The Impact of Code-switched Synthetic Data Quality is Task Dependent: Insights from MT and ASR

Injy Hamed,[1] **Ngoc Thang Vu,**[2] **Nizar Habash**[1,3]
[1]MBZUAI, [2]University of Stuttgart, [3]New York University Abu Dhabi
injy.hamed@mbzuai.ac.ae, thang.vu@ims.uni-stuttgart.de
nizar.habash@nyu.edu

Abstract

Code-switching, the act of alternating between languages, emerged as a prevalent global phenomenon that needs to be addressed for building user-friendly language technologies. A main bottleneck in this pursuit is data scarcity, motivating research in the direction of code-switched data augmentation. However, current literature lacks comprehensive studies that enable us to understand the relation between the quality of synthetic data and improvements on NLP tasks. We extend previous research conducted in this direction on machine translation (MT) with results on automatic speech recognition (ASR) and cascaded speech translation (ST) to test generalizability of findings. Our experiments involve a wide range of augmentation techniques, covering lexical replacements, linguistic theories, and back-translation. Based on the results of MT, ASR, and ST, we draw conclusions and insights regarding the efficacy of various augmentation techniques and the impact of quality on performance.

1 Introduction

Code-switching (CSW) is a worldwide phenomenon, involving the alternation between multiple languages in the same discourse.[1] Despite the need to process it effectively, language technologies still fall short when handling code-switched input compared to monolingual data (Doğruöz et al., 2021), where the lack of CSW resources is a main challenge. CSW data augmentation has thus been gaining attention as a workaround for alleviating this issue. Furthermore, the need for language technologies to not only process, but also generate CSW in human-computer interaction has been highlighted by researchers (Bawa et al., 2020) for the aim of building tools that cater to the needs and preferences of multilingual communities.

While considerable amount of research has been conducted on CSW data augmentation, we still lack comprehensive studies covering multiple augmentation techniques, human and extrinsic evaluations, and multiple NLP tasks. Such studies are needed to draw conclusions with regards to the improvements achieved by the different augmentation techniques on NLP tasks, the quality of the generated augmentations, and the relation between both; quality and improvements.

Several studies evaluate the effectiveness of different augmentation techniques extrinsically, however lack human evaluations assessing the quality of generations (Winata et al., 2018, 2019; Li and Vu, 2020; Gupta et al., 2021). Other studies include human evaluations, however, do not report results on downstream tasks (Pratapa and Choudhury, 2021; Kuwanto et al., 2024). Few studies involve both extrinsic and human evaluations. The study by Hussein et al. (2023) involved two augmentation approaches, however, their effectiveness was only reported in the scope of speech recognition. Hamed et al. (2022b) presented a study that is diverse in terms of extrinsic evaluations, covering MT, AST, and ST, however, the augmentation approaches were limited to lexical replacements only. Finally, Hamed et al. (2023) presented a comprehensive study covering multiple augmentation techniques, however, the extrinsic evaluation only covered the task of MT.

Given current literature, we cannot draw strong conclusions with regards to the effectiveness of the different techniques across different NLP tasks, as well as the relation between quality and improvements achieved on downstream tasks. In this paper, we aim at extending current literature with findings based on a more comprehensive setup in terms of investigated augmentation techniques and NLP tasks. To achieve that, we build on our experimental setup in Hamed et al. (2023), being the most comprehensive study in terms of augmentation ap-

[1]For survey papers on CSW in NLP, we refer the readers to the following papers: Sitaram et al. (2019); Doğruöz et al. (2021); Winata et al. (2023); Hamed et al. (2025).

Proceedings of the 7th Workshop on Computational Approaches to Linguistic Code-Switching, pages 6–17
May 3, 2025

proaches. We report results on ASR and cascaded ST, covering a wider range of NLP tasks investigated within the same experimental setup.

Our contributions are as follows:

- Following our previous experimental setup in Hamed et al. (2023), we report new results on ASR and ST. This allows us to make comparisons and draw conclusions based on five variations of augmentation approaches (covering linguistic-based approaches, lexical replacements, and back-translation) and three downstream tasks (MT, ASR, and ST).
- We present a discussion on the relation between the quality of generations and their effectiveness on NLP tasks in light of the results on ASR and ST as well as previous MT results. Our results show that with regards to the effectiveness of the techniques, some approaches are consistent in their performance across tasks, while others are task-dependent. Moreover, we find that the relation between data quality and NLP improvements, while confirmed for MT, does not hold for ASR.
- We explore and discuss other factors, besides quality of generations, that may affect results, including data diversity and task complexity.

The paper is organized as follows. Section 2 discusses related work. In Section 3, we provide an overview on the augmentation techniques included in the study. Section 4 is dedicated to the experimental setup. In Section 5, we present the ASR and ST results, as well as the correlations between quality of augmentations and ASR improvements. Finally, in Section 6, we provide further insights, discussing the possible impact of other factors.

2 Related Work

The majority of previous research on CSW data augmentation has addressed language modeling (LM), primarily for ASR. Various techniques have been investigated based on heuristics (Shen et al., 2011; Vu et al., 2012; Kuwanto et al., 2021), linguistic theories (Pratapa et al., 2018; Lee et al., 2019; Hussein et al., 2023), MT (Tarunesh et al., 2021), and generative models (Winata et al., 2018, 2019; Li and Vu, 2020) including large language models (LLMs) (Hu et al., 2023; Alharbi et al., 2024). MT has received less attention, where techniques mainly involved lexical replacements (Appicharla et al., 2021; Gupta et al., 2021; Xu and Yvon, 2021) and few efforts investigated back-translation (Kuwanto et al., 2023) and linguistic theories (Hamed et al., 2023).

With regards to studies conducting human evaluations without experimental results on downstream tasks, Pratapa and Choudhury (2021) compared lexical replacements and linguistic-based approaches, where higher human preference was observed for the latter approach. Recently, Kuwanto et al. (2024) investigated the use of the Equivalence Constraint theory (Poplack, 1980) when prompting LLMs by providing information on words that should be code-switched, showing slight improvements.

With regards to studies comparing different augmentation techniques through human evaluations as well as extrinsically, Hussein et al. (2023) compared random lexical replacements versus utilizing the Equivalence Constraint theory through human evaluation and ASR results. While the linguistic-based approach was found to be superior in the human evaluation, it was outperformed by random lexical replacements in language modeling and speech recognition. In Hamed et al. (2022b), we compared different approaches for lexical replacements. While the authors provide a comprehensive study, including human evaluation and results on MT, ASR, and ST tasks, the study is focused on experimental considerations for lexical replacements and does not include other augmentation approaches. In Hamed et al. (2023), we presented a comprehensive study covering multiple augmentation techniques, including linguistic-based approaches, lexical replacements, and back-translation. The study involved extrinsic evaluation on MT task in addition to human evaluation assessing the naturalness of the generations across techniques. A positive correlation was reported between the naturalness scores achieved by the different techniques and improvements on MT. However, given that the study is only focused on MT, it is still unclear whether the findings generalize to other NLP tasks. In this paper, we work towards filling the current research gap in comparative studies, extending literature with further findings in the area of CSW data augmentation.

3 Data Augmentation Techniques

We cover the same techniques and setup as in Hamed et al. (2023), where Arabic-English parallel sentences are utilized to generate CSW Arabic-English sentences using the approaches below.

3.1 Lexical Replacements

Dictionary Replacement (LEX$_{Dict}$): $X\%$ random Arabic words on the source side are replaced with their English gloss entries. The gloss entries are obtained using MADAMIRA (Pasha et al., 2014), a system that performs morphological analysis and disambiguation for Arabic. X is set to 19 based on the frequency found in naturally occurring CSW data (Hamed et al., 2022a).

Aligned with Random CSW Points (LEX$_{Rand}$): $X\%$ Arabic words on the source side are replaced with their counterpart English words on the target side based on alignments obtained using Giza++ (Och and Ney, 2003), as specified in Hamed et al. (2023). X is also set to 19.

Aligned with Predicted CSW Points (LEX$_{Pred}$): Instead of randomly choosing the words on the target side to be injected into the source side, a CSW predictive model is leveraged, where the model identifies the words on the target side that would be plausible CSW words on the source side. The CSW predictive model from Hamed et al. (2022b) is utilized for this task. The model is trained using ArzEn-ST corpus (Hamed et al., 2022a), containing CSW Arabic-English sentences and their English translations. In order to train the CSW predictive model, a matching algorithm was developed to tag the words on the target side that match the code-switched words on the source side. An mBERT model is then fine-tuned on this binary classification task, where given an English sentence, the model identifies which words are probable to be present in the CSW corresponding sentence. Afterwards, similar to the previous augmentation approach, target-to-source replacements are performed using alignments to inject these words into the source side sentence.

3.2 Linguistic Theories (EC and ML)

The GCM tool (Rizvi et al., 2021) is utilized to obtain CSW generations following two linguistic theories: Equivalence Constraint (EC) (Poplack, 1980) and Matrix Language Frame (MLF) (Myers-Scotton, 1997). The tool provides two approaches for sampling across the multiple generations it provides; random and Switch Point Fraction (SPF). In SPF sampling, the generations are ranked based on their SPF (Pratapa et al., 2018) distribution compared to a reference SPF distribution. The reference SPF (0.22) is calculated based on natural CSW data (Hamed et al., 2022a). Similar to the previous approaches, one generation is sampled per sentence for both sampling variants. We refer to the variants as EC$_{Rand}$, EC$_{SPF}$, ML$_{Rand}$, and ML$_{SPF}$.

3.3 Back-translation (BT)

A BT model (Hamed et al., 2023) is trained to translate English into CSW Arabic-English. The model is utilized to translate the target side of the Arabic-to-English parallel sentences to CSW sentences. The model is trained as a Transformer model using Fairseq (Ott et al., 2019) by utilizing the Arabic-English parallel corpora discussed in Section 4.1 in addition to ArzEn-ST corpus, where the approach is outlined in Hamed et al. (2023).

4 Experimental Setup

4.1 Data

In this Section, we specify the datasets used in (1) generating the augmentations and (2) training and evaluating the ASR systems.

For augmentation, we use the synthetic data generated in Hamed et al. (2023). The generations are obtained by augmenting $309k$ Arabic-English parallel sentences collected from the following corpora: Callhome Egyptian Arabic-English Speech Translation Corpus (Gadalla et al., 1997; LDC, 2002b,a; Kumar et al., 2014), LDC2012T09 (Zbib et al., 2012), LDC2017T07 (Chen et al., 2017), LDC2019T01 (Chen et al., 2019), LDC2021T15 (Tracey et al., 2021), and MADAR (Bouamor et al., 2018). Using the approaches outlined in Section 3, these monolingual parallel sentences are augmented into CSW Arabic-English sentences.

For ASR, we utilize ArzEn-ST, which is a CSW Arabic-English speech translation corpus. The corpus contains naturally occurring speech having frequent CSW (Hamed et al., 2020) along with its Arabic and English translations. The corpus is used in training, development and testing. ArzEn-ST train, dev, and test sets contain 3.3k, 1.4k, and 1.4k sentences (containing 2.2k, 0.9k, and 0.9k CSW sentences). For training, we also utilize Callhome (Gadalla et al., 1997) and MGB-3 (Ali et al., 2017) for Egyptian Arabic data, in addition to 5 hours from each of Librispeech (Panayotov et al., 2015) for English data, and MGB-2 (Ali et al., 2016) for Modern Standard Arabic (MSA) data. We perform Arabic Alif/Ya normalization, remove punctuation and corpus-specific annotations, and lower-case English words.

4.2 ASR Model

We use joint CTC/attention based end-to-end ASR systems using ESPnet (Watanabe et al., 2018). We apply SpecAugment (Park et al., 2019) and set the CTC/attention weight to 0.3. The encoder and decoder consist of 12 and 6 Transformer blocks with 4 heads, feed-forward inner dimension 2048 and attention dimension 256 (Karita et al., 2019). We use RNNLM consisting of 1 LSTM layer with 1000 hidden units trained for 20 epochs. For decoding, we set the beam size to 20 and CTC weight to 0.2. The LM is trained on the transcriptions of the ASR corpora, in addition to the synthetic CSW data in case of data augmentation experiments.

4.3 ST Model

We evaluate the effectiveness of the augmentation techniques on a cascaded ST task. We utilize our ASR models and the MT models from Hamed et al. (2023), where we train Transformer models using Fairseq. We report results on ArzEn-ST test set.

5 Results

We present ASR and ST results and discuss the relation between naturalness scores of the generations and improvements on ASR. For ASR, the full results are presented in Table 1. We present WER and CER on ArzEn-ST test set, for all sentences as well as CSW sentences only. We also report perplexity (PPL), out-of-vocabulary (OOV) rates, and the number of generations per technique. For ST, the full results are provided in Table 2, showing BLEU (Papineni et al., 2002), chrF, chrF++ (Popović, 2017), and BERTScore (F1) (Zhang et al., 2019), reported on all ArzEn-ST test set and the CSW sentences only. We provide the statistical significance for both tasks in Appendix A. The analysis in this section is based on the results on ArzEn-ST test set CSW sentences, using WER and chrF++, as CSW is our main concern. For easier comparison of results across ASR and MT, we also briefly discuss previous results obtained on MT.

5.1 ASR Results

We report results on the following two settings:

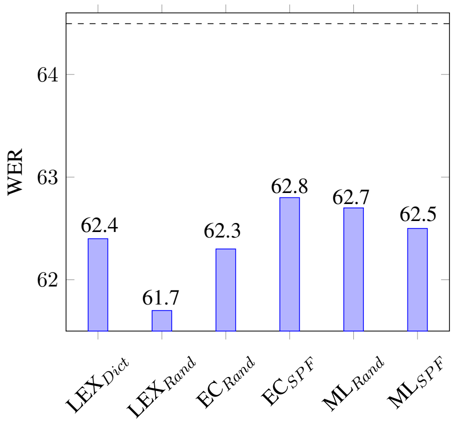
- Zero-shot setting: given the scarcity of CSW resources, we mimic the case of the lack of CSW corpora. We train a baseline model, ASR_BL$_{Mono}$, using the monolingual speech corpora for Egyptian Arabic, English, and MSA only. Data augmentation is performed using the techniques that do not require CSW parallel corpora: LEX$_{Dict}$, LEX$_{Rand}$, EC, and ML. The augmented CSW data along with the monolingual speech corpora transcriptions are used for LM rescoring.
- Non-zero-shot setting: this setting allows the use of CSW corpora. The baseline model, ASR_BL$_{All}$, is trained on the monolingual speech corpora in addition to ArzEn-ST. For augmentation, all techniques are applied.

We present WER results on ArzEn-ST test set CSW sentences in Figures 1a and 1b. The baseline models, ASR_BL$_{Mono}$ and ASR_BL$_{All}$, achieve 64.5% and 34.4% WER, respectively. For the zero-shot setting, among the linguistic theories, the best performance is achieved by EC$_{Rand}$. With regards to lexical replacements, LEX$_{Dict}$ provides comparable performance to linguistic theories and LEX$_{Rand}$ provides highest overall improvements, achieving absolute WER reduction of 2.8% over ASR_BL$_{Mono}$.

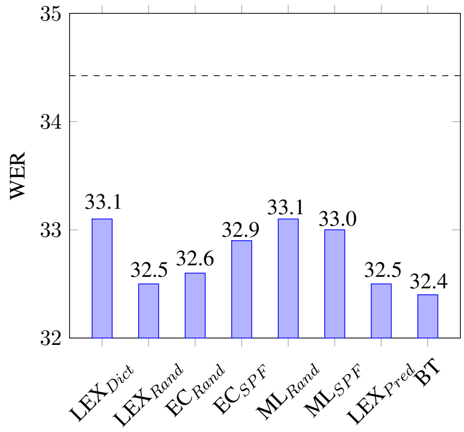
For the non-zero-shot setting, the best result is achieved by BT, achieving 2.0% absolute WER improvements over ASR_BL$_{All}$. By checking statistical significance, we find that LEX$_{Rand}$, LEX$_{Pred}$, and EC$_{Rand}$ provide equal performance to BT. This is followed by the other linguistic variants and LEX$_{Dict}$. It should be noted that LEX$_{Rand}$ proves to be a strong approach for ASR across both settings, while requiring no linguistic knowledge nor CSW data. This is in-line with the results of Hussein et al. (2023), where the superiority of random lexical replacements was demonstrated over the use of the Equivalence Constraint theory for ASR.

5.2 MT Results

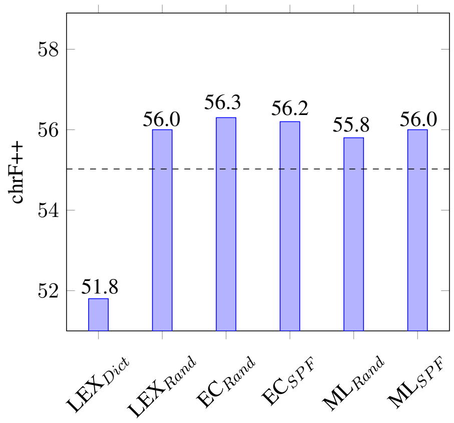
We include MT results from Hamed et al. (2023) in Figure 1. Similar to ASR, MT results cover zero-shot and non-zero-shot settings, with their respective baselines; MT_BL$_{Mono}$ and MT_BL$_{All}$. In case of the zero-shot setting, the MT models are trained on the Arabic-English parallel corpora outlined in Section 4.1, in addition to augmentations from the respective approaches. In case of the non-zero-shot setting, the training data of the MT models additionally included ArzEn-ST corpus. For a full discussion on MT experimental setup and results, we refer the readers to Hamed et al. (2023). Across both settings, LEX$_{Dict}$ degrades MT performance over baselines. We also report that linguistic-based approaches and LEX$_{Rand}$ perform



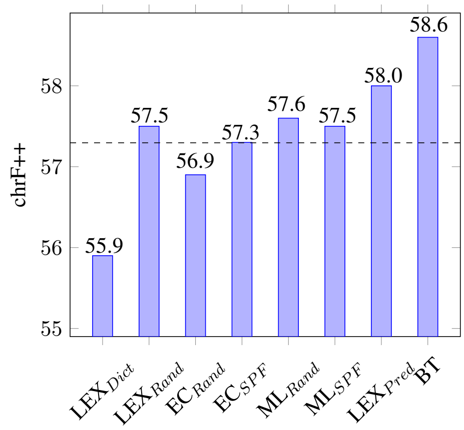
(a) The WER achieved by the ASR models in zero-shot setting. The dashed line represents ASR_BL$_{Mono}$.



(b) The WER achieved by the ASR models in non-zero-shot setting. The dashed line represents ASR_BL$_{All}$.



(c) The chrF++ scores achieved by the MT models in zero-shot setting. The dashed line represents MT_BL$_{Mono}$.



(d) The chrF++ scores achieved by the MT models in non-zero-shot setting. The dashed line represents MT_BL$_{All}$.

Figure 1: ASR and MT results on ArzEn-ST test set CSW sentences in zero-shot and non-zero-shot settings.

equally well, however, they are unable to achieve significant improvements over the baseline in the non-zero-shot setting. BT and LEX$_{Pred}$ show superiority, achieving +1.3 and +0.7 chrF++ points over the baseline, respectively.

5.3 ST Results

We present the chrF++ scores on ArzEn-ST test set CSW sentences for the non-zero-shot setting in Figure 2. The baseline, ST_BL$_{All}$, achieves 41.6 chrF++ points. We observe that LEX$_{Dict}$ does not outperform the baseline, where its overall performance on the ST task is affected by the low MT results. Among the linguistic theories, EC$_{SPF}$ performs best, and is the only variant that outperforms the baseline, providing similar performance to LEX$_{Rand}$. The best performance is achieved by BT followed by LEX$_{Pred}$, achieving improvements of +1.7 and +1.4 chrF++ points over ST_BL$_{All}$.

5.4 Effect of Quality on Performance

We examine the importance of generating natural CSW sentences in ASR LM rescoring. We utilize our human evaluation results from Hamed et al. (2023) and calculate the correlations against ASR scores. The human evaluation involved three annotators assessing 150 sentences across all augmentation techniques for naturalness on a scale of 1 to 5, following the rubrics introduced by Pratapa and Choudhury (2021). The mean opinion score (MOS) is calculated as the average of the annotators' scores for each sentence. The percentage of sentences perceived as natural (quite natural but rarely used - perfectly natural and frequently used) per technique is shown in Figure 3. We report correlations of 0.19 ($p = 0.73$) and -0.56 ($p = 0.15$) between the zero-shot and non-zero-shot ASR results (presented in Figure 1) and the percentage of sentences perceived as natural.

					All Test Sentences		CSW Test Sentences	
Exp	**Model**	**|Train|**	**PPL**	**OOV**	**WER**	**CER**	**WER**	**CER**
Baselines								
B1	ASR_BL$_{Mono}$	27,449	687.7	10.57	62.1	38.5	64.5	41.1
B2	ASR_BL$_{All}$	30,793	**415.1**	**5.57**	**34.7**	**20.0**	**34.4**	**20.0**
Zero-shot Experiments (ASR_BL$_{Mono}$+Augmentations)								
A1	+LEX$_{Dict}$	267,093	396.2	6.62	60.0	37.0	62.4	39.6
A2	+LEX$_{Rand}$	220,101	**364.6**	**5.70**	**59.5**	**36.9**	**61.7**	**39.4**
A3	+EC$_{Rand}$	169,549	460.0	6.23	60.2	37.2	62.3	39.6
A4	+EC$_{SPF}$	169,549	438.8	6.25	60.6	37.4	62.8	39.9
A5	+ML$_{Rand}$	125,681	455.3	6.37	60.5	37.3	62.7	39.8
A6	+ML$_{SPF}$	125,681	460.9	6.36	60.4	37.4	62.5	39.9
Non-zero-shot Experiments (ASR_BL$_{All}$+Augmentations)								
A7	+LEX$_{Dict}$	270,437	318.6	4.16	33.3	19.3	33.1	19.3
A8	+LEX$_{Rand}$	223,445	274.1	**3.88**	**32.9**	18.9	32.5	18.8
A9	+LEX$_{Pred}$	143,735	**270.4**	**3.88**	33.0	18.9	32.5	18.8
A10	+EC$_{Rand}$	172,893	301.0	3.95	33.1	18.9	32.6	18.8
A11	+EC$_{SPF}$	172,893	309.7	3.93	33.4	19.1	32.9	19.0
A12	+ML$_{Rand}$	129,025	313.7	4.11	33.7	19.3	33.1	19.2
A13	+ML$_{SPF}$	129,025	297.4	4.09	33.5	19.2	33.0	19.0
A14	+BT	181,868	275.3	3.96	**32.9**	**18.8**	**32.4**	**18.7**
Constrained Experiments (ASR_BL$_{All}$+Constrained[Augmentations])								
A15	+LEX$_{Dict}$	55,636	410.2	4.57	34.3	19.7	33.8	19.6
A16	+LEX$_{Rand}$	55,636	384.8	**4.39**	34.0	19.5	33.4	19.4
A17	+LEX$_{Pred}$	55,636	385.4	4.42	34.2	19.5	33.7	19.5
A18	+EC$_{Rand}$	55,636	394.5	4.50	34.2	19.6	33.6	19.5
A19	+EC$_{SPF}$	55,636	446.2	4.48	34.6	19.7	34.0	19.6
A20	+ML$_{Rand}$	55,636	435.5	4.54	34.6	19.8	34.2	19.8
A21	+ML$_{SPF}$	55,636	416.1	4.54	34.6	19.7	34.1	19.6
A22	+BT	55,636	**361.9**	4.41	**33.7**	**19.3**	**33.2**	**19.2**

Table 1: We report ASR results using WER and CER on ArzEn-ST test set, for all sentences as well as CSW sentences only. We also report PPL and OOV on all sentences of ArzEn-ST test set. We report the results of the baselines, zero-shot and non-zero-shot settings and well as the constrained settings. Given the varying amounts of generations produced by each technique, we also report the number of sentences used in training each model. The best performing models in each setting are bolded. The overall best performing model is underlined.

Unlike ASR, strong positive correlations of 0.92 ($p < 0.05$) and 0.97 ($p < 0.05$) were reported in zero-shot and non-zero-shot MT settings between chrF++ and naturalness scores.

To eliminate the factor of varying amounts of generations across techniques, we conduct constrained experiments (results in Table 1), where we only utilize the synthetic sentences augmented across all approaches ($= 24.8k$ sentences) for LM rescoring. We report a correlation of -0.26 ($p = 0.54$) between naturalness scores and ASR performance. Therefore, we conclude that for ASR, producing more natural synthetic data does not necessarily entail improvements in ASR LM rescoring.

6 Discussion

In this section, we share more insights to gain further understanding of other factors affecting results.

6.1 Consistency of Results Across Tasks

We discuss consistency of findings across tasks by comparing our ASR and MT results. With regards to the efficacy of the techniques, we observe that linguistic theories do not show superiority, and that the best results are achieved by BT followed by LEX$_{Pred}$. The performance of LEX$_{Dict}$ is found to be task-dependent, where it is effective in ASR but not suitable for MT, as the semantics of the original sentences may be altered. With regards

Exp	Model	All Test Sentences BLEU	chrF	chrF++	BertScore	CSW Test Sentences BLEU	chrF	chrF++	BertScore
Baseline									
	+ST_BL$_{All}$	15.9	42.2	40.3	0.335	16.4	43.7	41.6	0.318
Non-zero-shot Experiments									
A7	+LEX$_{Dict}$	15.7	42.1	40.2	0.343	16.1	43.2	41.2	0.322
A8	+LEX$_{Rand}$	15.9	42.7	40.7	0.347	16.5	44.1	42.0	0.329
A9	+LEX$_{Pred}$	**17.3**	43.5	41.7	**0.351**	**17.9**	44.9	43.0	0.335
A10	+EC$_{Rand}$	15.7	42.5	40.5	0.343	16.1	43.9	41.7	0.324
A11	+EC$_{SPF}$	16.5	42.8	40.9	0.348	17.1	44.2	42.2	0.334
A12	+ML$_{Rand}$	16.0	42.6	40.6	0.342	16.4	43.9	41.8	0.323
A13	+ML$_{SPF}$	16.0	42.6	40.6	0.346	16.5	44.0	41.8	0.330
A14	+BT	16.9	**43.7**	**41.8**	0.349	17.7	**45.4**	**43.3**	**0.337**

Table 2: We report ST results using BLEU, chrF, chrF++, and BertScore (F1) on ArzEn-ST test set, for all sentences as well as CSW sentences only. We report the results of the baseline and augmentation experiments.

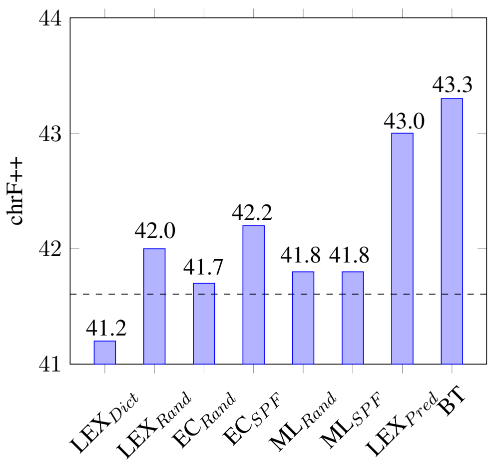


Figure 2: The chrF++ scores achieved in ST on ArzEn-ST test set CSW sentences in non-zero-shot setting. The dashed line represents the baseline model ST_BL$_{All}$.

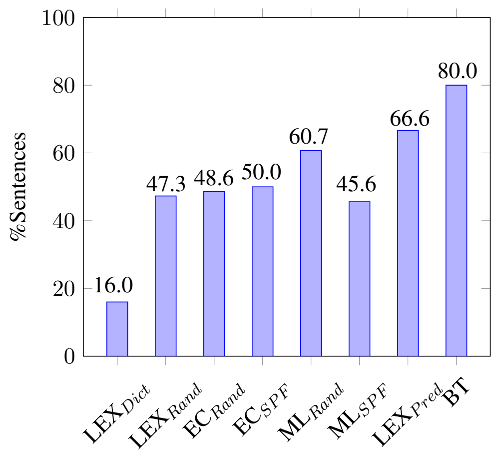


Figure 3: The human evaluation scores as obtained from (Hamed et al., 2023), showing the percentage of augmentations perceived as natural per technique.

to the relation between naturalness of generations and NLP improvements, a strong correlation was found for MT, but no correlation for ASR. The importance of quality is also seen in MT, where only BT and LEX$_{Pred}$ brought improvements over the baseline in the non-zero-shot setting, as opposed to all approaches in ASR.

6.2 Inconsistent Quality-Performance Relation Across Tasks

We further examine why the relation between quality and performance is not consistent across tasks. One factor that may affect this relation is the complexity of the NLP tasks and how well the baseline models perform on CSW. We conduct an error analysis on 100 sentences from ArzEn-ST dev set using the ASR and MT baseline models. We find that 70% of the sentences in the case of ASR have CSW-related issues as opposed to 25% in the case of MT. We provide examples in Table 3 demonstrating this disparity in performance. This may be a contributing factor, where quality might be less relevant to low-performing models. While CSW introduces further challenges to ASR, in the case of MT when translating to the primary/secondary language, the translation is partially present in the source sentence, allowing the model to perform better on CSW over monolingual sentences, as shown in Gaser et al. (2022). CSW quality can then be important for the model to not just retain words through translation but to learn the modifications often needed to achieve higher fluency.

	Examples
ASR Ref	بس ده ده one of my dream jobs يعني
ASR baseline	بس ده ده انا my dream job يعني
ASR BT	بس ده برضه my dream job يعني
MT Ref	But this is one of my dream jobs actually.
MT baseline	but this.. one of my dream jobs i mean
MT BT	but this.. this is one of my dream jobs
ASR Ref	و يعني i think اسوان most beautiful city في مصر كلها
ASR baseline	و يعني i think اصلا ال student في مصر كلها
ASR BT	و يعني i think ال student في مصر كلها
MT Ref	And actually, i think Aswan is the most beautiful city in all Egypt.
MT baseline	and i think aswan most beautiful city in all of egypt
MT BT	and i mean i think aswan is most beautiful city in all of egypt
ASR Ref	ال semester ده في ال graphic بنعمل posters كتير اوي
ASR baseline	semester نعمل ال دراستك ده في projects كتير اوي
ASR BT	semester بنعمل ال دراستك ده في posters كتير اوي
MT Ref	We're designing many posters this semester in graphic.
MT baseline	this semester in graphic we make posters a lot (We mark *posters a lot* as incorrect as the output incorrectly follows the same syntactic structure as the original CSW sentence, where the Arabic adjective كتير *ktyr* 'many' follows the English noun *posters*.)
MT BT	this semester in graphic we make a lot of posters
ASR Ref	انا بتعلم حاجات كتير جدا منها ان انا ازاي اعمل web applications ازاي اعمل applications ال mobile ازاي اعمل games ازاي حاجات hardware برضه embedded systems
ASR baseline	و بتعملي حاجات كتير جدا منها ان انا ازاي عامل ويبقي precautions ازاي اعمل applications ال mobile ازاي اعمل games ازاي حاجات hardware برضه implications
ASR BT	و اتعلمت حاجات كتير جدا منها ان انا ازاي عملوا يبقي precautions ازاي اعمل applications ال mobile ازاي اعمل games ازاي حاجات hardware برضه implications
MT Ref	I am learning a lot of things, including how to develop web applications, how to develop mobile applications, how to develop games .. as well as hardware things such as embedded systems
MT baseline	i learn a lot of things, including how to do web applications, how to make applications the mobile, how to make games, how.. hardware embedded systems
MT BT	i learn a lot of things, including how to make web applications, how to make applications the mobile, how to make games, how.. hardware also embedded systems

Table 3: Examples of outputs of ASR and MT systems. For each example, we show the reference transcription (ASR Ref) and translation (MT Ref) as well as the outputs of the baseline and BT augmentation models in the non-zero-shot setting. The words in the transcriptions/translations are highlighted according to whether they are correct (green) or incorrect (red and underlined) with regards to CSW. Given that Arabic is written from right to left, all sentences are displayed in a right-to-left orientation.

6.3 Other Factors Affecting Performance

We investigate other factors besides quality that may impact the effectiveness of the augmentation techniques, by checking their correlations against MT and ASR non-zero-shot results. With regards to the varying quantity of generated augmentations across techniques, while it may affect results, it holds a low correlation of -0.01 ($p = 0.98$) and -0.60 ($p = 0.12$) with ASR and MT results. We also check correlations against perplexity and OOV rate, where strong correlations of 0.89 ($p = 0.003$) and 0.84 ($p = 0.008$) are found for ASR. For MT, a lower correlation of -0.77 ($p = 0.027$) is found for perplexity (implementation details in Appendix B). We do not report correlations with OOV rate for MT, as it is the same value for the majority of augmentation techniques.[2] We agree with Hashimoto et al. (2019) that perplexity captures diversity but not quality, while human evaluation captures quality but not diversity, where we believe both criteria affect augmentation performance. Accordingly, the high performance achieved by BT and LEX$_{Pred}$ across ASR and MT tasks could be supported by their high performance on both criteria.

6.4 Perplexity as a Quality Measure

While perplexity has been previously used to measure the quality of generated CSW and monolingual augmented data (Winata et al., 2018; Feng et al., 2020; Evuru et al., 2024), we report a low correlation of -0.62 ($p = 0.10$) with naturalness scores. This highlights the importance of assessing naturalness through human evaluations as well as the need for further research towards developing automatic quality evaluation methods for CSW synthetic data.

7 Conclusions and Outlook

We investigate the efficacy of multiple CSW data augmentation approaches and the relation between quality of generations and improvements. We extend our previous work on MT with results on ASR and ST. We find that back-translation and predictive-based lexical replacements perform consistently well, however, quality of generations are found to be less important for ASR than MT models. We shed light on multiple factors that come into play, including diversity of generations as well as task complexity and model performance.

In future work, we plan on expanding the investigated approaches, with a focus on utilizing large language models. We also plan on exploring personalized CSW text generation.

[2] Both Arabic and English sentences of the parallel corpora are used on the source side when training the MT models, so no new words are introduced for LEX$_{Rand}$, LEX$_{Pred}$, and linguistic-based approaches.

Limitations

While this paper provides a comprehensive comparison of CSW augmentation techniques, in terms of the number of augmentation methods and the range of NLP tasks considered, we acknowledge that the coverage is limited to one language pair. Further research is needed to assess the generalizability of our findings across different languages. Additionally, we also acknowledge that LLM-based CSW generation is an interesting direction that is gaining attention (Yong et al., 2023; Potter and Yuan, 2024; Alharbi et al., 2024; Kuwanto et al., 2024). Further research is needed to assess its effectiveness compared to the approaches presented in this work, which we leave for future work.

Acknowledgments

We thank Sunayana Sitaram for her insightful discussions. We also thank the reviewers for their valuable comments and constructive feedback.

References

Sadeen Alharbi, Reem Binmuqbil, Ahmed Ali, Raghad Aloraini, Saiful Bari, Areeb Alowisheq, and Yaser Alonaizan. 2024. Leveraging LLM for augmenting textual data in code-switching ASR: Arabic as an example. In *Proceedings of SynData4GenAI*.

Ahmed Ali, Peter Bell, James Glass, Yacine Messaoui, Hamdy Mubarak, Steve Renals, and Yifan Zhang. 2016. The MGB-2 challenge: Arabic multi-dialect broadcast media recognition. In *Proceedings of SLT*.

Ahmed Ali, Stephan Vogel, and Steve Renals. 2017. Speech recognition challenge in the wild: Arabic MGB-3. In *Proceedings of ASRU*, pages 316–322.

Ramakrishna Appicharla, Kamal Kumar Gupta, Asif Ekbal, and Pushpak Bhattacharyya. 2021. IITP-MT at CALCS2021: English to Hinglish neural machine translation using unsupervised synthetic code-mixed parallel corpus. In *Proceedings of the Workshop on Computational Approaches to Linguistic Code-Switching (CALCS)*, pages 31–35.

Anshul Bawa, Pranav Khadpe, Pratik Joshi, Kalika Bali, and Monojit Choudhury. 2020. Do multilingual users prefer chat-bots that code-mix? let's nudge and find out! *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW1):1–23.

Houda Bouamor, Nizar Habash, Mohammad Salameh, Wajdi Zaghouani, Owen Rambow, Dana Abdulrahim, Ossama Obeid, Salam Khalifa, Fadhl Eryani, Alexander Erdmann, and Kemal Oflazer. 2018. The MADAR Arabic Dialect Corpus and Lexicon. In *Proceedings of LREC*, pages 3387–3396.

Song Chen, Dana Fore, Stephanie Strassel, Haejoong Lee, and Jonathan Wright. 2017. BOLT Egyptian Arabic SMS/chat and transliteration LDC2017T07. Philadelphia: Linguistic Data Consortium.

Song Chen, Jennifer Tracey, Christopher Walker, and Stephanie Strassel. 2019. BOLT Arabic discussion forum parallel training data LDC2019T01. Philadelphia: Linguistic Data Consortium.

A Seza Doğruöz, Sunayana Sitaram, Barbara E Bullock, and Almeida Jacqueline Toribio. 2021. A survey of code-switching: Linguistic and social perspectives for language technologies. In *Proceedings of ACL-IJCNLP*, pages 1654–1666.

Chandra Kiran Reddy Evuru, Sreyan Ghosh, Sonal Kumar, Utkarsh Tyagi, Dinesh Manocha, et al. 2024. Coda: Constrained generation based data augmentation for low-resource NLP. In *Proceeding of NAACL*.

Steven Y Feng, Varun Gangal, Dongyeop Kang, Teruko Mitamura, and Eduard Hovy. 2020. GenAug: Data augmentation for finetuning text generators. In *Proceedings of the Deep Learning Inside Out Workshop*.

Hassan Gadalla, Hanaa Kilany, Howaida Arram, Ashraf Yacoub, Alaa El-Habashi, Amr Shalaby, Krisjanis Karins, Everett Rowson, Robert MacIntyre, Paul Kingsbury, David Graff, and Cynthia McLemore. 1997. Callhome Egyptian Arabic transcripts. *Linguistic Data Consortium, Philadelphia*.

Marwa Gaser, Manuel Mager, Injy Hamed, Nizar Habash, Slim Abdennadher, and Ngoc Thang Vu. 2022. Exploring segmentation approaches for neural machine translation of code-switched Egyptian Arabic-English text. In *Proceedings of EACL*, pages 86–100.

Abhirut Gupta, Aditya Vavre, and Sunita Sarawagi. 2021. Training data augmentation for code-mixed translation. In *Proceedings of NAACL-HLT*, pages 5760–5766.

Injy Hamed, Nizar Habash, Slim Abdennadher, and Ngoc Thang Vu. 2022a. ArzEn-ST: A three-way speech translation corpus for code-switched Egyptian Arabic-English. In *Proceedings of the Arabic Natural Language Processing Workshop (WANLP)*.

Injy Hamed, Nizar Habash, Slim Abdennadher, and Ngoc Thang Vu. 2022b. Investigating lexical replacements for Arabic-English code-switched data augmentation. In *Proceedings of the Workshop on Technologies for Machine Translation of Low-Resource Languages (LoResMT)*, pages 86–100.

Injy Hamed, Nizar Habash, and Ngoc Thang Vu. 2023. Data augmentation techniques for machine translation of code switched texts: A comparative study. In *Findings of EMNLP*, pages 140–154.

Injy Hamed, Caroline Sabty, Slim Abdennadher, Ngoc Thang Vu, Thamar Solorio, and Nizar Habash. 2025. A survey of code-switched Arabic NLP: Progress, challenges, and future directions. In *Proceedings of COLING*, pages 4561–4585.

Injy Hamed, Ngoc Thang Vu, and Slim Abdennadher. 2020. ArzEn: A speech corpus for code-switched Egyptian Arabic-English. In *Proceedings of LREC*, pages 4237–4246.

Tatsunori B Hashimoto, Hugh Zhang, and Percy Liang. 2019. Unifying human and statistical evaluation for natural language generation. In *Proceedings of NAACL*, pages 1689–1701.

Ke Hu, Tara N Sainath, Bo Li, Yu Zhang, Yong Cheng, Tao Wang, Yujing Zhang, and Frederick Liu. 2023. Improving multilingual and code-switching asr using large language model generated text. In *Proceedings of ASRU*, pages 1–7.

Amir Hussein, Shammur Absar Chowdhury, Ahmed Abdelali, Najim Dehak, Ahmed Ali, and Sanjeev Khudanpur. 2023. Textual data augmentation for Arabic-English code-switching speech recognition. In *Proceedings of SLT*, pages 777–784.

Shigeki Karita, Nelson Enrique Yalta Soplin, Shinji Watanabe, Marc Delcroix, Atsunori Ogawa, and Tomohiro Nakatani. 2019. Improving transformer-based end-to-end speech recognition with connectionist temporal classification and language model integration. In *Proceedings of Interspeech*.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. In *Proceedings of ICLR*.

Gaurav Kumar, Yuan Cao, Ryan Cotterell, Chris Callison-Burch, Daniel Povey, and Sanjeev Khudanpur. 2014. Translations of the CALLHOME Egyptian Arabic corpus for conversational speech translation. In *Proceedings of IWSLT*, pages 244–248.

Garry Kuwanto, Chaitanya Agarwal, Genta Indra Winata, and Derry Tanti Wijaya. 2024. Linguistics theory meets LLM: Code-switched text generation via equivalence constrained large language models. *arXiv preprint arXiv:2410.22660*.

Garry Kuwanto, Afra Feyza Akyürek, Isidora Chara Tourni, Siyang Li, Alex Jones, and Derry Wijaya. 2023. Low-resource machine translation training curriculum fit for low-resource languages. In *Proceedings of the Pacific Rim International Conference on Artificial Intelligence*, pages 453–458.

Garry Kuwanto, Afra Feyza Akyürek, Isidora Chara Tourni, Siyang Li, and Derry Wijaya. 2021. Low-resource machine translation for low-resource languages: Leveraging comparable data, code-switching and compute resources. *arXiv preprint arXiv:2103.13272*.

LDC. 2002a. 1997 HUB5 Arabic transcripts – LDC2002T39. Web Download. Philadelphia: Linguistic Data Consortium.

LDC. 2002b. CALLHOME Egyptian Arabic transcripts supplement – LDC2002T38. Web Download. Philadelphia: LDC.

Grandee Lee, Xianghu Yue, and Haizhou Li. 2019. Linguistically motivated parallel data augmentation for code-switch language modeling. In *Proceedings of Interspeech*, pages 3730–3734.

Chia-Yu Li and Ngoc Thang Vu. 2020. Improving code-switching language modeling with artificially generated texts using cycle-consistent adversarial networks. In *Proceedings of Interspeech*, pages 1057–1061.

Carol Myers-Scotton. 1997. *Duelling languages: Grammatical structure in codeswitching*. Oxford University Press.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. FAIRSEQ: A fast, extensible toolkit for sequence modeling. In *Proceedings of NAACL: Demonstrations*, pages 48–53.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: an ASR corpus based on public domain audio books. In *Proceedings of ICASSP*, pages 5206–5210.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of ACL*, pages 311–318.

Daniel S Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D Cubuk, and Quoc V Le. 2019. SpecAugment: A simple data augmentation method for automatic speech recognition. In *Proceedings of Interspeech*, pages 2613–2617.

Arfath Pasha, Mohamed Al-Badrashiny, Mona Diab, Ahmed El Kholy, Ramy Eskander, Nizar Habash, Manoj Pooleery, Owen Rambow, and Ryan Roth. 2014. MADAMIRA: A fast, comprehensive tool for morphological analysis and disambiguation of Arabic. In *Proceedings of LREC*, pages 1094–1101.

Shana Poplack. 1980. Sometimes i'll start a sentence in spanish y termino en espanol: toward a typology of code-switching1. *Linguistics*, 18(7-8):581–618.

Maja Popović. 2017. chrF++: words helping character n-grams. In *Proceedings of the Conference on Machine Translation*, pages 612–618.

Tom Potter and Zheng Yuan. 2024. LLM-based code-switched text generation for grammatical error correction. In *Proceedings of EMNLP*.

Adithya Pratapa, Gayatri Bhat, Monojit Choudhury, Sunayana Sitaram, Sandipan Dandapat, and Kalika Bali. 2018. Language modeling for code-mixing: The role of linguistic theory based synthetic data. In *Proceedings of ACL*, pages 1543–1553.

Adithya Pratapa and Monojit Choudhury. 2021. Comparing grammatical theories of code-mixing. In *Proceedings of the Workshop on Noisy User-generated Text*, pages 158–167.

Mohd Sanad Zaki Rizvi, Anirudh Srinivasan, Tanuja Ganu, Monojit Choudhury, and Sunayana Sitaram. 2021. GCM: A toolkit for generating synthetic code-mixed text. In *Proceedings of EACL: System Demonstrations*, pages 205–211.

Han-Ping Shen, Chung-Hsien Wu, Yan-Ting Yang, and Chun-Shan Hsu. 2011. CECOS: A Chinese-English code-switching speech database. In *Proceedings of Oriental COCOSDA*, pages 120–123.

Sunayana Sitaram, Khyathi Raghavi Chandu, Sai Krishna Rallabandi, and Alan W Black. 2019. A survey of code-switched speech and language processing. *arXiv preprint arXiv:1904.00784*.

Ishan Tarunesh, Syamantak Kumar, and Preethi Jyothi. 2021. From machine translation to code-switching: Generating high-quality code-switched text. In *Proceedings of ACL-IJCNLP*, pages 3154–3169.

Jennifer Tracey et al. 2021. BOLT Egyptian Arabic SMS/chat parallel training data LDC2021T15. Web Download. Philadelphia: LDC.

Ngoc Thang Vu, Dau-Cheng Lyu, Jochen Weiner, Dominic Telaar, Tim Schlippe, Fabian Blaicher, Eng-Siong Chng, Tanja Schultz, and Haizhou Li. 2012. A first speech recognition system for Mandarin-English code-switch conversational speech. In *Proceedings of ICASSP*, pages 4889–4892.

Shinji Watanabe, Takaaki Hori, Shigeki Karita, Tomoki Hayashi, Jiro Nishitoba, Yuya Unno, Nelson Enrique Yalta Soplin, Jahn Heymann, Matthew Wiesner, Nanxin Chen, et al. 2018. ESPnet: End-to-end speech processing toolkit. In *Proceedings of Interspeech*, pages 2207–2207.

Genta Winata, Alham Fikri Aji, Zheng Xin Yong, and Thamar Solorio. 2023. The decades progress on code-switching research in NLP: A systematic survey on trends and challenges. In *Findings of ACL*, pages 2936–2978.

Genta Indra Winata, Andrea Madotto, Chien-Sheng Wu, and Pascale Fung. 2018. Learn to code-switch: Data augmentation using copy mechanism on language modeling. *CoRR*, abs/1810.10254.

Genta Indra Winata, Andrea Madotto, Chien-Sheng Wu, and Pascale Fung. 2019. Code-switched language models using neural based synthetic data from parallel sentences. In *Proceedings of CoNLL*, pages 271–280.

Jitao Xu and François Yvon. 2021. Can you traducir this? machine translation for code-switched input. In *Proceedings of the Workshop on Computational Approaches to Linguistic Code-Switching (CALCS)*.

Zheng Xin Yong, Ruochen Zhang, Jessica Forde, Skyler Wang, Arjun Subramonian, Holy Lovenia, Samuel Cahyawijaya, Genta Winata, Lintang Sutawika, Jan Christian Blaise Cruz, et al. 2023. Prompting multilingual large language models to generate code-mixed texts: The case of South East Asian languages. In *Proceedings of the Workshop on Computational Approaches to Linguistic Code-Switching (CALCS)*.

Rabih Zbib, Erika Malchiodi, Jacob Devlin, David Stallard, Spyros Matsoukas, Richard Schwartz, John Makhoul, Omar F. Zaidan, and Chris CallisonBurch. 2012. Machine translation of Arabic dialects. In *Proceedings of NAACL*, pages 49–59.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. BERTScore: Evaluating text generation with BERT. In *Proceedings of ICLR*.

A Statistical Significance Tests

We present the statistical significance for the ASR and ST experiments in Tables 4 and 5.

		LEX$_{Dict}$	LEX$_{Rand}$	EC$_{Rand}$	EC$_{SPF}$	ML$_{Rand}$
	WER	62.4	61.7	62.3	62.8	62.7
LEX$_{Dict}$	62.4					
LEX$_{Rand}$	61.7	0.009*				
EC$_{Rand}$	62.3	0.719	0.017*			
EC$_{SPF}$	62.8	0.124	<0.001*	0.016*		
ML$_{Rand}$	62.7	0.197	<0.001*	0.032*	0.719	
ML$_{SPF}$	62.5	0.764	0.003*	0.407	0.142	0.254

		LEX$_{Dict}$	LEX$_{Rand}$	LEX$_{Pred}$	EC$_{Rand}$	EC$_{SPF}$	ML$_{Rand}$	ML$_{SPF}$
	WER	33.1	32.5	32.5	32.6	32.9	33.1	33.0
LEX$_{Dict}$	33.1							
LEX$_{Rand}$	32.5	0.006*						
LEX$_{Pred}$	32.5	0.009*	0.928					
EC$_{Rand}$	32.6	0.057	0.503	0.535				
EC$_{SPF}$	32.9	0.337	0.095	0.114	0.238			
ML$_{Rand}$	33.1	0.865	0.003*	0.004*	0.018*	0.201		
ML$_{SPF}$	33.0	0.667	0.020*	0.026*	0.075	0.516	0.465	
BT	32.4	0.003*	0.589	0.509	0.177	0.018*	<0.001*	0.003*

Table 4: Statistical significance between ASR models in the zero-shot (upper) and non-zero-shot (lower) settings calculated on WER achieved on ArzEn-ST test set CSW sentences. We present the p-values and mark p-values < 0.05 with $*$, where the null hypothesis can be rejected. We include the WER figures for easier readability and comparison.

		LEX$_{Dict}$	LEX$_{Rand}$	LEX$_{Pred}$	EC$_{Rand}$	EC$_{SPF}$	ML$_{Rand}$	ML$_{SPF}$	BT
	chrF++	41.2	42.0	43.0	41.7	42.2	41.8	41.8	43.3
LEX$_{Dict}$	41.2								
LEX$_{Rand}$	42.0	0.0010*							
LEX$_{Pred}$	43.0	0.0010*	0.0010*						
EC$_{Rand}$	41.7	0.0100*	0.0490*	0.0010*					
EC$_{SPF}$	42.2	0.0010*	0.1598	0.0010*	0.0070*				
ML$_{Rand}$	41.8	0.0040*	0.0939	0.0010*	0.2687	0.0170*			
ML$_{SPF}$	41.8	0.0020*	0.1489	0.0010*	0.1798	0.0420*	0.2647		
BT	43.3	0.0010*	0.0010*	0.0410*	0.0010*	0.0010*	0.0010*	0.0010*	
ST_BL$_{All}$	41.6	0.0300*	0.0250*	0.0010*	0.2038	0.0040*	0.1518	0.0949	0.0010*

Table 5: Statistical significance between ST models in the non-zero-shot setting calculated on the chrF++ scores achieved on ArzEn-ST test set CSW sentences. We present the p-values and mark p-values < 0.05 with $*$, where the null hypothesis can be rejected. We include the chrF++ scores for easier readability and comparison.

B Perplexity in MT Setup

We report PPL in MT setups by training transformer-based LMs using Fairseq. The models are optimized with Adam (Kingma and Ba, 2014) using $\beta 1 = 0.9$, $\beta 2 = 0.98$. We set the dropout to 0.1 and the learning rate to 0.0005. We report perplexity for the non-zero-shot settings as follows: LEX$_{Dict}$ (163.2), LEX$_{Rand}$ (156.1), LEX$_{Pred}$ (148.6), EC$_{Rand}$ (146.0), EC$_{SPF}$ (150.5), ML$_{Rand}$ (147.0), ML$_{SPF}$ (150.5), and BT (143.2).