UM-DFKI Maltese Speech Translation

Aiden Williams* Kurt Abela* Rishu Kumar◇ Martin Bär*

Hannah Billinghurst* Kurt Micallef* Ahnaf Mozib Samin*

Andrea De Marco* Lonneke van der Plas† Claudia Borg*

*University of Malta, ◇DFKI, †IDIAP

aiden.williams.19@um.edu.mt, kurt.abela@um.edu.mt,
rishu.Kumar@dfki.de,martin.bar.22@um.edu.mt,
hannah.billinghurst.22@um.edu.mt, kurt.micallef@um.edu.mt,
ahnaf.samin.22@um.edu.mt, andrea.demarco@um.edu.mt,
lonneke.vanderplas@idiap.ch, claudia.borg@um.edu.mt

Abstract

For the 2023 IWSLT (Agarwal et al., 2023) Maltese Speech Translation Task, UM-DFKI jointly presents a cascade solution which achieves 0.6 BLEU. While this is the first time that a Maltese speech translation task has been released by IWSLT, this paper explores previous solutions for other speech translation tasks, focusing primarily on low-resource scenarios. Moreover, we present our method of fine-tuning XLS-R models for Maltese ASR using a collection of multi-lingual speech corpora as well as the fine-tuning of the mBART model for Maltese to English machine translation.

1 Introduction

Speech Translation (ST), or speech-to-text translation, involves converting speech in a source language into written text in a target language. With the rise of deep learning, steep progress has been made in this field and many other areas that fall under the Natural Language Processing (NLP) umbrella (Khurana et al., 2023; Qiu et al., 2020). However, development for low-resource languages has continued to present difficulties and obstacles due to a variety of factors, including the lack of sufficient training data, language experts and other resources (Magueresse et al., 2020; Hedderich et al., 2021).

The International Workshop on Spoken Language Translation (IWSLT) shared task is an annual competition that aims to foster research in the field of speech translation. With its low-resource track, it also contributes to advanced research for speech translation in low-resource scenarios. In this paper, we present our submission to the low resource track: a pipeline system for English-Maltese speech-to-text translation.

We begin by discussing the state of the art in speech translation and describe the two main approaches, cascade and end-to-end. Afterwards, we briefly summarise the challenges posed by low-resource languages and possible mitigation strategies. We then describe our system, a pipeline approach containing an internal Automatic Speech Recognition (ASR) component and the outward facing Machine Translation (MT) component. The ASR component can use one of five fine-tuned XLS-R (Babu et al., 2021) models, whereas the MT stage always uses an mBART-50 model.

2 Literature Review

The following literature review aims to provide an overview of previous IWSLT ST submissions, with a particular focus on low-resource scenarios. The review is divided into two sections; where the first explores the general approaches and challenges associated with low-resource ST, and the second section discusses previous approaches to low-resource ST as applied to IWSLT.

2.1 Previous IWSLT Approaches for Low-Resource Languages

The IWSLT (Anastasopoulos et al., 2022) set the task in 2022 to attempt to solve "the problem of developing speech transcription and translation tools for under-resourced languages". This problem involved translating Tamasheq into English and Tunisian Arabic into French. Three different teams attempted to solve the problem of the Tamasheq-English ST; Taltech publish an encoder-decoder ST model that used a pre-trained XLS-R that they fine-tuned on unlabelled Tamasheq as the encoder and mBART-50 as the decoder, GMU used the Fairseq s2t extension with its transformer archi-

Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023), pages 433–441
July 13-14, 2023

tecture in which they fine-tuned the pre-trained XLS-R 300M encoder on French and Arabic and then trained the whole model on the provided data from the task; finally, ON-TRAC had a primary submission which used a pre-trained Wav2Vec 2.0 base model trained on Tamasheq and a contrastive model which was comprised of a partial Wav2Vec 2.0 model, a linear layer used for down projecting the output of the Wav2Vec and a transformer decoder. All three submissions decided to focus on using large pre-trained models when approaching the task, which is the approach taken for our models as well. The results from the submissions showed that using powerful speech feature extractors such as Wav2Vec 2.0 and massive multilingual decoders such as mBART-50 does not stop low-resource ST from being a major challenge. Of the three submissions, training self-supervised models on the target data and producing artificial supervision seemed to be the most effective approach to solving the problem (Zanon Boito et al., 2022).

Previous, well-performing systems submitted to the IWLST offline and low-resource speech translation tracks made use of various methods to improve the performance of their cascade system. For the ASR component, many submissions used a combination of transformer and conformer models (Zhang et al., 2022; Li et al., 2022; Nguyen et al., 2021) or fine-tuned existing models (Zhang and Ao, 2022; Zanon Boito et al., 2022; Denisov et al., 2021). They managed to increase ASR performance by voice activity detection for segmentation (Zhang et al., 2022; Ding and Tao, 2021), training the ASR on synthetic data with added punctuation, noise-filtering and domain-specific fine-tuning (Zhang and Ao, 2022; Li et al., 2022) or adding an intermediate model that cleans the ASR output in terms of casing and punctuation (Nguyen et al., 2021). The MT components were mostly transformer-based (Zhang et al., 2022; Nguyen et al., 2021; Bahar et al., 2021) or fine-tuned on pre-existing models (Zhang and Ao, 2022). Additional methods used to improve MT performance were multi-task learning (Denisov et al., 2021), back-translation (Ding and Tao, 2021; Zhang et al., 2022; Zhang and Ao, 2022), domain adaption (Nguyen et al., 2021; Zhang et al., 2022), knowledge distillation (Zhang et al., 2022), making the MT component robust by training it on noisy ASR output data (Nguyen et al., 2021; Zhang et al., 2022; Zhang and Ao, 2022), re-ranking and de-noising techniques (Ding and Tao, 2021). Bahar et al. (2021) trained their ASR and MT components jointly by passing the ASR output to the MT component as a probability vector instead of a one-hot vector to attenuate error propagation and avoid information loss of the otherwise purely textual output.

2.2 Wav2Vec 2.0 XLS-R For Maltese ASR

One of the latest developments for the Wav2Vec system is the introduction of multilingual pre-training. Due to the robust architectural design of Wav2Vec 2.0, models are able to learn cross-lingual speech representations (XLSR) while pre-training on massive amounts of data. This is put in practice with the XLSR models, which are pre-trained on up to 53 different languages from the Mozilla Commonvoice (v. Nov. 2019), BABEL (Gales et al., 2014) and Multilingual LibriSpeech (Pratap et al., 2020) speech corpora, with the largest model, pre-trained on a total of 56 thousand hours of speech data (Conneau et al., 2021). To test out the XLSR approach, several Wav2Vec BASE models are pre-trained either monolingually or multilingually. Monolingual models follow the process previously taken, i.e. they are pre-trained using the same language on which they are fine-tuned. This process is changed slightly for multilingual models, which are pre-trained on ten languages; then at the fine-tuning stage, a model is fine-tuned for each language. The experiment also included the pre-training of the Wav2Vec LARGE XLSR-53 model, which was pre-trained on the entire dataset of unannotated data, and just like the multilingual models, a separate model is then created for each language it was evaluated on during fine-tuning. The performance of different approaches, evaluated on four languages; Assamese, Tagalog, Swahili, and Georgian, is shown in Table 1. In these languages, the multilingual models, XLSR even more so, outperform the monolingual model.

The work on the XLSR approach continues in (Babu et al., 2021) with the release of the XLS-R model, which saw an increase in both the size of the unannotated data and the languages included. BABEL, Multilingual LibriSpeech, and CommonVoice (v. Dec. 2020) are joined by the VoxPopoli (Wang et al., 2021) and VoxLingua107 (Valk and Alumäe, 2021) corpora for a total of 436 thousand unannotated hours.

Table 1: XLSR Wav2Vec 2.0 performance on low-resource settings when evaluated using WER. Assamese (AS), Tagalog (TL), Swahili (SW), and Georgian (KA) are the languages presented.

Language	AS	TL	SW	KA
Annotated Data (h)	55	76	30	46
XLSR-10	44.9	37.3	35.5	-
XLSR-53	44.1	33.2	36.5	31.1
XLS-R (0.3B)	42.9	33.2	24.3	28.0
XLS-R (1B)	40.4	30.6	21.2	25.1
XLS-R (2B)	39.0	29.3	21.0	24.3

2.3 mBART For Maltese to English Translation

According to (Liu et al., 2020), using mBART-25 as the pre-trained model has been shown to improve translations over a randomly initialized baseline in low/medium resource language. mBART-25 is a transformer model trained on the BART (Lewis et al., 2019) objective. It is trained on 25 different languages. mBART-25 was later extended to include 25 more languages and was called mBART-50 (Tang et al., 2020). However, neither model included Maltese - in fact, translation experiments on Maltese are very limited. In our experiments, in Section 3.2, we checked whether these performance gains expand to the Maltese language, and this claim appears to hold.

3 Methodology

For this task, we decided to use a cascade system where the ASR and MT components were trained separately but evaluated jointly. In this section, a detailed description of both components is given. First, the training data is described, followed by the pre-processing steps applied to said data. Next, the models are introduced, and lastly training, the training procedure is outlined.

3.1 Automatic Speech Recognition

The ASR component in this submission continues the previous work done in (Williams, 2022), and so the same annotated dataset consisting of 50 hours of Maltese speech is used for this task. We opted not to use data released for this task for two reasons. First was the additional annotation work that was required, mainly segmentation, for which we experienced issues attempting to do in a timely manner. Secondly, this submission includes models fine-tuned with non-Maltese data. Making use of the dataset in (Williams, 2022) as a base has made comparisons with previous experiments possible.

As described in Table 2, the Maltese speech corpus is made up of several segments from two main Maltese speech corpora, MASRI (Hernandez Mena et al., 2020), CommonVoice (CV) (Ardila et al., 2020) and an annotated set from publicly available parliamentary sittings. Previous research in ASR for Maltese has used English speech data with varying degrees of success (Mena et al., 2021). However, when applied in fine-tuning an XLS-R model, the effect was detrimental. To further observe the effect non-Maltese data would have on the translation task, we used three other subsets from the CommonVoice speech corpus. Selecting 50 hours of validated each from the Italian, French and Arabic sets.

Individually these speech corpora each amount to 50 hours, from which four models are trained. One with just the Maltese data and the other three trained on the extra language combined with the Maltese set. A fifth model is also trained with all the data included. Further combinations were not tried due to time concerns.

Table 2: Each corpus is listed along with its total length, sample count and average sample length.

Dataset	Length (h,m)	Samples	Average Length (s)
HEADSET	6, 40	4979	4.81
MEP	1, 20	656	7.11
Tube	13, 20	8954	5.34
MERLIN	19, 4	9720	6.14
Parlament	2, 30	1672	5.35
CV Validated	4, 57	3790	12.68
CV Other	5, 4	3833	4.71
CV French	50	-	-
CV Italian	50	-	-
CV Arabic	50	-	-
Validation	2, 32	1912	4.89
Test MASRI	1	668	5.39
Test CV	0, 54	670	4.74

The XLS-R model comes in three pre-trained variants; the small model with 300 million parameters, the medium model with a billion parameters and the large model with two billion parameters. Size on disk scales with size with the small model being roughly 1GB in size and the large model being roughly 8GB. All three of them have been pre-trained on roughly 500 thousand hours of un-

Table 3: ASR Models and the data used for fine-tuning.

Model	Corpora used
MT Only	All Maltese corpora
MT+All	All corpora presented
MT+AR	All Maltese corpora + Arabic subset
MT+FR	All Maltese corpora + French subset
MT+IT	All Maltese corpora + Italian subset

labelled, multilingual speech. Previous research (Williams, 2022), has shown that both the small and large models fare well when fine-tuned for the downstream Maltese ASR task. With this in mind, the small 300M XLS-R variant model was chosen for this task. The main reason was due to its smaller size, a larger batch size could be used which expedited the fine-tuning process, while the performance loss was expected to be minimal.

This submission follows the same training procedure as outlined in (Williams, 2022). Where the procedure was conducted utilising the Huggingface Trainer object with the following hyper-parameters. Each model is trained for 30 epochs, using the AdamW criterion with a starting learning rate of $3e-4$. To stabilise the training process, the first 500 training steps were used as warm-up steps. Gradient accumulation was also used to effectively quadruple the batch size. The batch size was dependent on the training set used, where due to some differences in sample lengths, different batch sizes had to be used. We fine-tune 5 XLS-R 300m models as presented in Table 3.

3.2 Machine Translation

The dataset used to train the machine translation systems comes from publicly available sources. The original data sources include datasets from Arab-Acquis (Habash et al., 2017), the European Vaccination Portal[1],the Publications Office of the EU on the medical domain[2], the European Medicines Agency[3], the COVID-19 ANTIBIOTIC dataset[4], the COVID-19 EC-EUROPA dataset[5], the COVID-19 EU press corner V2 dataset[6], the COVID-19 EUROPARL v2 dataset[7], the Digital Corpus of the European Parliament (Hajlaoui et al., 2014), the DGT-Acquis (Steinberger et al., 2014), ELRC[8], the Tatoeba corpus[9], OPUS (Tiedemann, 2012), EUIPO - Trade mark Guidelines[10], Malta Government Gazette[11], MaCoCu (Bañón et al., 2022), as well as data extracted from the Laws of Malta[12].

The different datasets were compiled into a single one. The total number of parallel sentences amounts to 3,671,287. The development and test set was kept the exact same as the OPUS dataset (Tiedemann, 2012), which amount to 2000 sentences each, and the rest of the data was placed in the training set, which amounts to 3,667,287 parallel sentences.

Before training the system, the data has to be further pre-processed. Firstly, a BPE tokenizer is trained on the training set only. The MosesDecoder[13] package is used to pre-process the dataset, by normalising punctuation and training a true case on the training set and applying it to the whole dataset. In the case of Maltese data, a tokenizer specifically designed for Maltese was used because the regular English tokenizer does not tokenize everything correctly. For this, the tokenizer from MLRS[14] was used, which utilises regular expressions to tokenize linguistic expressions that are specific to Maltese, such as certain prefixes and articles. The dataset is then encoded using the previously trained BPE encoder.

The machine translation model is built and trained using Fairseq (Ott et al., 2019). Fairseq is a library that allows for easy implementation of a machine translation system through CLI commands, meaning minimal code is needed to create a fully working machine translation system.

For this system, a pre-trained mBART-50 model (Tang et al., 2020) was used and fine-tuned on our

[1] https://bit.ly/3dLbGX9
[2] https://bit.ly/3R2G5OH
[3] https://bit.ly/3QWIjPM
[4] https://bit.ly/3pBCg7u
[5] https://bit.ly/3AcjIzR
[6] https://bit.ly/3wmCyTD
[7] https://bit.ly/3wl3brZ
[8] https://www.lr-coordination.eu/node/2
[9] https://bit.ly/3cejoIU
[10] https://bit.ly/3AB01Tr
[11] https://bit.ly/3QDXm1a
[12] https://legislation.mt/
[13] https://www.statmt.org/moses/
[14] https://mlrs.research.um.edu.mt/

data. An mBART-25 (Liu et al., 2020) model, as well as a randomly initialised baseline Transformer model, were also experimented with, however after training a system using a subset of the dataset, it was apparent that the mBART-50 model outperforms them both. Due to limited resource constraints, only one MT model was trained on the full dataset.

The maximum number of steps was set out to be 1,000,000, yet the validation was performed every 10,000 steps with a patience value of 10. This means that if the BLEU score on the validation set does not improve after ten validation steps, then the model stops training. After multiple experiments using a smaller subset of the dataset, it was seen that increasing max-tokens tended to result in higher overall performance. However, due to resource constraints, the maximum number of tokens per batch was set to 1024. The learning rate is set to $1e^{-3}$, but the initial learning rate is smaller at $1e^{-7}$ and increases using an inverse square root learning rate scheduler to linearly increase the rate after 10,000 steps. For inference, a beam size of five is used to generate predictions.

The total number of updates using mBART-50 was 990,000, with an early stop since the validation didn't improve in the last 10 validation epochs. This amounts to exactly three full epochs on the whole training set.

3.3 Completed Pipeline

To create a speech-to-text translation system, a Huggingface pipeline is set up to accept an audio file that is passed to the ASR system. The test set provided for this task is a single file of over one hour. Due to its size, the file needs to be segmented for inference and evaluation due to its size. The XLS-R model automatically returns a timestamp for each output word. These timestamps are used to create segments that align with the segments file provided with the test set.

This means that the ASR component returns a list of text strings. Each segment is an item in the list of strings. Each string is passed to the MT system. Before passing through the MT component, the resultant strings are pre-processed. The aforementioned MosesDecoder package is used to transform the strings using the same rules that have been applied to the MT training data. This means that the strings have their punctuation normalised, then true cased and finally tokenized. The processed strings are then passed to the mBART model to be inferred and the BPE model to encode the inputs. The beam size is set to five. The resulting tokens are then detokenized and saved.

4 Evaluation and Results

Table 4 contains the official results for our submission for the Maltese → English spoken language translation track. While we observed better scores during training and validation, our models struggled with the official test set. In this section, we note our few observations and qualitative analysis of results to highlight the errors.

The test set proved to be difficult for both the ASR and MT systems to get right due to the type of language used as well as the speed of the speech in general. Table 5 shows the reference transcription of the beginning of the file, accompanied by the MT Only and MT+All ASR transcription, and lastly, the machine translation of the mt-50 model. The monolingually fine-tuned MT Only model was our primary submission from the five submitted ASR models, with BLEU scores of 0.6.

The mt-50 output is relatively similar to the reference sentence, except for a few minor errors, including the misspelling of the name "Mark". However, this should still be a good sentence to input into the machine translation system. In stark contrast to the MT+All system outputs.

The main issue here is that this system does not output Maltese characters and completely omits them, which presents an issue for the downstream translation task since the meaning of the word is lost in these cases.

Machine translation also had similar issues. The training set contained data coming from legal texts, so the data is very formal, making it very difficult to evaluate since the input text is very informal and unlike the legal text data seen.

Unfortunately, most of this is unrelated to what

Table 4: Official Results for our models for Maltese → English SLT task

Submission Name	BLEU Score
MT Only	0.6
MT+All	0.7
MT+AR	0.4
MT+FR	0.3
MT+IT	0.4

Table 5: Reference transcription sample from the IWSLT 2023 test set along with the MT Only and MT+All automatic transcription and the machine translation of the MT Only output.

Reference	*merħba' għal- podcast ieħor din id- darba ma bniedem kemxejn polemikuż mhux għax jien għandi wisq xi ngħid però Mark Camilleri huwa il- mexxejj kemxejn kontroversjali tal- kunsill nazzjonali tal- ktieb*
MT Only	merba' l- pot kast ieħor din id- darba ma bniedem kemxejn polemikuż mhux għax jien għandi wisq xi ngħid però mar Camilleri huwa il- mexxejj kemxejn kontroversjali tal- kunsill nazzjonali tal- ktieb
MT+All	meba l Pold cast ieor din id-darba ma bniedem kemmxejn polemiku mhux gax jien Gandi wisq xi ngid per mar kamileri huwai - mexxejk emxejh kontroversjali tal- kunsill nazzjonali tal-ktieb
Translation MT Only	four of the other potential this time does not work very slightly at all , but not at all , the same time , it is the slightly cross- sectoral leader of the national when the book is also of humane

was actually said. Looking into the translations deeper, one can see the reasoning behind certain translations. For example, the dataset does not contain a lot of conversational data, so general greetings like "merħba" may not be present. This case is represented by the translation of the token "merba", which was translated to "four". Here the token "merba" (welcome) was mistaken for "erba" (four). Other mistakes include those that are phonetically plausible but grammatically incorrect output, such as the transcription for "podcast" which was transcribed as "pot kast". Certain expressions like "din id-darba" were correctly translated to "this time", however rarer words such as "polemikuż" and "kontroversjali", both of which have the same meaning as "controversial", seemed to not appear in the translation.

Continuing the trend observed in (Williams, 2022), the use of additional languages when fine-tuning an XLS-R model proved to be detrimental towards the final output. As observed in Section 4, some models trained with additional data lost the ability to transcribe Maltese-specific alphabetic characters. So far, the character-to-sound pair was always made with the source language in mind. For example, the French 'Ç' is transformed into the 'C' character, which itself is only present in the Maltese alphabet when English words are loaned and used directly. It's important to note that code-switching to English is very common in Maltese speech. Future work should explore these character-to-sound pairs.

5 Conclusion and Future Work

This paper showcased the results of a speech-to-text translation system in the direction of Maltese to English. A cascade system is chosen, where ASR and MT models are pipelined together.

The automatic speech recognition system chosen is based on XLS-R and is fine-tuned on data from different languages. The best-performing model was the XLS-R 300M model fine-tuned on 50 hours of Maltese speech. The machine translation system chosen is based on mBART-50, and it was fine-tuned on parallel Maltese - English data. Aside from fine-tuning, no modifications were made to the pre-trained models.

For future work, we have various potential avenues for improvement. For machine translation, since mBART-50 was not pre-trained on Maltese data, extending the vocabulary to include Maltese-specific tokens would improve the representation and potentially the downstream performance as well. Moreover, our approach solely relied on parallel data and did not investigate techniques which leverage monolingual data, such as back-translation. Monolingual corpora, such as Korpus Malti v4 (Micallef et al., 2022), not only provide significantly more data but also have more diversity in terms of domains. Apart from this, it might be beneficial to perform more quality checks on the parallel dataset since some portions of the publicly available datasets are automatically crawled and, in some cases, contain noise.

Regarding ASR improvement, other systems, such as Whisper and, most recently Meta's Massively Multilingual Speech (MMS) project should be tried and evaluated. The research made in multilingual fine-tuning needs to be more focused. One idea we can explore is the transliteration of foreign alphabetic characters into Maltese characters, e.g. 'h' in English would be transliterated as 'ħ'. It is also the case that no language model is used to correct the ASR output mistakes; this is currently our next milestone.

Acknowledgements

We acknowledge LT-Bridge Project (GA 952194). Rishu Kumar was supported financially by the EMLCT[15] programme during this entire work.

References

Milind Agarwal, Sweta Agrawal, Antonios Anastasopoulos, Ondřej Bojar, Claudia Borg, Marine Carpuat, Roldano Cattoni, Mauro Cettolo, Mingda Chen, William Chen, Khalid Choukri, Alexandra Chronopoulou, Anna Currey, Thierry Declerck, Qianqian Dong, Yannick Estève, Kevin Duh, Marcello Federico, Souhir Gahbiche, Barry Haddow, Benjamin Hsu, Phu Mon Htut, Hirofumi Inaguma, Dávid Javorský, John Judge, Yasumasa Kano, Tom Ko, Rishu Kumar, Pengwei Li, Xutail Ma, Prashant Mathur, Evgeny Matusov, Paul McNamee, John P. McCrae, Kenton Murray, Maria Nadejde, Satoshi Nakamura, Matteo Negri, Ha Nguyen, Jan Niehues, Xing Niu, Atul Ojha Kr., John E. Ortega, Proyag Pal, Juan Pino, Lonneke van der Plas, Peter Polák, Elijah Rippeth, Elizabeth Salesky, Jiatong Shi, Matthias Sperber, Sebastian Stüker, Katsuhito Sudoh, Yun Tang, Brian Thompson, Kevin Tran, Marco Turchi, Alex Waibel, Mingxuan Wang, Shinji Watanabe, and Rodolfo Zevallos. 2023. Findings of the IWSLT 2023 Evaluation Campaign. In *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023)*. Association for Computational Linguistics.

Antonios Anastasopoulos, Loïc Barrault, Luisa Bentivogli, Marcely Zanon Boito, Ondřej Bojar, Roldano Cattoni, Anna Currey, Georgiana Dinu, Kevin Duh, Maha Elbayad, Clara Emmanuel, Yannick Estève, Marcello Federico, Christian Federmann, Souhir Gahbiche, Hongyu Gong, Roman Grundkiewicz, Barry Haddow, Benjamin Hsu, Dávid Javorský, Věra Kloudová, Surafel Lakew, Xutai Ma, Prashant Mathur, Paul McNamee, Kenton Murray, Maria Nădejde, Satoshi Nakamura, Matteo Negri, Jan Niehues, Xing Niu, John Ortega, Juan Pino, Elizabeth Salesky, Jiatong Shi, Matthias Sperber, Sebastian Stüker, Katsuhito Sudoh, Marco Turchi, Yogesh Virkar, Alexander Waibel, Changhan Wang, and Shinji Watanabe. 2022. Findings of the IWSLT 2022 Evaluation Campaign. In *Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT 2022)*, pages 98–157, Dublin, Ireland (in-person and online). Association for Computational Linguistics.

[15] https://mundus-web.coli.uni-saarland.de/

Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber. 2020. Common voice: A massively-multilingual speech corpus. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4218–4222, Marseille, France. European Language Resources Association.

Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick von Platen, Yatharth Saraf, Juan Pino, Alexei Baevski, Alexis Conneau, and Michael Auli. 2021. XLS-R: self-supervised cross-lingual speech representation learning at scale. *CoRR*, abs/2111.09296.

Parnia Bahar, Patrick Wilken, Mattia A. Di Gangi, and Evgeny Matusov. 2021. Without Further Ado: Direct and Simultaneous Speech Translation by AppTek in 2021. In *Proceedings of the 18th International Conference on Spoken Language Translation (IWSLT 2021)*, pages 52–63, Bangkok, Thailand (online). Association for Computational Linguistics.

Marta Bañón, Miquel Esplà-Gomis, Mikel L. Forcada, Cristian García-Romero, Taja Kuzman, Nikola Ljubešić, Rik van Noord, Leopoldo Pla Sempere, Gema Ramírez-Sánchez, Peter Rupnik, Vít Suchomel, Antonio Toral, Tobias van der Werff, and Jaume Zaragoza. 2022. MaCoCu: Massive collection and curation of monolingual and bilingual data: focus on under-resourced languages. In *Proceedings of the 23rd Annual Conference of the European Association for Machine Translation*, pages 303–304, Ghent, Belgium. European Association for Machine Translation.

Alexis Conneau, Alexei Baevski, Ronan Collobert, Abdelrahman Mohamed, and Michael Auli. 2021. Unsupervised Cross-Lingual Representation Learning for Speech Recognition. In *Proc. Interspeech 2021*, pages 2426–2430.

Pavel Denisov, Manuel Mager, and Ngoc Thang Vu. 2021. IMS' Systems for the IWSLT 2021 Low-Resource Speech Translation Task. In *Proceedings of the 18th International Conference on Spoken Language Translation (IWSLT 2021)*, pages 175–181, Bangkok, Thailand (online). Association for Computational Linguistics.

Liang Ding and Dacheng Tao. 2021. The USYD-JD Speech Translation System for IWSLT2021. In *Proceedings of the 18th International Conference on Spoken Language Translation (IWSLT 2021)*, pages 182–191, Bangkok, Thailand (online). Association for Computational Linguistics.

Mark JF Gales, Kate M Knill, Anton Ragni, and Shakti P Rath. 2014. Speech recognition and keyword spotting for low-resource languages: Babel project research at cued. In *Fourth International workshop on spoken language technologies for under-resourced languages (SLTU-2014)*, pages 16–23. International Speech Communication Association (ISCA).

Nizar Habash, Nasser Zalmout, Dima Taji, Hieu Hoang, and Maverick Alzate. 2017. A parallel corpus for evaluating machine translation between arabic and european languages. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 235–241.

Najeh Hajlaoui, David Kolovratnik, Jaakko Väyrynen, Ralf Steinberger, and Daniel Varga. 2014. Dcep-digital corpus of the european parliament. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*.

Michael A. Hedderich, Lukas Lange, Heike Adel, Jannik Strötgen, and Dietrich Klakow. 2021. A Survey on Recent Approaches for Natural Language Processing in Low-Resource Scenarios. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2545–2568, Online. Association for Computational Linguistics.

Carlos Daniel Hernandez Mena, Albert Gatt, Andrea DeMarco, Claudia Borg, Lonneke van der Plas, Amanda Muscat, and Ian Padovani. 2020. MASRI-HEADSET: A Maltese corpus for speech recognition. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 6381–6388, Marseille, France. European Language Resources Association.

Diksha Khurana, Aditya Koli, Kiran Khatter, and Sukhdev Singh. 2023. Natural language processing: State of the art, current trends and challenges. *Multimedia tools and applications*, 82(3):3713–3744.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Yinglu Li, Minghan Wang, Jiaxin Guo, Xiaosong Qiao, Yuxia Wang, Daimeng Wei, Chang Su, Yimeng Chen, Min Zhang, Shimin Tao, Hao Yang, and Ying Qin. 2022. The HW-TSC's Offline Speech Translation System for IWSLT 2022 Evaluation. In *Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT 2022)*, pages 239–246, Dublin, Ireland (in-person and online). Association for Computational Linguistics.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation.

Alexandre Magueresse, Vincent Carles, and Evan Heetderks. 2020. Low-resource languages: A review of past work and future challenges. *arXiv preprint arXiv:2006.07264*.

Carlos Daniel Hernandez Mena, Andrea DeMarco, Claudia Borg, Lonneke van der Plas, and Albert Gatt. 2021. Data augmentation for speech recognition in maltese: A low-resource perspective. *CoRR*, abs/2111.07793.

Kurt Micallef, Albert Gatt, Marc Tanti, Lonneke van der Plas, and Claudia Borg. 2022. Pre-training data quality and quantity for a low-resource language: New corpus and BERT models for Maltese. In *Proceedings of the Third Workshop on Deep Learning for Low-Resource Natural Language Processing*, pages 90–101, Hybrid. Association for Computational Linguistics.

Tuan Nam Nguyen, Thai Son Nguyen, Christian Huber, Ngoc-Quan Pham, Thanh-Le Ha, Felix Schneider, and Sebastian Stüker. 2021. KIT's IWSLT 2021 Offline Speech Translation System. In *Proceedings of the 18th International Conference on Spoken Language Translation (IWSLT 2021)*, pages 125–130, Bangkok, Thailand (online). Association for Computational Linguistics.

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-HLT 2019: Demonstrations*.

Vineel Pratap, Qiantong Xu, Anuroop Sriram, Gabriel Synnaeve, and Ronan Collobert. 2020. MLS: A Large-Scale Multilingual Dataset for Speech Research. In *Proc. Interspeech 2020*, pages 2757–2761.

Xipeng Qiu, Tianxiang Sun, Yige Xu, Yunfan Shao, Ning Dai, and Xuanjing Huang. 2020. Pre-trained models for natural language processing: A survey. *Science China Technological Sciences*, 63(10):1872–1897.

Ralf Steinberger, Mohamed Ebrahim, Alexandros Poulis, Manuel Carrasco-Benitez, Patrick Schlüter, Marek Przybyszewski, and Signe Gilbro. 2014. An overview of the european union's highly multilingual parallel corpora. *Language resources and evaluation*, 48(4):679–707.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2020. Multilingual translation with extensible multilingual pretraining and finetuning. *arXiv preprint arXiv:2008.00401*.

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

Jörgen Valk and Tanel Alumäe. 2021. Voxlingua107: A dataset for spoken language recognition. In *2021 IEEE Spoken Language Technology Workshop (SLT)*, pages 652–658.

Changhan Wang, Morgane Riviere, Ann Lee, Anne Wu, Chaitanya Talnikar, Daniel Haziza, Mary Williamson, Juan Pino, and Emmanuel Dupoux. 2021. VoxPopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 993–1003, Online. Association for Computational Linguistics.

Aiden Williams. 2022. The applicability of Wav2Vec 2.0 for low-resource Maltese ASR. B.S. thesis, University of Malta.

Marcely Zanon Boito, John Ortega, Hugo Riguidel, Antoine Laurent, Loïc Barrault, Fethi Bougares, Firas Chaabani, Ha Nguyen, Florentin Barbier, Souhir Gahbiche, and Yannick Estève. 2022. ON-TRAC Consortium Systems for the IWSLT 2022 Dialect and Low-resource Speech Translation Tasks. In *Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT 2022)*, pages 308–318, Dublin, Ireland (in-person and online). Association for Computational Linguistics.

Weitai Zhang, Zhongyi Ye, Haitao Tang, Xiaoxi Li, Xinyuan Zhou, Jing Yang, Jianwei Cui, Pan Deng, Mohan Shi, Yifan Song, Dan Liu, Junhua Liu, and Lirong Dai. 2022. The USTC-NELSLIP Offline Speech Translation Systems for IWSLT 2022. In *Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT 2022)*, pages 198–207, Dublin, Ireland (in-person and online). Association for Computational Linguistics.

Ziqiang Zhang and Junyi Ao. 2022. The YiTrans Speech Translation System for IWSLT 2022 Offline Shared Task. In *Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT 2022)*, pages 158–168, Dublin, Ireland (in-person and online). Association for Computational Linguistics.