Guerreiro, Pedro Henrique Martins, Joao Alves, Amin Farajian, Ben Peters, Ricardo Rei, Patrick Fernandes, Sweta Agrawal, Pierre Colombo, Jos’e G. C. de Souza, and André Martins. Tower: An Open Multilingual Large Language Model for Translation-Related Tasks. *ArXiv*, abs/2402.17733, 2024. URL `https://api.semanticscholar.org/CorpusID:268031976`.

Bouthors, Maxime, Josep Crego, and François Yvon. Towards Example-Based NMT with Multi-Levenshtein Transformers. In Bouamor, Houda, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1830–1846, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.113. URL `https://aclanthology.org/2023.emnlp-main.113`.

Bulté, Bram and Arda Tezcan. Neural Fuzzy Repair: Integrating Fuzzy Matches into Neural Machine Translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1800–1809, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1175. URL `https://aclanthology.org/P19-1175`.

Cai, Deng, Yan Wang, Huayang Li, Wai Lam, and Lemao Liu. Neural Machine Translation with Monolingual Translation Memory. In Zong, Chengqing, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *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 7307–7318, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.567. URL `https://aclanthology.org/2021.acl-long.567`.

Cao, Qian and Deyi Xiong. Encoding gated translation memory into neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3042–3047, 2018. doi: 10.18653/v1/D18-1340.

Dubey, Abhimanyu, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab Al-Badawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Mathew Oldham, Mathieu Rita,

Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaoqing Ellen Tan, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aaron Grattafiori, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alex Vaughan, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Franco, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, Danny Wyatt, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Firat Ozgenel, Francesco Caggioni, Francisco Guzmán, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Govind Thattai, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Karthik Prasad, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kun Huang, Kunal Chawla, Kushal Lakhotia, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng

Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Maria Tsimpoukelli, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikolay Pavlovich Laptev, Ning Dong, Ning Zhang, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Rohan Maheswari, Russ Howes, Ruty Rinott, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Kohler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vítor Albiero, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaofang Wang, Xiaojian Wu, Xiaolan Wang, Xide Xia, Xilun Wu, Xinbo Gao, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yuchen Hao, Yundi Qian, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, and Zhiwei Zhao. The Llama 3 Herd of Models, 2024. URL `https://arxiv.org/abs/2407.21783`.

Edunov, Sergey, Myle Ott, Michael Auli, and David Grangier. Understanding Back-Translation at Scale. In Riloff, Ellen, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii, editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 489–500, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1045. URL `https://aclanthology.org/D18-1045`.

Esplà-Gomis, Miquel, Víctor M. Sánchez-Cartagena, Juan Antonio Pérez-Ortiz, and Felipe Sánchez-Martínez. Cross-lingual neural fuzzy matching for exploiting target-language monolingual corpora in computer-aided translation. In Goldberg, Yoav, Zornitsa Kozareva, and Yue Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7532–7543, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.511. URL `https://aclanthology.org/2022.emnlp-main.511`.

Fadaee, Marzieh, Arianna Bisazza, and Christof Monz. Data Augmentation for Low-Resource Neural Machine Translation. In Barzilay, Regina and Min-Yen Kan, editors, *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 567–573, Vancouver, Canada, July 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-2090. URL `https://aclanthology.org/P17-2090`.

Feng, Fangxiaoyu, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. Language-agnostic BERT Sentence Embedding. In Muresan, Smaranda, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.62. URL `https://aclanthology.org/2022.acl-long.62`.

Feng, Yang, Shiyue Zhang, Andi Zhang, Dong Wang, and Andrew Abel. Memory-augmented Neural Machine Translation. In Palmer, Martha, Rebecca Hwa, and Sebastian Riedel, editors, *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1390–1399, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/D17-1146. URL `https://aclanthology.org/D17-1146`.

Gu, Jiatao, Yong Wang, Kyunghyun Cho, and Victor O. K. Li. Search engine guided neural machine translation. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence*, pages 5133–5140, New Orleans, Louisiana, USA, 2018. Association for the Advancement of Artificial Intelligence. doi: 10.1609/aaai.v32i1.12013.

Gu, Jiatao, Changhan Wang, and Junbo Zhao. Levenshtein Transformer. In Wallach, H., H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL `https://proceedings.neurips.cc/paper_files/paper/2019/file/675f9820626f5bc0afb47b57890b466e-Paper.pdf`.

Gulcehre, Caglar, Orhan Firat, Kelvin Xu, Kyunghyun Cho, Loïc Barrault, Huei-Chi Lin, Fethi Bougares, Holger Schwenk, and Y. Bengio. On Using Monolingual Corpora in Neural Machine Translation, 03 2015.

He, Qiuxiang, Guoping Huang, Qu Cui, Li Li, and Lemao Liu. Fast and Accurate Neural Machine Translation with Translation Memory. In Zong, Chengqing, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *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 3170–3180, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.246. URL `https://aclanthology.org/2021.acl-long.246`.

Jean, Sébastien, Orhan Firat, Kyunghyun Cho, Roland Memisevic, and Yoshua Bengio. Montreal Neural Machine Translation Systems for WMT'15. In Bojar, Ondřej, Rajan Chatterjee, Christian Federmann, Barry Haddow, Chris Hokamp, Matthias Huck, Varvara Logacheva, and Pavel Pecina, editors, *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 134–140, Lisbon, Portugal, September 2015. Association for Computational Linguistics. doi: 10.18653/v1/W15-3014. URL `https://aclanthology.org/W15-3014`.

Jiang, Albert Q., Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7B, 2023.

Jiao, Wenxiang, Wenxuan Wang, Jen tse Huang, Xing Wang, Shuming Shi, and Zhaopeng Tu. Is ChatGPT A Good Translator? Yes With GPT-4 As The Engine, 2023.

Johnson, J., M. Douze, and H. Jegou. Billion-Scale Similarity Search with GPUs. *IEEE Transactions on Big Data*, 7(03):535–547, jul 2021. ISSN 2332-7790. doi: 10.1109/TBDATA.2019.2921572.

Khandelwal, Urvashi, Angela Fan, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. Nearest Neighbor Machine Translation, 2021.

Klein, Guillaume, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander M Rush. OpenNMT: Open-source toolkit for neural machine translation. *Computing Research Repository*, arXiv:1701.02810, 2017. doi: 10.18653/v1/P17-4012. URL https://arxiv.org/abs/1701.02810.

Kocmi, Tom, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Philipp Koehn, Benjamin Marie, Christof Monz, Makoto Morishita, Kenton Murray, Makoto Nagata, Toshiaki Nakazawa, Martin Popel, Maja Popović, and Mariya Shmatova. Findings of the 2023 Conference on Machine Translation (WMT23): LLMs Are Here but Not Quite There Yet. In Koehn, Philipp, Barry Haddow, Tom Kocmi, and Christof Monz, editors, *Proceedings of the Eighth Conference on Machine Translation*, pages 1–42, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.wmt-1.1. URL https://aclanthology.org/2023.wmt-1.1.

Kocmi, Tom, Eleftherios Avramidis, Rachel Bawden, Ondrej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Benjamin Marie, Kenton Murray, Masaaki Nagata, Martin Popel, Maja Popovic, Mariya Shmatova, Steinþór Steingrímsson, and Vilém Zouhar. Preliminary WMT24 Ranking of General MT Systems and LLMs, 2024. URL https://arxiv.org/abs/2407.19884.

Koehn, Philipp. Statistical Significance Tests for Machine Translation Evaluation. In Lin, Dekang and Dekai Wu, editors, *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 388–395, Barcelona, Spain, July 2004. Association for Computational Linguistics. URL https://aclanthology.org/W04-3250.

Kudo, Taku and John Richardson. SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium, November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-2012. URL https://aclanthology.org/D18-2012.

Moslem, Yasmin, Rejwanul Haque, John D. Kelleher, and Andy Way. Adaptive Machine Translation with Large Language Models. In Nurminen, Mary, Judith Brenner, Maarit Koponen, Sirkku Latomaa, Mikhail Mikhailov, Frederike Schierl, Tharindu Ranasinghe, Eva Vanmassenhove, Sergi Alvarez Vidal, Nora Aranberri, Mara Nunziatini, Carla Parra Escartín, Mikel Forcada, Maja Popovic, Carolina Scarton, and Helena Moniz, editors, *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 227–237, Tampere, Finland, June 2023a. European Association for Machine Translation. URL https://aclanthology.org/2023.eamt-1.22.

Moslem, Yasmin, Rejwanul Haque, and Andy Way. Fine-tuning Large Language Models for Adaptive Machine Translation, 2023b.

Mu, Yongyu, Abudurexiti Reheman, Zhiquan Cao, Yuchun Fan, Bei Li, Yinqiao Li, Tong Xiao, Chunliang Zhang, and Jingbo Zhu. Augmenting Large Language Model Translators via Translation Memories. In Rogers, Anna, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Findings of the Association for Computational Linguistics: ACL 2023*, pages 10287–10299, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.653. URL `https://aclanthology.org/2023.findings-acl.653`.

Ng, Nathan, Kyra Yee, Alexei Baevski, Myle Ott, Michael Auli, and Sergey Edunov. Facebook FAIR's WMT19 News Translation Task Submission. In Bojar, Ondřej, Rajen Chatterjee, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, André Martins, Christof Monz, Matteo Negri, Aurélie Névéol, Mariana Neves, Matt Post, Marco Turchi, and Karin Verspoor, editors, *Proceedings of the Fourth Conference on Machine Translation* (*Volume 2: Shared Task Papers, Day 1*), pages 314–319, Florence, Italy, August 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-5333. URL `https://aclanthology.org/W19-5333`.

Ouyang, Long, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In Koyejo, S., S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744. Curran Associates, Inc., 2022. URL `https://proceedings.neurips.cc/paper_files/paper/2022/file/b1efde53be364a73914f58805a001731-Paper-Conference.pdf`.

Pagliardini, Matteo, Prakhar Gupta, and Martin Jaggi. Unsupervised Learning of Sentence Embeddings Using Compositional n-Gram Features. 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 528–540, New Orleans, Louisiana, USA, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1049. URL `https://www.aclweb.org/anthology/N18-1049`.

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, page 311–318, USA, 2002. Association for Computational Linguistics. doi: 10.3115/1073083.1073135. URL `https://doi.org/10.3115/1073083.1073135`.

Peng, Keqin, Liang Ding, Qihuang Zhong, Li Shen, Xuebo Liu, Min Zhang, Yuanxin Ouyang, and Dacheng Tao. Towards Making the Most of ChatGPT for Machine Translation. In *The 2023 Conference on Empirical Methods in Natural Language Processing*, 2023. doi: 10.18653/v1/2023.findings-emnlp.373. URL `https://openreview.net/forum?id=fxdvWG4rJe`.

Pham, Minh Quang, Jitao Xu, Josep Crego, François Yvon, and Jean Senellart. Priming Neural Machine Translation. In Barrault, Loïc, Ondřej Bojar, Fethi Bougares, Rajen Chatterjee, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Alexander Fraser, Yvette Graham, Paco Guzman, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, André Martins, Makoto Morishita, Christof Monz, Masaaki Nagata, Toshiaki Nakazawa, and Matteo Negri, editors, *Proceedings of the Fifth Conference on Machine Translation*, pages

516–527, Online, November 2020. Association for Computational Linguistics. URL https://aclanthology.org/2020.wmt-1.63.

Popović, Maja. chrF: character n-gram F-score for automatic MT evaluation. In Bojar, Ondřej, Rajan Chatterjee, Christian Federmann, Barry Haddow, Chris Hokamp, Matthias Huck, Varvara Logacheva, and Pavel Pecina, editors, *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal, September 2015. Association for Computational Linguistics. doi: 10.18653/v1/W15-3049. URL https://aclanthology.org/W15-3049.

Post, Matt. A Call for Clarity in Reporting BLEU Scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Belgium, Brussels, October 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-6319. URL https://www.aclweb.org/anthology/W18-6319.

Reheman, Abudurexiti, Tao Zhou, Yingfeng Luo, Di Yang, Tong Xiao, and Jingbo Zhu. Prompting neural machine translation with translation memories. In *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence and Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence and Thirteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'23/IAAI'23/EAAI'23. AAAI Press, 2023. ISBN 978-1-57735-880-0. doi: 10.1609/aaai.v37i11.26585. URL https://doi.org/10.1609/aaai.v37i11.26585.

Rei, Ricardo, Craig Stewart, Ana C Farinha, and Alon Lavie. COMET: A Neural Framework for MT Evaluation. In Webber, Bonnie, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.213. URL https://aclanthology.org/2020.emnlp-main.213.

Scao, Teven Le, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilic, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, Jonathan Tow, Alexander M. Rush, Stella Biderman, Albert Webson, Pawan Sasanka Ammanamanchi, Thomas Wang, Benoît Sagot, Niklas Muennighoff, Albert Villanova del Moral, Olatunji Ruwase, Rachel Bawden, Stas Bekman, Angelina McMillan-Major, Iz Beltagy, Huu Nguyen, Lucile Saulnier, Samson Tan, Pedro Ortiz Suarez, Victor Sanh, Hugo Laurençon, Yacine Jernite, Julien Launay, Margaret Mitchell, Colin Raffel, Aaron Gokaslan, Adi Simhi, Aitor Soroa, Alham Fikri Aji, Amit Alfassy, Anna Rogers, Ariel Kreisberg Nitzav, Canwen Xu, Chenghao Mou, Chris Emezue, Christopher Klamm, Colin Leong, Daniel van Strien, David Ifeoluwa Adelani, and et al. BLOOM: A 176B-Parameter Open-Access Multilingual Language Model. *CoRR*, abs/2211.05100, 2022. doi: 10.48550/ARXIV.2211.05100. URL https://doi.org/10.48550/arXiv.2211.05100.

Sennrich, Rico, Barry Haddow, and Alexandra Birch. Improving Neural Machine Translation Models with Monolingual Data. In Erk, Katrin and Noah A. Smith, editors, *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany, August 2016. Association for Computational Linguistics. doi: 10.18653/v1/P16-1009. URL https://aclanthology.org/P16-1009.

Son, Jungha and Boyoung Kim. Translation Performance from the User's Perspective of Large Language Models and Neural Machine Translation Systems. *Information*, 14(10), 2023.

ISSN 2078-2489. doi: 10.3390/info14100574. URL https://www.mdpi.com/2078-2489/14/10/574.

Steinberger, Ralf, Andreas Eisele, Szymon Klocek, Spyridon Pilos, and Patrick Schlüter. DGT-TM: A freely available Translation Memory in 22 languages. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 454–459, Istanbul, Turkey, May 2012. European Language Resources Association (ELRA).

Tamura, Takuya, Xiaotian Wang, Takehito Utsuro, and Masaaki Nagata. Target Language Monolingual Translation Memory based NMT by Cross-lingual Retrieval of Similar Translations and Reranking. In *Proceedings of Machine Translation Summit XIX, Vol. 1: Research Track*, pages 313–323, 2023.

Tezcan, Arda and Bram Bulté. Evaluating the impact of integrating similar translations into neural machine translation. *INFORMATION*, 13(1):33, 2022. ISSN 2078-2489. doi: 10.3390/info13010019. URL http://doi.org/10.3390/info13010019.

Tezcan, Arda, Bram Bulté, and Bram Vanroy. Towards a Better Integration of Fuzzy Matches in Neural Machine Translation through Data Augmentation. *Informatics*, 8(1), 2021. ISSN 2227-9709. doi: 10.3390/informatics8010007. URL https://www.mdpi.com/2227-9709/8/1/7.

Vanroy, Bram, Arda Tezcan, and Lieve Macken. MATEO: MAchine Translation Evaluation Online. In Nurminen, Mary and Brenner, Judith and Koponen, Maarit and Latomaa, Sirkku and Mikhailov, Mikhail and Schierl, Frederike and Ranasinghe, Tharindu and Vanmassenhove, Eva and Alvarez Vidal, Sergi and Aranberri, Nora and Nunziatini, Mara and Parra Escartín, Carla and Forcada, Mikel and Popovic, Maja and Scarton, Carolina and Moniz, Helena, editor, *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 499–500. European Association for Machine Translation (EAMT), 2023. URL https://lt3.ugent.be/mateo/.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

Wei, Jason, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. Emergent Abilities of Large Language Models. *Transactions on Machine Learning Research*, 2022. ISSN 2835-8856. URL https://openreview.net/forum?id=yzkSU5zdwD. Survey Certification.

Xu, Jitao, Josep Crego, and Jean Senellart. Boosting Neural Machine Translation with Similar Translations. In Jurafsky, Dan, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1580–1590, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.144. URL https://aclanthology.org/2020.acl-main.144.

Xu, Jitao, Minh Quang Pham, Sadaf Abdul Rauf, and François Yvon. LISN @ WMT 2021. In Barrault, Loic, Ondrej Bojar, Fethi Bougares, Rajen Chatterjee, Marta R. Costa-jussa, Christian Federmann, Mark Fishel, Alexander Fraser, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Paco Guzman, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp

Koehn, Tom Kocmi, Andre Martins, Makoto Morishita, and Christof Monz, editors, *Proceedings of the Sixth Conference on Machine Translation*, pages 232–242, Online, November 2021. Association for Computational Linguistics. URL https://aclanthology.org/2021.wmt-1.22.

Xu, Jitao, Josep Crego, and François Yvon. Integrating Translation Memories into Non-Autoregressive Machine Translation. In Vlachos, Andreas and Isabelle Augenstein, editors, *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1326–1338, Dubrovnik, Croatia, May 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.eacl-main.96. URL https://aclanthology.org/2023.eacl-main.96.

Xu, Nuo, Yinqiao Li, Chen Xu, Yanyang Li, Bei Li, Tong Xiao, and Jingbo Zhu. Analysis of Back-Translation Methods for Low-Resource Neural Machine Translation. In *Natural Language Processing and Chinese Computing: 8th CCF International Conference, NLPCC 2019, Dunhuang, China, October 9–14, 2019, Proceedings, Part II*, page 466–475, Berlin, Heidelberg, 2019. Springer-Verlag. ISBN 978-3-030-32235-9. doi: 10.1007/978-3-030-32236-6_42. URL https://doi.org/10.1007/978-3-030-32236-6_42.

Zhang, Jingyi, Masao Utiyama, Eiichro Sumita, Graham Neubig, and Satoshi Nakamura. Guiding Neural Machine Translation with Retrieved Translation Pieces. 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 1325–1335, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1120. URL https://www.aclweb.org/anthology/N18-1120.

A. Appendix

A.1. Sent2vec hyper-parameters

To train sent2vec models, we used the same hyper-parameters that are suggested in the description paper (Pagliardini et al., 2018) for a sent2vec model trained on Wikipedia data containing both unigrams and bigrams. The hyper-parameters values are provided in Table 8.

Hyper-Parameter	Value
embedding dimension	700
minimum word count	8
minimum target word count	20
initial learning rate	0.2
epochs	9
sub-sampling hyper-parameter	5×10^{-6}
bigrams dropped per sentence	4
number of negatives sampled	10

Table 8. Hyper-parameters for training sent2vec models.

A.2. FAISS hyper-parameters

We created a Flat index with an inner product metric for brute-force search. We used cosine similarity as the match metric effectively by adding the L2-normalised vectors of the sentence representation to the index and using an L2-normalised sentence vector as an input query. For more information on FAISS, please see `https://github.com/facebookresearch/faiss/wiki`.

A.3. NMT training data and vocabulary sizes

System	EN→UK	EN↔FR
BASE	286417	300000
NFR	572731	600000
NFR$_{mono}$	572731	600000
BT	1747737	1799436
BT+NFR	3491066	3598872
BASE_HQ	1747737	1799436
NFR_HQ	3491066	3598872

Table 9. The total number of bilingual sentence pairs used for training the NMT systems using all available data, per language direction.

For training the NFR systems by using the minimum FM similarity threshold of $\lambda = 0.5$, we were able to retrieve FMs for all source sentences in our experiments. As a result, combining the augmented and the non-augmented sentence pairs in the NFR approach simply doubled the training data sizes for all systems and all language directions (e.g. *BASE* vs. *NFR*, and *BT* vs. *BT+NFR*). It is also worth highlighting that *NFR* and NFR_{mono} use the same NMT training data. The key difference is that NFR_{mono} additionally employs FM retrieval and augmentation (only) on the test set using the additional available monolingual data.

System	**EN→UK**	**EN→FR**	**EN→FR**
BASE	21906/17612	36252/35550	35550/36252
NFR	33075/17612	39924/35550	39731/36252
NFR_{mono}	33075/17612	39924/35550	39731/36252
BT	21906/30683	36253/50611	35551/51089
BT+NFR	38466/30683	47726/50611	47501/51089
BASE_HQ	–	51089/50611	50611/51089
NFR_HQ	–	53874/50611	53592/51089

Table 10. Vocabulary sizes (source/target) of the NMT systems using all available data, per language direction.

A.4. NMT hyper-parameters

Hyper-Parameter	**Value**
source/target embedding dimension	512
size of hidden layers	512
feed-forward layers	2048
number of heads	8
number of layers	6
batch size	32
gradient accumulation	4
dropout	0.1
warm-up steps	8000
optimizer	Adam
validation steps	2000

Table 11. Common hyper-parameter values used for training the NMT systems.

A.5. LLM implementation details

All LLMs in the experiments were prompted using consistent templates and configurations to ensure fair comparison. Each model's native chat template format was utilised while maintaining identical prompt content across all models. The default sampling parameters were used for inference, as their respective developers recommended. Table 12 shows the prompt template used across all models. While the actual formatting varied according to each model's chat template, the content structure remained consistent:

```
Translate the following text from {source_language} into {target_language}.
{source_language}: {source_sentence}
{target_language}:
```

Table 12. Prompt template used across all models. The actual formatting followed each model's specific chat template while maintaining this content structure.

In some cases, the models exhibited consistent patterns of overgeneration, such as adding parenthetical notes (e.g., "\n Note that...") after the translation itself. These extra generations followed predictable patterns and were systematically filtered out before evaluation. The final hypotheses used for comparison against the references contained only the models' core translations.

A.6. Back-translation performance

System	BLEU	chrF	COMET
UK→EN 100%	59.54	76.66	86.27
UK→EN 66%	55.67	74.12	85.04
UK→EN 33%	50.69	70.60	82.89
FR→EN 100%	53.95	71.61	85.01
FR→EN 66%	51.30	69.52	83.49
FR→EN 33%	46.90	66.05	80.56
EN→FR 100%	51.50	71.21	84.22
EN→FR 66%	48.14	68.77	82.48
EN→FR 33%	43.71	65.26	77.88

Table 13. Results of the automatic evaluations performed on back-translation systems using different sizes of bilingual data, in reverse language direction, on the reversed test sets.

A.7. System performance using reduced datasets

System	**EN→UK**			**EN→FR**			**FR→EN**		
	BLEU	chrF	COMET	BLEU	chrF	COMET	BLEU	chrF	COMET
BT 100%	56.21	76.93	92.18	54.88	73.34	85.27	56.99	74.23	86.72
BT 80%	56.22	76.95	91.98	54.91	73.56	85.74	57.60	74.77	86.83
BT 60%	56.68	77.24	92.04	54.58	73.28	85.63	56.95	74.47	86.78
BT 40%	57.33	77.53	92.14	53.70	72.79	85.41	56.48	73.94	86.35
BT 20%	56.49	76.96	91.79	53.48	72.88	85.20	55.36	72.92	86.07
NFR_{mono} 100%	60.39	78.89	92.03	52.39	71.68	84.45	55.54	72.62	85.44
NFR_{mono} 80%	60.03	78.72	92.02	52.32	71.58	84.37	55.19	72.43	85.39
NFR_{mono} 60%	59.44	78.39	91.97	52.11	71.47	84.37	55.11	72.37	85.36
NFR_{mono} 40%	58.87	78.08	91.92	52.07	71.51	84.29	54.92	72.27	85.40
NFR_{mono} 20%	58.42	77.85	91.90	51.83	71.34	84.29	54.72	72.16	85.38
BT+NFR 100%	66.95	82.39	92.78	61.91	77.54	87.13	64.69	78.65	87.79
BT+NFR 80%	65.40	81.53	92.66	60.44	76.76	86.85	63.62	78.02	87.32
BT+NFR 60%	65.11	81.27	92.69	59.38	76.04	86.39	62.51	77.14	87.06
BT+NFR 40%	62.68	80.14	92.57	58.24	75.39	86.40	61.25	76.26	86.87
BT+NFR 20%	59.42	78.42	92.22	56.04	74.04	85.77	59.15	74.92	86.55

Table 14. Results of the automatic evaluations performed on systems using the original bilingual data and gradually decreasing sizes of additional monolingual data in the target language.

System	**EN→UK**			**EN→FR**			**FR→EN**		
	BLEU	chrF	COMET	BLEU	chrF	COMET	BLEU	chrF	COMET
BASE 100%	54.85	75.95	91.23	51.5	71.21	84.22	53.95	71.61	85.01
BASE 66%	53.18	74.62	90.18	48.14	68.77	82.48	51.30	69.52	83.49
BASE 33%	46.76	70.23	87.47	43.71	65.26	77.88	46.90	66.05	80.56
NFR 100%	57.73	77.52	91.78	52.67	71.82	84.50	54.43	71.94	85.26
NFR 66%	53.56	74.86	90.43	48.60	69.03	82.26	50.61	69.05	83.10
NFR 33%	46.31	70.05	87.07	46.90	67.70	81.01	45.96	65.58	80.23
BT 100%	56.21	76.93	92.18	54.88	73.34	85.27	56.99	74.23	86.72
BT 66%	54.56	75.83	91.43	53.75	72.72	85.19	56.25	73.65	86.21
BT 33%	51.52	73.79	90.32	51.80	71.11	83.92	53.88	72.28	85.06
NFR_{mono} *100%*	60.39	78.89	92.03	52.39	71.68	84.45	55.54	72.62	85.44
NFR_{mono} *66%*	53.96	75.16	90.54	49.77	69.77	82.93	50.62	69.09	83.16
NFR_{mono} *33%*	46.15	70.03	87.28	43.72	65.27	78.68	46.14	65.65	80.17
BT+NFR 100%	66.95	82.39	92.78	61.91	77.54	87.13	64.69	78.65	87.79
BT+NFR 66%	63.89	80.29	92.01	59.75	76.20	86.40	61.99	76.38	86.49
BT+NFR 33%	61.83	78.96	91.30	58.01	74.75	85.23	60.62	76.11	85.92

Table 15. Results of the automatic evaluations performed on systems using all additional monolingual data in the target language and gradually decreasing sizes of the original bilingual data.

Address for correspondence:
Arda Tezcan
arda.tezcan@ugent.be
LT3, Ghent University
Groot-Brittanniëlaan 45, 9000 Ghent, Belgium

PBML

The Prague Bulletin of Mathematical Linguistics

NUMBER 122 DECEMBER 2024 43–70

Towards Automated Spoken Language Assessment: A Study of ASR Transcription of Examinations for Non-Native Speakers of Czech

Michal Novák, Peter Polák, Kateřina Rysová, Magdaléna Rysová, Ondřej Bojar

Charles University
Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics
Prague, Czechia
[mnovak|polak|rysova|magdalena.rysova|bojar]@ufal.mff.cuni.cz

Abstract

The article investigates the effectiveness of Automatic Speech Recognition (ASR) systems for transcribing Czech language proficiency exams, targeting non-native speakers. It explores the potential of ASR technology as the first step in developing an automated assessment tool aligned with the Common European Framework of Reference for Languages (CEFR). We analyze transcriptions from various ASR systems, refined by human annotators, to evaluate the effectiveness of this approach and the extent of manual correction required for accuracy. Focusing on A2 level exam recordings, we compare different transcription methodologies, including human-only transcription, to understand the influence of the human element in the process. The paper also presents a quantitative analysis that addresses the efficiency of manual post-editing versus direct transcription and the impact of post-editing on transcript consistency and potential biases. A case study demonstrates the challenges of transcribing non-native spoken language in a setting where recording errors is essential, discussing both advantages and limits of human transcription and the variability among transcribers, especially in low audio quality scenarios.

Corresponding author: mnovak@ufal.mff.cuni.cz

Cite as: Michal Novák, Peter Polák, Kateřina Rysová, Magdaléna Rysová, Ondřej Bojar. Towards Automated Spoken Language Assessment: A Study of ASR Transcription of Examinations for Non-Native Speakers of Czech. The Prague Bulletin of Mathematical Linguistics No. 122, 2024, pp. 43–70. doi: 10.14712/00326585.031.

1. Introduction

The advent of neural networks in recent years has significantly impacted various tasks in natural language processing, including Automatic Speech Recognition (ASR). ASR systems such as Whisper (Radford et al., 2023) from OpenAI deliver high-quality transcriptions across many languages, showing remarkable robustness to nuances and variations in speech. This robustness is advantageous for applications like natural language understanding, where the goal is to grasp the speaker's intended meaning. However, this robustness can be problematic when evaluating language competency, as it can mask errors in pronunciation, grammar, or vocabulary that are crucial for the evaluation. Consequently, while advances in ASR technology have enhanced many aspects of natural language processing, they also pose challenges when evaluating linguistic proficiency.

Our article investigates the effectiveness of ASR systems in transcribing spoken parts of Czech language proficiency exams for non-native speakers. It is an initial step towards creating an automated tool capable of assessing spoken language proficiency according to the Common European Framework of Reference for Languages (Ivanová, 2006) standards. The intended tool aims to support human evaluators in determining whether candidates meet the certification requirements for Czech language exams. In addition, it could benefit learners by allowing them to regularly assess their performance.

The objectives of our work are two-fold. Firstly, we aim to explore typical examples of how current ASR systems may mask errors in spoken language produced by non-native speakers of Czech and contrast these with the challenges that human annotators face. By identifying specific instances where ASR systems fail to capture errors that human annotators would notice, we can better understand the gap between automated and human evaluations.

Secondly, we are focused on the more concrete and short-term goal of creating a dataset of exam recordings and their transcripts. This dataset will serve as a valuable resource for further research and development in the field of ASR and language assessment. To achieve this, we are investigating the most practical transcribing approach, considering factors such as time efficiency, consistency, and avoiding bias towards existing ASR systems. Our aim is to develop a transcription methodology that is both efficient and reliable, ensuring that the transcripts produced are of high quality and useful for evaluating spoken language proficiency.

Specifically, we examine two primary methods for manual transcription and annotation of recordings: (1) *direct transcription* (without ASR assistance) and (2) *human post-editing of ASR-generated transcripts*. We hypothesize that the former is time-consuming, and we could thus afford it only for a portion of our data. On the other hand, the latter, although faster, may introduce biases reflecting the specific ASR system used and, most importantly, obscure differences in speech output quality of the examined candidates. To mitigate potential biases specific for individual ASR systems, we combine outputs from multiple ASR systems during automatic transcription, which annotators subsequently post-edit (typically reintroduce errors, in fact), to make them closer to the original recordings.

In this study, we focus only on recordings of speakers examined at A2 level of the CEFR standards. This level represents an upper basic proficiency in the language, where learners are expected to handle simple and routine tasks and is required for non-native speakers who wish to obtain permanent residency in the Czech Republic.

The article is structured as follows. After presenting the related work in Section 2, we introduce the methods that we used to acquire, transcribe, and annotate recordings in Section 3. Later in Section 4, we summarize the data that we collected and annotated using these methods for the purpose of further analysis. Using examples from the collected data, we show the most interesting types of differences between the transcripts obtained by different methods in Section 5. In Section 3, we also propose the algorithm to automatically align the transcripts of the same recording. The algorithm is key for the analysis in Section 6, in which we quantitatively assess the usability of the proposed methods to transcribe and annotate the recordings. Finally, in Section 7 we summarize our results and discuss potential directions for future research.

2. Related Work

For Czech, an automatic transcription system was recently developed specifically for non-native speakers (Holaj, 2023) which produces a representation of phonemes. Transcribing the speech of foreign speakers is a challenging task because their utterances contain errors that native speakers do not make.

In designing this system, three different annotation schemes were developed (two using attribute-based annotation and one using synthetic annotation). Using machine learning methods, a total of five speech recognition models were developed. The best-performing model achieved a 77% phoneme-level transcription accuracy on test data (recordings of isolated words or short phrases of non-native speakers). Currently, there is no other tool capable of automatically transcribing non-native Czech speech, including sound segments not present in standard Czech.

The model is trained using the Persephone speech recognition library (Adams et al., 2019). Persephone was developed as a speech recognition tool for transcribing recordings in languages with limited data, which is advantageous given the uncertainty regarding the required sample size for successfully training a model capable of recognizing non-native Czech speech.

Holaj (2023) analyzes over 100 hours of audio recordings, although about one-third of this total duration consists of researcher instructions, pauses, or background noise. These recordings were collected from 254 respondents with proficiency levels ranging from A0 to C1, the majority of whom were at levels A0 to A2. The data contain isolated words or phrases recorded. The data collection process involved a native

Czech speaker reading a line of data in standard Czech (typically a sound and three example words or phrases) from printed materials. The respondents then repeated the line (using a paper with the printed dataset). In the second step, respondents slowly read individual sounds, words, or phrases independently.

As part of the project, the annotation tool ANOPHONE (Holaj and Pořízka, 2023) was used for manual data processing. The tool serves as an online database of recordings from non-native speakers, along with an overview of their annotations for available annotation tasks. In addition, the tool facilitates the annotation of these recordings within custom annotation tasks.

In addition to the mentioned automatic system focused directly on transcribing spoken Czech of non-native speakers, there are other applications designed for converting Czech speech into written text. These applications include tools like ČESKY.AI[1] or UWebASR (Švec et al., 2018). While the former is a commercial solution, the latter is freely available for research purposes. The recorded audio is automatically transcribed and stored in a structured XML format, allowing for further manual post-processing. Another work, Lehečka et al. (2023), focuses on the transcription of oral history archives in Czech. The model is available online for use with UWebASR.

Finally, there are highly multilingual ASR models available. Whisper (Radford et al., 2023) is a family of encoder-decoder-based models capable of transcribing audio in 96 languages. Additionally, the model generates transcripts as unnormalized (natural) text, i.e., with casing and punctuation, without any need for an inverse text normalization tool. MMS (Pratap et al., 2024) covers more than 1000 languages. MMS models are based on wav2vec 2.0 (Baevski et al., 2020) encoder model with CTC (Graves et al., 2006) decoding. According to the authors, these models are expected to surpass the performance of the Whisper Large model on the FLEURS (Conneau et al., 2023) dataset. Another notable system is Phonexia,[2] which specializes in voice biometrics and speech recognition technologies. Phonexia's solutions can identify a speaker's voice after just a few seconds of natural speech, detect gender, estimate age, and identify languages and keywords in conversations. Their latest Speech Platform includes a new generation of Language Identification technology, capable of recognizing 140 languages. Phonexia's systems are designed to convert spoken words into text and offer commercial solutions.

Regarding datasets, there are relatively few spoken corpora containing Czech of non-native speakers. An example is the corpus by Kubanek-German (2000), which includes recordings of children (16 boys and 16 girls) aged 10 years, whose first language is German and who are learning Czech as a foreign language. The recordings capture 25-minute interviews consisting of three parts. In the first part, the conversation covered topics familiar to the children, the second part included questions based

[1] `https://cesky.ai/`

[2] `https://www.phonexia.com/`

on an unfamiliar picture book on the theme of water, and in the third part, the children collaborated in a group on an assigned task. Conversations with the children were led by an adult investigator. The corpus also includes transcripts of the recordings and its data is available online.[3].

Another spoken corpus that includes Czech of non-native speakers is the corpus by Schmiedtova (2000–2001), which contains both recordings and their transcripts. The speakers captured in these recordings are adults.

There is also a corpus that captures the Czech language of migrants (Bučková, 2023). This corpus includes informal spoken Czech and German from Czech-German bilingual speakers born in Czechoslovakia around 1955, who moved to Germany after the age of 12.

The primary advantage of these corpora for our long-term goals lies in their potential to support the development of ASR systems capable of identifying pronunciation errors.

3. Methodology

The paper focuses on the analysis of spoken data, specifically recordings of non-native speakers taking the Czech language proficiency exam at the A2 level according to the Common European Framework of Reference for Languages. The basis of our research lies in the audio recordings of exams, which were transcribed into written text and enriched with additional annotation. This has been done either entirely manually from scratch, or semi-automatically by manual post-editing of the ASR outputs. In order to compute metrics based on edit distance of transcripts, we introduce a method for aligning them at the utterance level.

3.1. Audio Data Acquisition

Audio data was provided by the Institute for Language and Preparatory Studies of Charles University (ÚJOP).[4] These recordings represent the oral part of the Czech Language Certificate Exam (CCE; Pečený, 2012, 2013), administered by the ÚJOP. A portion of the data was also supplied by the National Pedagogical Institute of the Czech Republic,[5] which oversees the Exam in the Czech Language for Permanent Residence (Cvejnová and Geppert, 2022).

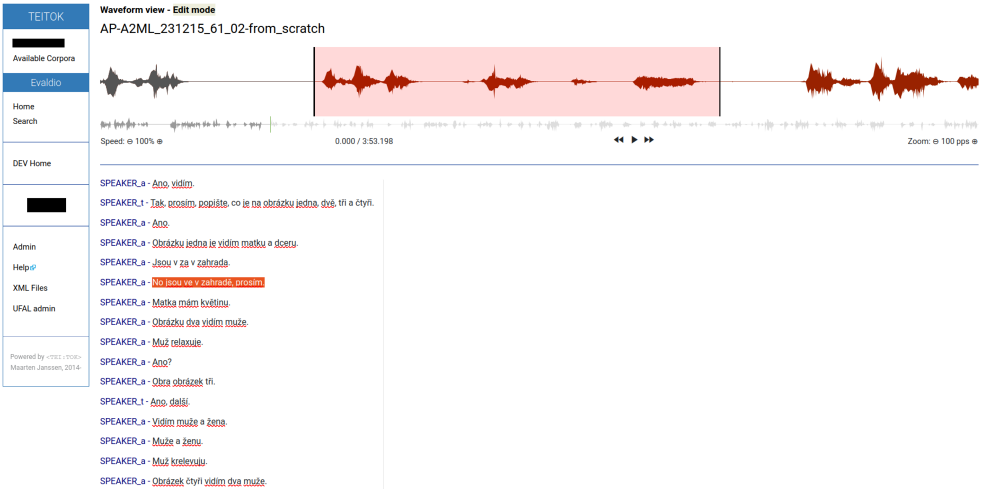
3.2. Manual Annotation

To ensure transcription quality and accuracy, a manual annotation process was implemented. We have six trained students and graduates of Czech philology to work

[3] http://talkbank.org/DB/

[4] http://ujop.cuni.cz

[5] http://npi.cz



AP-A2ML_231215_61_02-from_scratch

SPEAKER_a - Ano, vidím.
SPEAKER_t - Tak, prosím, popište, co je na obrázku jedna, dvě, tři a čtyři.
SPEAKER_a - Ano.
SPEAKER_a - Obrázku jedna je vidím matku a dceru.
SPEAKER_a - Jsou v za v zahrada.
SPEAKER_a - No jsou ve v zahradě, prosím.
SPEAKER_a - Matka mám květinu.
SPEAKER_a - Obrázku dva vidím muže.
SPEAKER_a - Muž relaxuje.
SPEAKER_a - Ano?
SPEAKER_a - Obra obrázek tři.
SPEAKER_t - Ano, další.
SPEAKER_a - Vidím muže a žena.
SPEAKER_a - Muže a ženu.
SPEAKER_a - Muž krelevuju.
SPEAKER_a - Obrázek čtyři vidím dva muže.

Figure 1. A screenshot of the TEITOK environment while editing a transcript.

with the annotation tool described below and provide manual annotation according to the guidelines. The annotation process included (1) transcription, (2) time alignment of utterances, and (3) speaker identification. Subsequently, all the transcripts have been reviewed by a single annotator in order to fix errors and achieve better agreement of the transcripts produced by different annotators.

Annotation Tool The annotations have been collected using the TEITOK platform (Janssen, 2021).[6] TEITOK is a web-based environment for viewing, creating and editing datasets of various types, including multimodal data combining text and audio.

Figure 1 shows the annotation screen for one of the transcripts. The top part displays the recording's waveform and the controls to play the recording while the bottom part contains its transcript split into a sequence of utterances. TEITOK allows the annotator to align a transcribed utterance with a particular segment in the recording. This can be achieved either by selecting a region in the waveform or specifying the exact start and end times of the utterance. In addition, the annotation tool allows for labeling the utterances with identifiers of speakers. In the background, the tool also logs the metadata about each editing session: the name of the edited file and the annotator, and the timestamps when the annotator starts and ends the editing session.

[6] `https://gitlab.com/maartenes/TEITOK/`

This information can be used to calculate the duration the annotators spent on the transcripts.[7]

It is possible both to annotate transcripts in TEITOK from scratch as well as to load already annotated transcripts for further post-editing.

Transcription In our pursuit of developing a tool for automated evaluation of language proficiency, it is crucial to obtain speech transcripts that adequately capture what individual speakers produced during Czech language exams. Our goal is to avoid artificially enhancing the transcriptions, which some systems designed for native speakers might do to increase readability. Instead, we aim to preserve language deficiencies. Thus, transcriptions should include all speech errors, for example, word repetitions, incomplete words, incorrect word forms, or filler expressions. At the same time, we aim to keep the transcriptions simple both for annotators and users and do not distinguish acoustic and articulatory details. We transcribe recordings into written Czech according to the guidelines that the annotators should follow. A sample transcript of the recording is included in Appendix.

As shown in Figure 1, the annotators technically split transcribed text into utterances during the annotation process. Each utterance must be produced by a single speaker and the utterances may overlap if necessary, i.e. when more speakers speak at the same time. Because we did not specify any further constraints for the utterances, the number and size of utterances may vary across the annotators.

Time Alignment The annotators were asked to align each utterance with a particular segment in the recording by specifying the start and end time of the utterance.

Speaker Identification Each utterance was assigned an anonymized speaker identifier. From the identifier, it is possible to distinguish the examiner from the candidate in a given session. However, the identifiers are not unique across transcripts, i.e. the same identifier can be assigned to two different speakers in different transcripts or the same speaker can be labeled by different identifiers in distinct transcripts.

Manual Review All the transcripts have been reviewed by a single annotator who produced none of the original transcripts. The reviewer is a researcher in linguistics and one of the co-authors of this study.

The reason for additional review was to fix potential errors and enhance the adherence of the transcripts produced by different annotators to the annotation guidelines. The subsequent analysis in Section 6 was mainly carried out on the reviewed tran-

[7] Unfortunately, this feature was implemented to the TEITOK only after some of the transcripts had been already annotated. We thus do not have this information for every transcript.

scripts. However, it was more convenient to use the original non-reviewed transcripts for some experiments (e.g. in Section 6.2).

3.3. Automatic Annotation

As this study compares the manual way of acquiring transcripts with the semi-automatic or fully automatic way, we process the recordings with automatic tools. Specifically, we use the WhisperX system to provide us with an initial version of all the necessary information followed by three ASR systems on the segments of the recording specified by time alignment to observe which of the ASRs is better fit for our purposes.

Annotation by WhisperX WhisperX[8] (Bain et al., 2023) is a toolkit that combines fast ASR with voice activity detection, word-level timestamps, and speaker diarization.[9] For ASR, it uses the *faster-whisper* tool,[10] a time and memory efficient reimplementation of OpenAI's Whisper model. In particular, we use the *Large V2* model as the default for transcription by WhisperX. Speaker diarization is provided by the *pyannote.audio* tool in version 2.1.1 (Bredin and Laurent, 2021).[11]

The WhisperX toolkit can therefore automatically provide all information that is annotated manually: (1) transcripts segmented into utterances, (2) time alignment of utterances, and (3) speaker identifiers. Although it is also possible to deliver time alignment on the word level, we do not need such fine granularity for this study.

Speaker identifiers produced by WhisperX do not capture the role of the speaker. Therefore, we apply a post-processing heuristics that attempts to recognize which speaker is an examiner. The heuristics is based on the proportion of words typically said by the examiners (e.g. "úloha" 'exercise', "otázka" 'question'). The other speakers are then assigned the candidate role with different co-indexing. The speaker identifiers are renamed accordingly.

ASR Systems Having a transcript segmented into utterances by the *pyannote.audio* tool in the WhisperX toolkit, we run additional ASR systems on the transcribed utterances. We use the single segmentation by WhisperX across other ASR systems because we want to be able to easily combine the different ASR outputs.

[8] `https://github.com/m-bain/whisperX`

[9] This is the task of recognizing the segments of recordings when the same speaker is speaking. Unlike in Section 3.2, the tool does not attempt to identify who is the examiner and who is the candidate, it only distinguishes the speakers.

[10] `https://github.com/SYSTRAN/faster-whisper`

[11] `https://huggingface.co/pyannote/speaker-diarization`

Out of the ASR systems capable of transcribing Czech speech (see Section 2), we selected the following models for our experiments. Besides the Whisper model *Large V2*, which is the default of the WhisperX toolkit, we also tried the Whisper model *Large V3*, and MMS models *mms-1b-all* and *mms-1b-fl102*.

We do not use the tool developed by Holaj (2023) because it is trained on one-word utterances and provides a phonetic transcript only. We also do not use UWebASR because the tool is available only via an online interface and thus not very suitable for our batch processing.

3.4. Manual Post-edits of Automatic Annotation

In the semi-automatic way of annotation, we first run the automatic annotation tools (see Section 3.3) and then ask the human annotators to post-edit the produced outputs. We used two variants of transcripts as the input for manual post-editing: (1) *WhisperX*, and (2) *mixed* transcripts.

Whereas *WhisperX* transcripts are the transcripts exactly as produced by the WhisperX toolkit, in the *mixed* transcripts, each utterance is randomly selected from the four available ASR outputs by the four ASR systems mentioned above. We hypothesize that evaluating against references created by post-editing a single ASR system will be biased towards the particular system. The *mixed* setup (supposedly unbiased) allows us to study the contrast with the supposedly biased approach (*WhisperX*). We suggest the bias can be alleviated by mixing multiple ASR outputs to form the basis for manual post-editing.

The automatic transcripts were then loaded to the TEITOK environment for the manual post-editing. To limit any potential annotators' bias, the annotators were not given the information on the source of the automatic transcripts.

3.5. Automatic Alignment of Transcripts

In our study, we compare different versions of transcripts. All automatic transcripts, regardless of the used ASR model, follow the same segmentation and speaker diarization by WhisperX (see Section 3.3). Therefore, their comparison is straightforward. However, the transcripts written from scratch or post-edited by human annotators exhibit different segmentation and even diarization, i.e. the guess of the speaker. Therefore, we use a specific evaluation protocol to compare the automatic transcripts and those annotated by humans. Within this protocol, we work at the utterance level, comparing only the matching utterances and aggregating the statistics at the document level.

Our evaluation protocol follows several steps:

1. *We sort all utterances by their start timestamp.* This guarantees that overlapping speech from two speakers appears in the same sequence in both documents.

2. In the case of manually annotated transcripts, *we concatenate neighboring utterances of the same speaker into one continuous utterance*. This step is omitted for automatic transcripts due to inaccuracies in automatic speaker diarization.
3. *We compute a 1-to-1 alignment between utterances.* The alignment aims at pairing the utterances that are the most likely matches based on their content and timing. The heuristic used for this is discussed in the following section.
4. *We extend the 1-to-1 alignment to many-to-1 using a heuristic*. Many-to-1 alignment is crucial since automatic transcripts are divided into short segments based on predicted sentence boundaries, while manual annotations segment utterances only according to the change in speaker (refer to Step 2), i.e. much less frequently.
5. If several utterances are matched to a single one, *we combine them together*.
6. *We compute relevant statistics* as needed (e.g., character error rate).

1-to-1 Alignment: We compute the 1-to-1 alignment using a global alignment function in `Bio.pairwise2`[12] package. We construct a specialized scoring function for the two segments S_a and S_m, where S_a stands for an automatic segment and S_m for a manual one:

$$\text{score}(S_a, S_m) = \begin{cases} -\infty & \text{overlap}(S_a, S_m) = 0 \\ \frac{\text{overlap}(S_a,S_m)}{\text{duration}(S_a)} \cdot \frac{\text{edit\_distance}(S_a,S_m)}{\text{length}(S_a)} & \text{else.} \end{cases} \tag{1}$$

Note that the order of the arguments is important as the score is not symmetric. The alignment algorithm then maximizes Equation (1). Segments that do not overlap in time are assigned a score of $-\infty$, which means that they should never be aligned. In cases where the segments overlap, the score is calculated as the product of the relative overlap and the relative edit distance, given that several S_a are anticipated to be subsegments of S_m. Relative overlap helps to match the segment S_a to a segment S_m that encompasses the segment S_a to the greatest extent. The second term, the relative edit distance, helps to match the segment S_a with the segment S_m that most closely resembles its content. This is particularly important when two speakers are talking simultaneously, causing two utterances in a single transcript to overlap in time, and thus we have to rely on transcribed words rather than on the timespan similarity.

Many-to-1 Alignment: As previously discussed, the automatic transcripts are segmented into short utterances, whereas the manual or post-edited versions are segmented based on speaker boundaries, resulting in longer segments. This implies that several utterances in the automatic transcript typically correspond to a single utterance in the human-edited transcript. Therefore, we extend the 1-to-1 alignment to a many-to-1 alignment. To achieve this, we gather each unaligned S_a and aggregate all segments S_m that either overlap with or are "near" the segment S_a. We define

[12] `https://biopython.org/docs/1.76/api/Bio.pairwise2.html`

two segments as being near if their beginnings or endings are within 0.5 seconds of each other, accommodating annotation variations in segment time alignment. The segment S_a is then matched with the segment S_m that achieves the highest score, as specified in Equation (1). There might not be any corresponding segment S_m, which can occur when the ASR model generates text from background noise. Additionally, it is possible that no segment S_a is linked to a segment S_m, which can happen if the ASR skips some utterances.

4. Data Description

Using the methods described in Section 3, we collected a set of recordings and their transcripts, which we analyze further in the following sections. Here we present basic properties and statistics of the collected data. The subset of the data authorized for publication by the recordings' providers is publicly available for download,[13] browsing, and querying.[14]

4.1. Recordings

In this work, we examine recordings of Czech language proficiency exams for non-native speakers or their pretests. Specifically, these exams are at the A2 level according to the CEFR (Vodičková et al., 2012).

The exam consists of several parts: reading comprehension, listening comprehension, writing, and speaking, each evaluated separately. For our study, we focus on data from the speaking part of the Czech language exam for foreigners.

The recorded interactions have the form of a dialogue, capturing conversations between the examiner (a native Czech speaker) and the exam candidate (a non-native speaker learning Czech as a foreign language). The conversation follows predefined tasks. For instance, the exam candidate responds to examiner questions such as "Where are you from?" or "What do you do in your free time?"

Furthermore, the candidate must engage in various communication situations. During these interactions, both examiner and candidate assume different roles. For example, they might play the roles of two friends deciding what gift to buy for a mutual friend or discussing travel plans. The goal is information exchange between the examiner and the exam candidate. Thus, during this conversation, the candidate must answer the examiner's questions and also ask questions related to the given topic.

Communication situations can be initiated using cards (the candidate holds one card related to a specific topic, while the examiner holds another card on the same topic), which guide the information exchange during the conversation.

[13] http://hdl.handle.net/11234/1-5731

[14] https://lindat.mff.cuni.cz/services/teitok-live/evaldio/index.php

Proficiency CEFR Level A2 in Czech as a Foreign Language The A2 proficiency level, also referred to as upper Basic User, is required for non-native speakers who wish to obtain permanent residency in the Czech Republic. At this level, candidates can effectively communicate in everyday situations. They possess basic vocabulary and can use it appropriately. A2 speakers engage in simple interactions, exchanging information about common topics. Because they are not yet fully independent language users, their success in communication also depends on the engagement of their conversation partner. The following speaking abilities characterize speakers at the A2 level:

Answering Simple Questions: A2 candidates can respond to straightforward questions.

Describing Everyday Situations: They can describe everyday situations, e.g., related to family, school, work, or leisure time. Additionally, they can discuss their plans, habits, personal experiences, and activities they will undertake (e.g., for the evening or weekend).

Expressing Preferences: A2 speakers can express their likes and dislikes, as well as what does or does not appeal to them.

Seeking Information: They can ask for information in various contexts, such as in stores, post offices, banks, restaurants, or hotels. They inquire about prices, quantities, and travel-related details.

Clear Pronunciation: Despite their foreign accent, A2 speakers articulate clearly and understandably.

Basic Sociolinguistic Competence: In basic communication situations, they appropriately use fundamental language means.

4.2. Transcripts

For all collected recordings, we have provided one or multiple transcripts. All transcripts are assigned a speaker role and segmented into utterances, which are aligned with the recording using timestamps.

First, for each recording we have automatically created four transcripts using the four ASR systems listed in Section 3.3. WhisperX was employed to assign the speaker roles and align the utterances with the recording. Each ASR system was then used to transcribe all utterances.

Second, we distinguish three types of human-annotated transcripts depending on the annotation method:

- *From scratch*: manual annotation with no pre-annotation;
- *WhisperX*: manual post-editing of the automatic transcripts produced by WhisperX;
- *Mixed*: manual post-editing of the automatic transcripts combined from the outputs of all four ASR systems.

We keep three versions of each transcript corresponding to the specific stages during the annotation process:

1. *Initial*: the transcripts are either completely empty (in the case of the *From scratch* method) or produced by the selected automatic method with no manual intervention;
2. *Before review*: the result of the first stage of manual annotation;
3. *Final*: the reviewed transcripts.

4.3. Statistics on the Recordings and their Transcripts

We analyse 65 recordings with the total duration over 397 minutes.

Most of the recordings feature two speakers: one examiner and one candidate.[15] The examiner is a native speaker of Czech and the candidate is a non-native speaker being examined for the A2 level. While the candidate is different in each recording, the examiners naturally often reappear. The total number of candidates is thus around 65 and the total number of examiners is in the range of 2–5.[16] In total, the examiners and candidates speak for 198 and 184 minutes, respectively.

For each recording, we collected at least one manual or manually post-edited transcript. In order to allow for various experiments, some of the recordings have been transcribed multiple times. However, no annotator has transcribed the same recording twice. In total, we produced 90 transcripts, with 16 recordings transcribed more than once. See Figure 2 for a full histogram of the number of transcripts per recordings.

Table 1 shows the basic statistic on the manually annotated transcripts across the annotation methods and speaker roles. Notably, recordings transcribed from scratch tend to be shorter on average. Another interesting observation is that candidates' utterances contain fewer characters than those of the examiners, even though their durations are about the same. This is likely a consequence of the candidates' less fluent speech.

5. Qualitative Analysis of Transcription Methods

Transcribing spoken language accurately is a challenging task, whether performed by humans or automated speech recognition systems. To get a better insight into the differences between various transcribing approaches, we perform a comparative analysis of the most common differences. We contrast the output of the WhisperX-large-v2 model, with its manually post-edited variant and with the same recordings fully transcribed from scratch.

[15] One recording presents another format of the A2 exam that features two candidates. Another recording contains an additional examiner, who only comments on technical issues.

[16] The recordings have not yet been accompanied with metadata.

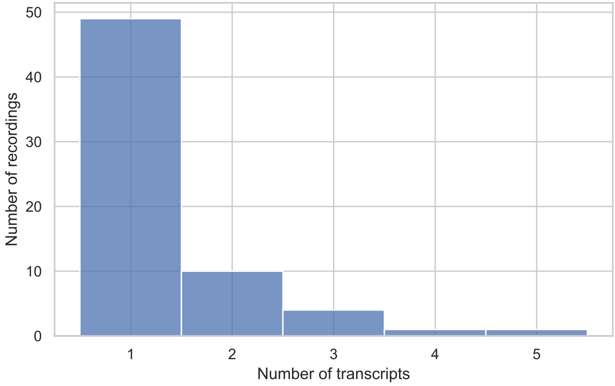


Figure 2. The histogram of the number of manually post-edited or fully manual transcripts per recording.

Method	Transcript count	Avg. duration (s)			Avg. char. count		
		Examiners	Candidates	All	Examiners	Candidates	All
WhisperX	23	191.42	156.41	347.83	1752.91	1004.74	2757.65
Mixed	41	174.09	171.15	345.24	1590.02	964.20	2554.22
From scratch	26	127.34	160.70	288.04	1115.62	861.85	1977.46
All	90	165.01	164.36	329.38	1494.60	944.99	2439.59

Table 1. The statistics on manually annotated transcripts across the annotation methods and speaker roles.

5.1. Typical Challenges of ASR in L2 Scenario

In the following, we list the most typical errors introduced by using ASR for transcription of non-native speakers and accompany them with the real examples found in the transcripts.

Corrections The Czech word for ‘thank you’ can be expressed as either “děkuju” or “děkuji.” The former is commonly used in spoken discourse, while the latter is considered stylistically more formal. In Example 1, the ASR system initially transcribed the spoken word “děkuju” as “děkuji,” thereby elevating it to a more formal register. However, human annotators accurately captured the utterance as “děkuju,” aligning with the speaker’s actual expression. Interestingly, in a subsequent instance within the same recording (Example 2), the ASR system correctly identified the term as “děkuju,” demonstrating consistency with both annotators.

(1) *WhisperX ASR:* “Děkuj**i**.”
WhisperX post-edit: “Děkuj**u**.”
From scratch: “Děkuj**u**.”

(2) *WhisperX ASR:* “Takže děkuj**u**.”
WhisperX post-edit: “Takže děkuj**u**.”
From scratch: “Takže děkuj**u**.”

Word Boundaries In Example 3, WhisperX incorrectly merged the Czech words “spolu” (‘together’) and “mluvit” (‘to speak’), creating a non-existent word amalgamation. This error was not replicated by human annotators, who correctly identified and separated the two words. Interestingly, the ASR system did not repeat this mistake in another segment of the recording (Example 4), successfully recognizing and separating the words “spolu” and “mluvit”.

(3) *WhisperX ASR:* “Zase budeme **spolumluvit**. A zase **spolumluvíme**.”
WhisperX post-edit: “Zase budeme **spolu mluvit**. A zase **spolu mluvíme**.”
From scratch: “Zase budeme **spolu mluvit**. A zase **spolu mluvíme**.”

(4) *WhisperX ASR:* “Teď budeme **spolu mluvit**.”
WhisperX post-edit: “Teď budeme **spolu mluvit**.”
From scratch: “Teď budeme **spolu mluvit**.”

Word Repetitions The intricacies of ASR systems in capturing spoken language nuances are highlighted in instances where repetition occurs in speech, a common phenomenon during impromptu discourse. In Example 5, the phrase “na tu na tu” was repeated in a recording, likely because the speaker searched for the right words. While the ASR failed to document this repetition, human annotators did not overlook it. Furthermore, the annotators diverged in their transcription of certain words: one captured “koukám,” (‘I see‘), the correct Czech form of the verb ”koukat” (’to see’), while the other heard “kukám”. Similarly, the pronoun “ona” (‘she‘) was recorded by one annotator, whereas the other noted “vona,” a colloquial variant. In addition, the word ″ještě″ (‘still‘) was transcribed by one annotator as a colloquial variant ″eště″ (in contrast to the another annotator and to the WhisperX system that captured the standard form).

(5) *WhisperX ASR:* ″Takže koukám, že **na tu** holku, ona ještě chodila někam do knihovny...″

WhisperX post-edit: ″Takže koukám, že **na tu na tu** holku, ona ještě chodila někam do knihovny...″

From scratch: ″Takže kukám, že **na tu na tu** holku, vona eště chodila někam do knihovny...″

Filler Words Furthermore, the ability of the ASR system to detect filler words becomes a point of interest. In Example 6, the WhisperX system (as well as both annotators) was successful at identifying the word “jakoby” that serves merely as a verbal filler due to its semantic redundancy in the captured statement. Conversely, the ASR omitted the filler word “no” (‘well’), a term frequently employed in spoken Czech, that was not overlooked by human annotators.

On the other hand, two annotators presented differing transcriptions of some other words in the utterance. While one annotator captured the standard Czech word “děti” (‘children’), the other annotator recorded “dětí”, reflecting the actual pronunciation by the non-native speaker, albeit incorrect in formal Czech usage (in the given context). Moreover, one annotator noted the pronunciation ″sedm,″ while the other documented ″sedum.″ Both are acceptable pronunciation variants of the word “sedm” (‘seven’) in Czech.

(6) *WhisperX ASR:* ″**Jakoby** to bude jenom děti, sedm osob.″

WhisperX post-edit: ″**No jakoby** to bude jenom děti, sedm osob.″

From scratch: ″**No jakoby** to bude jenom dětí. Sedum osob.″

Omitting Utterances The challenges of capturing spoken language are exemplified by the occasional failure to transcribe entire utterances. Example 7 illustrates this phenomenon in a case where both human annotators confirmed the presence of a response to the question “What did you do at Christmas?” in the recording. However, the ASR system omitted this response (″O Vánocích jsem byl v práci.″ ‘I was at work over Christmas.‘) entirely, instead proceeding to the subsequent question, “Kdy v vaší zemi nejvíc prší?” ‘When does it rain the most in your country?‘ Interestingly, a similar error was made also by a human, as only the *from scratch* annotator transcribed the intermediate question(“Kdy v vaší zemi nejvíc prší?”), while the other skipped it.

(7) *WhisperX ASR:* “Kdy v vaší zemi nejvíc prší? Kdy?”
WhisperX post-edit: “O Vánocích jsem byl v práci. Kdy, víc prší...”
From scratch: “O Vánocích jsem byl v práci. Kdy ve vaší zemi nejvíc prší? Kdy, víc prší.”

Ungrammatical Sentences The transcription of ungrammatical utterances also poses a challenge for ASR systems. Example 8 illustrates it in the sentence “To je pršet víc,” (lit. ‘It is rain more.‘) which was uniformly transcribed by two annotators despite its ungrammaticality in Czech. The correct grammatical form should be “To prší víc.” (‘It rains more.‘). The ASR system, confronted with the incorrect structure, substituted it with a grammatically correct but contextually unrelated sentence, “To je první věc.” (‘It is the first thing.‘).

(8) *WhisperX ASR:* “To je první věc.”
WhisperX post-edit: “To je pršet víc.”
From scratch: “To je pršet víc.”

Non-Existent Words In Example 9, WhisperX transcribed the non-existent Czech word “buzin,” presumably similar in sound to the utterance on the recording. This part was interpreted differently by two annotators: one captured it as “bazén” (‘pool’), a real Czech word, albeit inappropriate in its case form, and the other as “bazénu” (into ‘pool‘), which fits both contextually and morphologically. Additionally, the verb form “skočím” (‘I will jump‘) was recorded by the ASR system in its standard form and kept the same in the post-edited version by the annotator, while the annotator transcribing from scratch noted it as “skočim,” a non-standard spoken variant.

(9) *WhisperX ASR:* “Skočím někam do **buzin**.”
WhisperX post-edit: “Skočím někam do **bazén**.”
From scratch: “Skočim někam do **bazénu**.”

Incomprehensible Utterances Both the ASR system and human annotators encounter difficulties with transcription when faced with barely comprehensible or incomprehensible audio. This issue is illustrated in Example 10 where WhisperX and annotators produced slightly different transcriptions in an attempt to capture the exact utterance of the speaker. The resulting sentences from all three parties lacked naturalness, grammatical correctness, and meaningful content in the Czech language. In this instance, the non-native speaker failed to construct a cohesive and coherent segment of discourse.

(10)	*WhisperX ASR:*	"Se jenom zkoušet nikam nevidět."
	WhisperX post-edit:	"Sem na zkoušek nikam nevijdět."
	From scratch:	"Jsem zkoušet někam na výlet."

5.2. Summary

The transcription of spoken recordings into written text is a challenge, as evidenced by a case study that reveals both machines and humans are prone to errors in this complex task. Humans, however, tend to be more careful in their transcriptions, often capturing nuances that machines may overlook. Interestingly, even among native speakers, discrepancies in transcription can arise, reflecting individual differences in auditory perception and interpretation. This variability is particularly noticeable in instances of low audio intelligibility, where neither machine nor human can definitively guarantee transcription quality.

6. Quantitative Analysis of Transcription Methods

In order to figure out the most convenient approach for acquiring transcripts of recordings of Czech language exams for L2 speakers, we formulate three main research questions which can be answered by observing quantitative characteristics of the created dataset:

RQ1: Is manual post-editing of ASR outputs more efficient than fully manual transcription?

RQ2: Does post-editing enhance transcripts' consistency?

RQ3: Is the human post-edited transcription biased towards the ASR system it was based on?

The final transcription method must be efficient in time and human resources (RQ1). At the same time, the transcripts produced by different annotators should not be too diverse (RQ2). Last but not least, the human-annotated data, which are expected to be primarily used for testing purposes, should not be overly dependent (biased) on the system that has been used during the annotation process (RQ3).

Method	Annot. time per recording (s)	RTF	Standardized RTF	CER (%)		
				Examiners	Candidates	All
WhisperX	3983.71 ± 2067.94	**10.78 ± 4.28**	**-0.34 ± 0.84**	9.93 ± 5.56	24.36 ± 15.80	14.68 ± 8.16
Mixed	4571.11 ± 1822.95	12.58 ± 5.99	0.26 ± 1.11	19.87 ± 6.35	47.31 ± 11.83	30.06 ± 7.78
From scratch	**3577.38 ± 2121.30**	12.65 ± 6.15	-0.17 ± 0.81	100.00 ± 0.00	100.00 ± 0.00	100.00 ± 0.00
All	4116.97 ± 2027.19	12.27 ± 5.81	0.00 ± 1.00	40.48 ± 38.49	56.67 ± 31.25	46.33 ± 35.40

Table 2. Statistics on annotation efficiency across the annotation methods. The numbers are aggregated over the corresponding transcripts, for which the given statistic is available. Best results in bold.

6.1. RQ1: Is manual post-editing of ASR outputs more efficient than manual transcription?

Annotators are paid for the total time they spend on their annotation work. Therefore, we seek to answer this research question by measuring the annotation speed.

As mentioned in Section 3.2, the TEITOK environment records basic information about each editing session, including its start and end times. Summing over all editing sessions for a given transcript, we can then calculate how much time the manual annotation work takes. Note that as this feature have not been implemented in TEITOK from the beginning of manual annotation, we have collected this information for 75 out of 90 transcripts.

Table 2 shows the statistics of the annotation times and speed with regard to the transcription method. Focusing on average annotation times per recording, it shows that on average it takes around 70 minutes to process one recording (average duration 6 minutes). It also suggests that annotating from scratch is faster than post-editing WhisperX output. Post-editing a mixed output seems to be the slowest.

However, it must be accounted for that the recordings substantially differ in their length. We thus measure the *real-time factor* (*RTF*), which is a ratio between the annotation time and the recording's duration. For example, the RTF of 12 means that it takes 12 seconds to annotate a second of the recording. As seen in Table 2, the *WhisperX* method seems to be around 15% faster in terms of RTF than the *mixed* and *from scratch* methods that perform on par.

Nevertheless, RTF still ignores potentially different work pace across annotators. Figure 3 discloses that the fastest annotator works twice as fast as the slowest: Their RTFs are 9 and 18, respectively. Thus, we standardize each transcript's RTF by subtracting the annotator's mean RTF and dividing the difference by the standard deviation of the annotator's RTFs.[17] The standardized RTFs in Table 2 paint yet another

[17] One of the annotators was no longer working, when we introduced timestamp logging to TEITOK. Therefore, we talk about six annotators in total in Section 3.2 while Figure 3 shows density estimates only for five annotators.

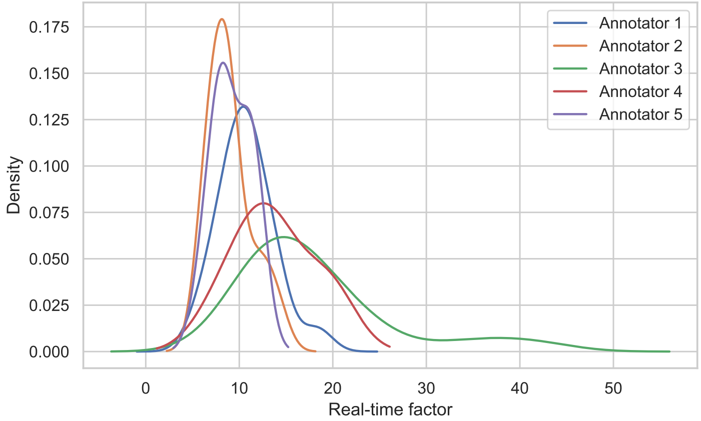


Figure 3. Density estimates of the distribution of real-time factors over transcripts across annotators.

picture: while *WhisperX* remains the fastest method, the annotation from scratch is actually faster than the *mixed* method.

The efficiency of annotation methods can also be approximated by the relative number of post-editing operations. We compute the number of additions, deletions, and substitutions relative to the length of the final text, all at the character level. Thus, it effectively corresponds to *Character Error Rate* (*CER*), where the post-edited transcript serves as the reference. It cannot be used for comparison of transcripts annotated from scratch by definition, as its CER is always 100%. However, the numbers in Table 2 confirm that the *mixed* method is much more demanding than the *WhisperX* method: it requires twice as many post-editing operations. There is also a substantial difference in CER across the speaker roles. The utterances spoken by candidates require almost 2.5 times more edits than those uttered by examiners. The high number of edit operations for L2 speakers likely makes the methods based on post-editing ASR outputs less efficient compared to the annotation from scratch.

6.2. RQ2: Does post-editing enhance transcripts' consistency?

It is important to strive for high consistency of the transcripts. That is, transcripts processed by different annotators or by the same annotator at different times should always follow the annotation guidelines to the highest possible extent. While transcript consistency is somewhat maintained by having a single annotator review all

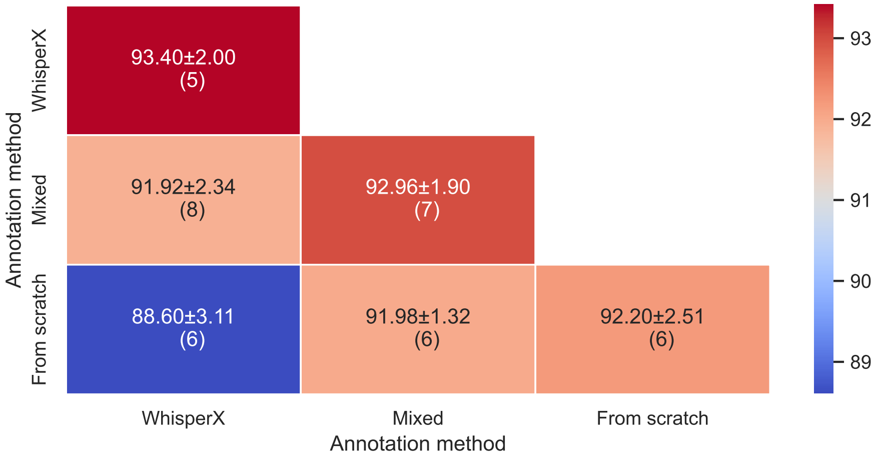


Figure 4. Inter-annotator agreement across annotation methods. The number in parentheses is the number of transcript pairs the statistics is based on.

transcripts during the second stage of manual annotation (see Section 3.2), the method used in the first stage should also aim to contribute significantly to this consistency.

To this end, we measure *Inter-annotator agreement* (*IAA*) across annotation methods. We limit ourselves only to recordings with multiple transcripts and calculate IAA between each pair of transcripts of the same recording. IAA for a pair of transcripts is computed as an average of two 1−CER scores, symmetrically taking each transcript in the pair as a reference.

For each combination of annotation methods, Figure 4 shows the mean IAA over all pairs sharing the combination. First, let us focus on the diagonal, which answers RQ2. The highest agreement is achieved if the transcripts are based on the WhisperX system, followed by the *mixed* method. Although annotating transcripts from scratch seems to be the least consistent, its difference to the IAA for post-edited WhisperX transcripts is not too big.

Comparison of IAA across annotation methods in Figure 4 shows that ASR-based transcripts are slightly shifted away from those produced from scratch. This is mostly pronounced for WhisperX transcripts which is the method resulting in lowest agreement with fully manual transcripts. The agreement of mixed transcripts is about the same as that of the manual transcripts and that of the WhisperX transcripts. It can be justified by its mixed nature, where part of it is formed by the Whisper model outputs. At the same time, some utterances may have been completely rewritten as suggested

ASR version	Annotation method			Mean
	WhisperX	Mixed	Scratch	
mms-1b-all	**40.8±9.7**	46.4±4.2	49.0±9.2	45.4±8.5
mms-1b-fl102	**31.2±7.9**	32.7±6.0	39.5±8.9	34.4±8.2
whisperX-large-v2	**13.3±5.6**	16.1±4.6	18.7±5.1	16.1±5.4
whisper-large-v3	**19.9±5.6**	24.6±8.2	25.7±6.7	23.5±7.3
Mean	**28.2±13.3**	31.1±13.8	37.4±15.1	

Table 3. Character error rates (CER) for each ASR system using reference transcripts derived from different annotation methods. ***Bold*** *indicates the minimum CER for each ASR system, while underline highlights the highest CER for each ASR system.*

by a relatively high number of editing operations in Table 2, which makes it similar to the annotation from scratch.

6.3. RQ3: Is human post-edited transcription biased towards the ASR system it was based on?

One potential downside of post-annotating ASR outputs is a bias toward the underlying transcript. This is especially important in our application, since we aim to capture possible deficiencies in pronunciation and language. The potential risk in our application stems from the fact that the ASR systems might not be robust to the difficult domain of non-native speakers and might strive for an artificially polished transcript (e.g., polished-out mispronunciation or stuttering).

We measure the extent of the bias towards a particular ASR system by measuring the character error rate (CER) of the ASR system transcript towards a particular annotation. The results are in Table 3. First, we note that WhisperX-large-v2 exhibits the smallest CER across all annotation methods, including *from scratch*. This indicates that WhisperX-large-v2 is the most robust ASR system in our study. The second most reliable ASR is then whisper-large-v3. The MMS models perform the worst. Second, we see that the post-edited transcripts from WhisperX observed the smallest CER across all ASR models, outperforming the *mixed* method and the *from scratch* method. This is surprising as the *mixed* method used random segments from all four ASR systems, leading us to anticipate that all ASR systems except whisperX-large-v2 would get a lower CER when evaluated against the reference annotated by the *mixed* method, compared to the *WhisperX* method. One possible explanation of the counter-intuitive observation is that a lot of segments predicted by the MMS models tend to be empty or with a very high CER, which gives the annotators no clues and makes the post-editing more tedious. This is also supported by our findings in Section 6.2, where we observe that the post-edited WhisperX transcripts show the highest inter-annotator agreement, and in Section 6.1, where we observed that post-editing WhisperX outputs is approximately 15% faster than post-editing mixed outputs and requires half as many editing operations. Consequently, it is apparent that there is some tendency of the underlying ASR transcript to influence the annotators.

To better understand possible bias, we perform a comparison of *mixed* and *WhisperX* post-edited transcripts with their *from-scratch* counterparts and compare how many times the annotator was influenced by the ASR transcript. We run pairwise word-level alignment on three versions of the transcript (the ASR, post-edited, and *from scratch* ones), and analyze the different situations for all word triplets found by the alignment. The most important case for our purpose is when, during the post-editing, the annotator agrees with the ASR but disagrees with the *from scratch* transcript, i.e. when the ASR has probably led to an error oversight. For example, the annotator kept "nějakou" from the ASR whereas the *from scratch* transcript contains "nějak". Another very common type is an omission of a false start in the post-edited transcript, e.g., "pívem" vs. "pí pívem".

The results are in Figure 5. As we can see in the figure, the case where the annotator is potentially influenced by the ASR (the "ASR-P=A" and "P-S=D" cells) is relatively rare among the various ASR models. For instance, the number 13 in the cell "ASR-P=A" and "P-S=D" means that there were 13 cases (aligned word triplets) when the post-editor kept the word as proposed by the ASR ("ASR-P=A") while the post-edit differs from the transcription from scratch ("P-S=D").

The highest number of possible influences is in the case of the *WhisperX* method (7% vs. 5% and 1% in other models). During a manual inspection of the triplets, we observed that the annotator was indeed influenced by the ASR system. The most common case was the influence of the correct spelling of an incorrectly pronounced word, for example:

(11) *WhisperX ASR:* "dob**ř**e děkuju"
WhisperX post-edit: "dob**ř**e děkuju"
From scratch: "dob**ž**e děkuju"

Another common type of influence was the omission of false starts, as shown in the following example:

(12) *WhisperX ASR:* "...mlékem a ten dont"
WhisperX post-edit: "...mlékem a ten dort"
From scratch: "...mlékem **a te** a ten dort"

A similar influence involved filler words that were captured in the from-scratch transcription but were missing in the post-edited version, as shown here:

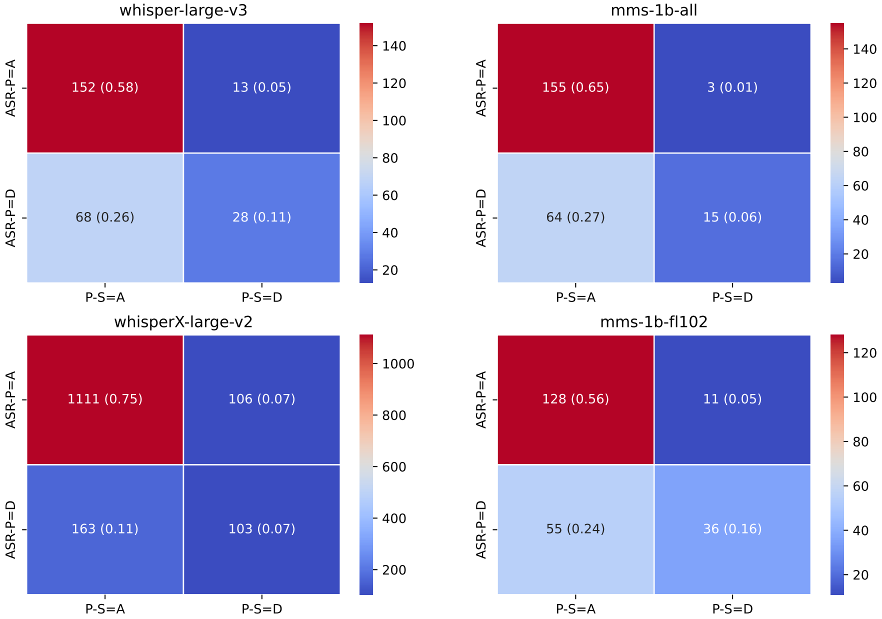


Figure 5. Analysis of the influence of the ASR transcript on the post-editing. Each confusion matrix represents agreement and disagreement between aligned word triplets (ASR↔post-edited↔from-scratch) with the relative frequencies in the brackets. P represents post-edited, S signifies from-scratch, and A and D indicate agreement and disagreement, respectively. Higher agreement between the ASR and post-edited transcripts (ASR-P=A), coupled with their disagreement with the from-scratch transcript (P-S=D), suggests a potential bias.

(13) *WhisperX ASR:* “kolik stojí ten dort”
WhisperX post-edit: “kolik stojí ten dort”
From scratch: “**hm** kolik stojí ten dolt ”

Therefore, we conclude that the ASR systems clearly tend to influence the annotator during post-editing. Strictly speaking, the bias implied by a particular ASR system may not be harmful to the final purpose of automatic spoken language assessment – an ASR may be ignoring speech errors that are irrelevant to the assessment – but such a situation still remains risky as the overall system would rely on the particular version of the ASR. Also, the biased post-edited transcript could be difficult to use for other purposes. Consequently, it remains uncertain whether the observed level of influence will impact the downstream task.

6.4. Summary

Annotating from scratch has the significant advantage of being inherently free from bias towards any ASR system. However, it may exhibit lower consistency and be less efficient to acquire. Our analysis shows that the decrease in these aspects is not substantial. Therefore, annotating from scratch is a good choice, guaranteeing no bias at the small cost of slightly lower speed and consistency.

The *mixed* method demonstrates decent inter-annotator agreement and appears less biased than the *WhisperX* method, likely due to the positive effect of mixing ASR outputs. However, the poor performance of MMS systems sometimes results in outputs so bad that post-editing essentially becomes annotating from scratch. This is reflected in the relatively lower speed of post-editing. Consequently, we do not recommend the *mixed* method for further transcription of exam recordings.

The *WhisperX* method shows the highest consistency and efficiency. However, we also observe a bias that might negatively impact the use of such transcripts as references for future evaluations of ASR systems, including new ones specialized in L2 speakers. Nonetheless, this bias may be less significant for the downstream task, which we cannot evaluate in the current setup.

When dealing with recordings of the A2-level candidates, it is safer to annotate from scratch. Nevertheless, the negative aspects of the WhisperX method may become less critical if we progress to transcribing recordings at higher levels of language competency.

7. Conclusion

In our article, we focused on the usability of ASR systems for transcribing spoken parts of Czech language proficiency exams for non-native speakers. The objectives of the study were two-fold: (1) to explore the most common cases where ASR masks errors in L2 speech, and (2) to compare fully manual and semi-automatic methods for obtaining reference transcriptions of the exams. The study was limited to exams at the A2 level.

Our analysis shows that it is safer, albeit slightly less efficient, to annotate transcriptions fully manually from scratch. Manual post-editing of WhisperX outputs proved to be competitive, especially in terms of efficiency and consistency. From comparing individual examples, we observed that the potential bias might be less significant for the downstream task. Moreover, we expect this bias to decrease with rising levels of speakers' language competence.

As we plan to continue annotating exam recordings for higher levels of language competence in the near future, we should repeat these experiments on a smaller scale to verify if the findings for the A2 level hold for higher levels as well.

Acknowledgements

The publication was funded by the Programme to Support Applied Research in the Area of the National and Cultural Identity for the Years 2023 to 2030 (NAKI III) of the Ministry of Culture of the Czech Republic within the project Automated Speech Scoring in Czech (DH23P03OVV037). We also thank Maarten Janssen for his assistance with the TEITOK environment and two anonymous reviewers for their insightful comments.

Bibliography

Adams, Oliver, Trevor Cohn, Graham Neubig, Hilaria Cruz, Steven Bird, and Alexis Michaud. Evaluating phonemic transcription of low-resource tonal languages for language documentation. In *11th International Conference on Language Resources and Evaluation, LREC 2018*, pages 3356–3365. European Language Resources Association (ELRA), 2019.

Baevski, Alexei, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460, 2020.

Bain, Max, Jaesung Huh, Tengda Han, and Andrew Zisserman. WhisperX: Time-Accurate Speech Transcription of Long-Form Audio. *INTERSPEECH 2023*, 2023. doi: 10.21437/Interspeech.2023-78.

Bredin, Hervé and Antoine Laurent. End-to-end speaker segmentation for overlap-aware resegmentation. In *Proc. Interspeech 2021*, Brno, Czech Republic, August 2021. doi: 10.21437/Interspeech.2021-560.

Bučková, Aneta. Languages in Migration. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University, 2023. URL `http://hdl.handle.net/11372/LRT-4777`.

Conneau, Alexis, Min Ma, Simran Khanuja, Yu Zhang, Vera Axelrod, Siddharth Dalmia, Jason Riesa, Clara Rivera, and Ankur Bapna. Fleurs: Few-shot learning evaluation of universal representations of speech. In *2022 IEEE Spoken Language Technology Workshop (SLT)*, pages 798–805. IEEE, 2023. doi: 10.1109/SLT54892.2023.10023141.

Cvejnová, Jitka and Ondřej Geppert. *Zkouška z češtiny pro trvalý pobyt v ČR (úroveň A2)*. Národní pedagogický institut České republiky, Praha, Czechia, 2022.

Graves, Alex, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd international conference on Machine learning*, pages 369–376, 2006. doi: 10.1145/1143844.1143891.

Holaj, Richard. *Nástroj pro automatický přepis řeči nerodilých mluvčí českého jazyka*. Disertační práce, Masarykova univerzita, Filozofická fakulta, 2023. URL `https://is.muni.cz/th/io67a/`.

Holaj, Richard and Petr Pořízka. ANOPHONE: An Annotation Tool for Phonemes and L2 Annotation Systems for Czech. *Journal of Linguistics/Jazykovedný casopis*, 74(1):333–344, 2023. doi: 10.2478/jazcas-2023-0050. URL `https://doi.org/10.2478/jazcas-2023-0050`.

Ivanová, Jaroslava, editor. *Společný evropský referenční rámec pro jazyky : jak se učíme jazykům, jak je vyučujeme a jak v jazycích hodnotíme*. Univerzita Palackého, Olomouc, 2. české vyd. edition, 2006.

Janssen, Maarten. A Corpus with Wavesurfer and TEI: Speech and Video in TEITOK. In Ekštein, Kamil, František Pártl, and Miloslav Konopík, editors, *Text, Speech, and Dialogue*, pages 261–268, Cham, 2021. Springer International Publishing. ISBN 978-3-030-83527-9. doi: 10.1007/978-3-030-83527-9_22.

Kubanek-German, Angelika. Early Language Programmes in Germany. In *An Early Start: Young Learners and Modern Languages in Europe and beyond*, Strasbourg, 2000. Council of Europe Publishing.

Lehečka, Jan, Jan Švec, Josef V. Psutka, and Pavel Ircing. Transformer-based Speech Recognition Models for Oral History Archives in English, German, and Czech. In *Proc. INTERSPEECH 2023*, pages 201–205, 2023. doi: 10.21437/Interspeech.2023-872.

Pečený, Pavel. Jak se připravovat k Certifikované zkoušce z češtiny pro cizince (CCE). In *Sborník Asociace učitelů češtiny jako cizího jazyka (AUČCJ)*, Praha, Czechia, 2012. Akropolis.

Pečený, Pavel. Oblasti zvyšování kvality jazykové zkoušky na příkladu Certifikované zkoušky z češtiny pro cizince (CCE). In *Zvyšování kvality výuky a testování cizích jazyků (včetně češtiny pro cizince)*, pages 87–92, Poděbrady, Czechia, 2013. ÚJOP UK.

Pratap, Vineel, Andros Tjandra, Bowen Shi, Paden Tomasello, Arun Babu, Sayani Kundu, Ali Elkahky, Zhaoheng Ni, Apoorv Vyas, Maryam Fazel-Zarandi, et al. Scaling speech technology to 1,000+ languages. *Journal of Machine Learning Research*, 25(97):1–52, 2024.

Radford, Alec, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR, 2023.

Schmiedtova, Barbara. Item "L2 Czech" in collection "Barbsch-L2 data". The Language Archive, 2000–2001. URL `https://hdl.handle.net/1839/00-0000-0000-0000-5D9B-2`.

Švec, Jan, Martin Bulín, Aleš Pražák, and Pavel Ircing. UWebASR – Web-based ASR engine for Czech and Slovak. In *Proceedings of CLARIN Annual Conference 2018*, pages 190–193, Pisa, Italy, 2018. CLARIN.

Vodičková, Kateřina, Pavel Pečený, and Jana Nováková. Specifikace Certifikované zkoušky z češtiny pro cizince a Společný evropský referenční rámec pro jazyky. In *Výuka a testování cizích jazyků v kontextu Společného evropského referenčního rámce (SERR)*, pages 78–90, Poděbrady, Czechia, 2012. ÚJOP UK.

Appendix

Here is a sample transcript of the recording. EXAM is the examiner (native Czech speaker) and CAND is the exam candidate (non-native Czech speaker).

EXAM: Dobrý den, můžu vám nějak pomoci?
CAND: Dobry den, můžete pomoct, je nějaká dobře dobře restaurac?
EXAM: Hm, já bych doporučila restauraci U Vejvodů.
CAND: A dobře a kde j zde je?
CAND: Ta restaurace.
EXAM: Hm, je na hlavním náměstí.
EXAM: Vidím, že jste autem.
EXAM: Pojedete rovně a doprava.
EXAM: A u nádraží zahnete vlevo.
CAND: Aha.
CAND: A jaké je tam je jídlo, je vegetariání?
CAND: Jídlo.
EXAM: Hm, ano, mají tam jídlo české, ale mají jídlo i pro vegetariány.
CAND: Dobře a tam je parkoviště?
EXAM: No, parkoviště je hned vedle restaurace, takže snadno zaparkujete.
CAND: Dobže, děkuju.

Address for correspondence:
Michal Novák
mnovak@ufal.mff.cuni.cz
Institute of Formal and Applied Linguistics
Faculty of Mathematics and Physics
Charles University
Malostranské náměstí 25
118 00 Praha 1, Czechia