Rosetta Balcanica: Deriving a “Gold Standard” Neural Machine Translation (NMT) Parallel Dataset for Western Balkan Languages

Edmon Begoli
Oak Ridge National Laboratory
begolie@ornl.gov

Maria Mahbub
Oak Ridge National Laboratory
mahbubm@ornl.gov

Sudarshan Srinivasan
Oak Ridge National Laboratory
srinivasans@ornl.gov

Abstract

The *Rosetta Balcanica* is an ongoing effort to expand the resources for low-resource western Balkan languages. This effort focuses on discovering and using accurately translated, officially mapped, and curated parallel language resources and their preparation and use as neural machine translation (NMT) datasets. Some of the guiding principles, practices, and methods employed by *Rosetta Balcanica* are generalizable and could apply to other low-resource language resource expansion efforts. With this goal in mind, we present our rationale and approach to discovering and using meticulously translated and officially curated low-resource language resources and our use of these resources to develop a parallel “gold standard” translation training resource. Secondly, we describe our specific methodology for NMT dataset development from these resources and its publication to a widely-used and accessible repository for natural language processing (*Hugging Face Hub*). Finally, we discuss the trade-offs and limitations of our current approach and the roadmap for future development and expansion of the current *Rosetta Balcanica* language resource.

[1]

1 Introduction

Many underresourced languages are spoken by ethnic groups residing in regions affected by economic, social, or other crises. These circumstances frequently necessitate the involvement of international institutions focused on economic assistance, the promotion of human rights, the advancement of democratic structures and processes, humanitarian aid, peacekeeping, and regional economic and political stabilization. As part of these activities, these institutions issue reports and studies that support their missions and goals.

The documents produced by international organizations operating in regions affected by these factors require precise translation to: a) preserve the original meaning of the master communique, and b) accurately convey that meaning across the languages of the regional groups. Such documents constitute a “golden standard” for any downstream cross-lingual computational linguistics tasks, particularly machine translation training. For a data set to fit such a standard, it must map across parallel data sets at the phrase and paragraph levels with high accuracy, be produced by professional translators, and be issued by an authoritative publication body.

Neural machine translation (NMT) (Bahdanau et al., 2014) is a state-of-the-art approach to machine translation that uses an end-to-end encoder-decoder architecture with attention mechanisms (Vaswani et al., 2017), to translate source language sentences into target language sentences. Attention-based models are particularly well-suited for translation tasks as they support the training of models that can maintain attention on complex relationships and dependencies between two related sequences. For example, attention-based models excel at learning relationships between words and phrases in two languages, even when the languages do not have a one-to-one mapping.

Accurate and complete sentence-to-sentence alignment between the source and target languages in a training dataset is particularly important for effectively training neural machine translation models because it helps the NMT model form accurate patterns of attention. Therefore, a “golden”

[1] The United States Government retains and the publisher, by accepting the article for publication, acknowledges that the United States Government retains a non- exclusive, paid-up, irrevocable, world-wide license to publish or reproduce the published form of this manuscript, or allow others to do so, for United States Government purposes. The Department of Energy will provide public access to these results of federally sponsored research in accordance with the DOE Public Access Plan (http://energy.gov/downloads/doe-public-access-plan).

Proceedings of the The Seventh Workshop on Technologies for Machine Translation of Low-Resource Languages (LoResMT 2024), pages 186–192
August 15, 2024 ©2024 Association for Computational Linguistics

training set in low-resource languages is of great value for effective low-resource NMT model development, as the training content is meticulously translated and aligned at the phrase, sentence, and paragraph levels.

This effort focuses on the Western Balkan languages, a region in South Central Europe comprising countries formed by the breakup of former Yugoslavia and Albania. The languages spoken in this region include dialects of Southern Slavic languages (Serbian, Croatian, Bosnian, Macedonian, Slovenian, etc.), Albanian, and others (Friedman, 2011). These languages and resources were selected due to our familiarity with the region's cultural, historical, and ethnic circumstances, and our fluency in several of these language groups (including Croatian, Bosnian, Serbian, and to some degree, Macedonian).

Furthermore, given the significant similarity between Croatian, Serbian, and Bosnian languages (Ljubesic et al., 2007), some of these parallel resources, with translations from English, Shqip, and Macedonian to either Serbian, Croatian, or Bosnian, can potentially be used to cross-map the translations between all three languages. This significantly boosts these three language resources and their use for other multilingual tasks (Pourdamghani and Knight, 2017).

The development of high-quality, meticulously translated parallel datasets for low-resource languages such as those in the Western Balkans is crucial for advancing neural machine translation. In this paper, we outline the methodology, challenges, and future directions for creating these datasets, which serve as foundational resources for enhancing multilingual and cross-lingual computational linguistics.

2 Background

Many of the languages of the Western Balkans are spoken by small national groups or ethnic minorities and are considered low-resource languages (Tyers and Alperen, 2010; Mati et al., 2021; Kunchukuttan and Bhattacharyya, 2021). Albanian, or Shqip, is spoken by 7.5 million people worldwide, with approximately 4.5 million speakers in the region. Although an Indo-European language, it is unique and is not closely related to any other Indo-European language. Macedonian (Masson and Davies, 2016), an east-south Slavic language, is spoken by approximately two million people.

Furthermore, the region has been severely affected by a series of regional wars, which, in addition to causing significant population losses, have been followed by economic depopulation (Lukic et al., 2012), a trend that continues today (Lutz and Gailey, 2020). It is estimated that 4.4 million people have emigrated from the region (World Bank Group and WIIW, 2018, p.42).

2.1 Related Work

A number of parallel language resources for the languages of the Western Balkans. Many of these resources are the results of volunteer or officially sanctioned projects where news sites and individual works of well-known literature or other resources were collected into a corpus.

META-SHARE is the web site that curates and maintains most of these of parallel language resources for Western Balkan languages. Macedonian-Croatian Parallel Corpus (**mk-hr_pcorp**) (Cebović and Tadić, 2016) is a parallel corpus consisting of fictional synchronic prose texts with over 500,000 tokens in each language and 39,735 aligned sentences. South-East European Parallel Corpus (Tyers and Alperen, 2010) is a corpus of South-East European languages derived from The South-East European Times website. The website is a collection of regional news that was sponsored by the United States European Command, dedicated to coverage of Southeast Europe (it ended publication in March 2015). The South-East European Parallel Corpus includes, among others, resources in Albanian (41,741,782 tokens), Macedonian (37,623,521 tokens) and Croatian (34,968,453 tokens).

A parallel corpus for the tourism domain focusing on English and Croatian languages was created by leveraging automatic data crawlers to collect parallel data from the web (Toral et al., 2017). The dataset was used to train machine translation (MT) systems for the tourism domain, aiming to optimize translation tasks relevant to this specific field.

Parallel Global Voices (PGV) (Prokopidis et al., 2016) is a parallel language dataset, derived from the Global Voices multilingual group of websites, where volunteers publish and translate news stories in more than 40 languages.

Parallel Data, Tools and Interfaces in OPUS is a growing resource of freely accessible parallel corpora. It also provides tools for processing parallel and monolingual data, as well as several interfaces to search the data, making it a unique resource for various research activities (Tiedemann, 2012). Albanian from Taoteba, a collection of volunteer contributed sentences and translations that has 2,532 Albanian, 77,988 Macedonian, 5,301 Croatian, 45,786 Serbian and 619 Bosnian sentences. A recent survey of resources and methods for Serbian Language was presented in (Marovac et al., 2023).

However, most of the available resources such as Common Crawl[2] or Internet Archive[3] that are commonly used by the research community to build parallel datasets (Banón et al., 2022; El-Kishky et al., 2019) suffer from limitations. They often comprise language structures that are either inadequately translated or lack meticulous alignment across languages. In contrast, the resources utilized in *Rosetta Balcanica* stand out for their rigorous curation process. They have been curated from texts translated by official sources and meticulously mapped to multiple languages, ensuring high-quality and accurate representations to serve as resources for a wide range of applications.

2.2 Data Source

The Organization for Security and Cooperation in Europe (OSCE) (Galbreath, 2007) is the world's largest intergovernmental security-oriented organization. Its mission includes conflict prevention, arms control, the promotion of democratic values and processes, and the protection and promotion of human rights, among other objectives. The OSCE has several missions in the Western Balkans, specifically in Albania, Bosnia and Herzegovina, Kosovo, and Serbia. As part of its mission, the OSCE publishes and curates multilingual official documents relevant to its activities in the region. Their website [4] hosts an extensive resource library of reports and white papers, with a significant section dedicated to the Western Balkans.

For its translators, the OSCE requires a university degree in interpretation or a related field, with a requisite perfect command of the languages of interest and professional proficiency in English (preferably as a mother tongue). Pertinent to the region and languages of interest, the OSCE publishes reports, studies, and white papers on human rights, democratic elections, hate crime statistics, and other topics within the OSCE's scope and mission. Given the scarcity of language resources and the high quality of multilingual translations, we have included all these resources.

3 Approach

Our overall approach to the development of Rosetta Balcanica was to identify the specific, officially translated multilanguage resources amenable to raw language processing and then convert them into a general and neutrally formatted parallel language resource. From there, we chose to post-format this neutral resource into a specific format (e.g. Hugging Face Hub format), with the intention of making it accessible to a large group of NLP practitioners.

We started the collection process by attempting to automate the search, retrieval, and preprocessing of downloaded resources. We encountered a number of errors related to standard document formatting (page numbers, header/footer dividers) that made automation complex, uncertain and error prone, so we delayed the entire automation process in favor of manual collection and preprocessing to make initial progress in corpus development (first stage/release in a roadmap (see Figure 4). We intend to develop and proceed with a more robust automation starting with the second stage.

3.1 Selection and Preparation

We selected OSCE resources by language, starting with Shqip, assuming that there will be at least one translation (English) for the Shqip resources. In most cases, we found parallel translations between English, Shqip and the combination of Macedonian and Serbian, and in some instances, such as regional publications or annual OSCE reports, the documents were translated into multiple languages.

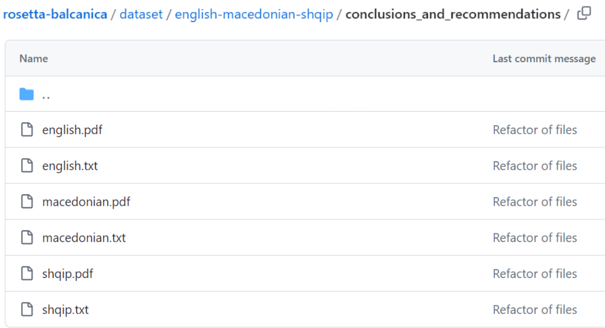
3.2 Parallel Dataset Creation

Following best practices for the development of parallel datasets, we create a folder for each document and then convert each PDF into a plain text representation. In particular, all document folders reside in a *language combination* directory named corresponding to the language of the documents it contains. For example, if a directory contains folders for documents that involve the three languages English, Macedonian, and Shqip, our language

[2] https://commoncrawl.org/
[3] https://archive.org/
[4] https://www.osce.org/resources/documents

combination directory would be named *english-macedonian-shqip*. This folder structure allows us to add more documents to an already existing directory or to add new language combinations with new documents in the future.

We applied minimal formatting to remove page numbers and other non-linguistic features such as header artifacts. Then, we made sure that language files are aligned with each other at the paragraph and sentence level. The resulting structure is a folder named after each document, with each file entry representing a translation of the same content as seen in figure 1.



rosetta-balcanica / dataset / english-macedonian-shqip / conclusions_and_recommendations /

Name	Last commit message
..	
english.pdf	Refactor of files
english.txt	Refactor of files
macedonian.pdf	Refactor of files
macedonian.txt	Refactor of files
shqip.pdf	Refactor of files
shqip.txt	Refactor of files

Figure 1: A folder structure and sample content for parallel language document.

This shows the structure for the content extracted from the document titled "Conclusions and Recommendations" from the 5th Annual South East Media Conference[5]. Table 1 shows some random samples extracted from this document for the three languages.

3.3 Validation

Given that the sources covered under Rosetta Balcanica are officially translated materials, we only performed spot validation to ensure alignment between the texts. Specifically, we validated the translations by randomly sampling parallel sentences and verifying their translations in Google Translate.

3.4 *Hugging Face Hub*

Hugging Face is the largest online NLP community engaged around the use and sharing of state-of-the-art models. A part of this community is a Hugging Face Hub where NLP researchers can upload, share, and access data sets and models. At the time of the writing of this manuscript, there are over 900 datasets and more than 25 metrics available. Data on Hugging Face hub follows a *Dataset* convention. *Datasets* is a library to easily access and share data sets and evaluation metrics for Natural Language Processing (NLP), computer vision, and audio tasks. Once part of the hub, the data sets can be loaded into a Hugging Face pipeline with a single line of code and used to train neural machine translation models (NMT).

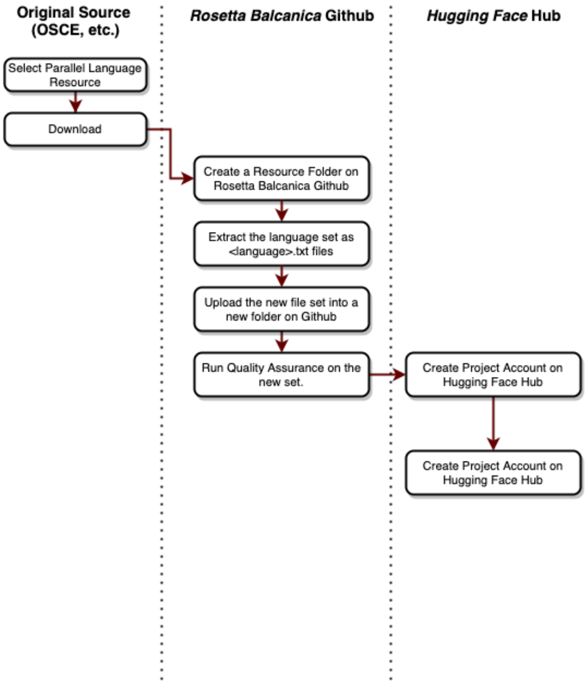

Original Source (OSCE, etc.)
Rosetta Balcanica Github
Hugging Face Hub
Select Parallel Language Resource
Download
Create a Resource Folder on Rosetta Balcanica Github
Extract the language set as <language>.txt files
Upload the new file set into a new folder on Github
Run Quality Assurance on the new set.
Create Project Account on Hugging Face Hub
Create Project Account on Hugging Face Hub

Figure 2: Rosetta Balcanica dataset development and registration workflow (Release 1).

We use a script that automatically scrapes all the document folders within a language combination directory and then extracts and aggregates all the sentences in each language text file to create a single text file corresponding to each language. This script then parses these language files to create a Hugging Face compliant dataset. Specifically, we create a temporary folder, which contains language-pair folders for each language. For example, the language combination directory *english-macedonian-shqip* containing documents from the three languages would result in two language pair folders *en-ma* and *en-sh*. We followed the convention of using the first two letters of the language to name the folder. Each language-pair directory contains 4 files corresponding to training and testing files for each language in the language pair following Hugging Face's machine translation dataset convention. These are then zipped up into

[5] https://www.osce.org/representative-on-freedom-of-media/384549

English	Macedonian	Shqip
OSCE SOUTH EAST EUROPE MEDIA CONFERENCE CONCLUSIONS AND RECOMMENDATIONS	КОНФЕРЕНЦИЈА НА ОБСЕ ЗА МЕДИУМИ ВО ЈУГОИСТОЧНА ЕВРОПА ЗАКЛУЧОЦИ И ПРЕПОРАКИ	KONFERENCA E OSBE-SË PËR MEDIA NË EVROPËN JUGLINDORE KONKLUZIONE DHE REKOMANDIME
Trade unions need to be recognized as legitimate representatives of journalists.	Синдикатите треба да се признаат како легитимни претставници на новинарите.	Sindikatat duhet të njihen si përfaqësues legjitim të gazetarëve.
Encourage investigative pieces and journalism also in PSM in line with best practices of the sector.	Поттикнување истражувачки стории и новинарство и во ЈМС, согласно најдобрите практики во секторот	Inkurajim i reportazheve dhe gazetarisë hulumtuese edhe në transmetuesit publikë në pajtim me praktikat më të mira.

Table 1: Sample parallel language document content from the OSCE 5th Annual South East Media Conference

rosetta_balcanica.tar.gz which is saved at the root of the repository.

All these files are currently hosted on Github at our repository[6]. The data sets are made accessible via the compressed zip file through the Hugging Face Hub by directly accessing the Github repository. Finally, we upload the dataset to HuggingFace Hub [7] for easy access.

4 Dataset Statistics

The first release of the *Rosetta Balcanica* focuses on the parallel documents in Shqip and Macedonian. For this release, we used only Shqip resources available through OSCE.

4.1 Corpus Statistics

Counts	English	Macedonian	Shqip
# of Sentences	8567	8567	8567
# of Tokens	137620	148459	168202
# of Unique Tokens	7839	14565	18911

Table 2: Parallel English-Macedonian-Shqip Corpus Statistics

4.2 Topics Represented

As expected, there is an obvious and deliberate bias in the topics presented in the OSCE corpus figure 3. These topics reflect the subjects of the publications and reports that are within the scope of OSCE's mission in the Western Balkan countries - human rights watch (and related violations), elections, conflicts and incidents, and overall sociocultural and economic development. These topics are expected to dominate the corpus for the first two to three releases of the dataset while we are focusing on OSCE and Hague Tribunal original sources. We plan to diversify the topics over time, but, overall, we are deliberately choosing a narrowly focused dataset that is of a highest quality, and the tradeoff of that approach will be lower topic diversity.

5 Challenges and Limitations

Maceodonian and Serbian entries are written in Cyrillic script. To compare the entries between Serbian, and, for example, Croatian and Bosnian, which are otherwise very similar languages, an additional transliteration step is required. OSCE is a trusted dataset so we did not perform transliteration of all the data. We checked 1% of randomly selected samples and found those accurately translated.

6 Future Work

The scope of the work presented in this encompasses the work on the OSCE dataset, which is topically narrow, and primarily performing manual preprocessing to minimize "debugging" time. Future work will encompass a) inclusion of other, similarly structured parallel language datasets, and b) automation of dataset retrieval and preprocessing.

Once we process OSCE resources for all Western languages, we will work to process Hague tribunal documents which are, like OSCE, profes-

[6] https://github.com/ebegoli/rosetta-balcanica

[7] https://huggingface.co/datasets/sudarshan85/rosetta_balcanica

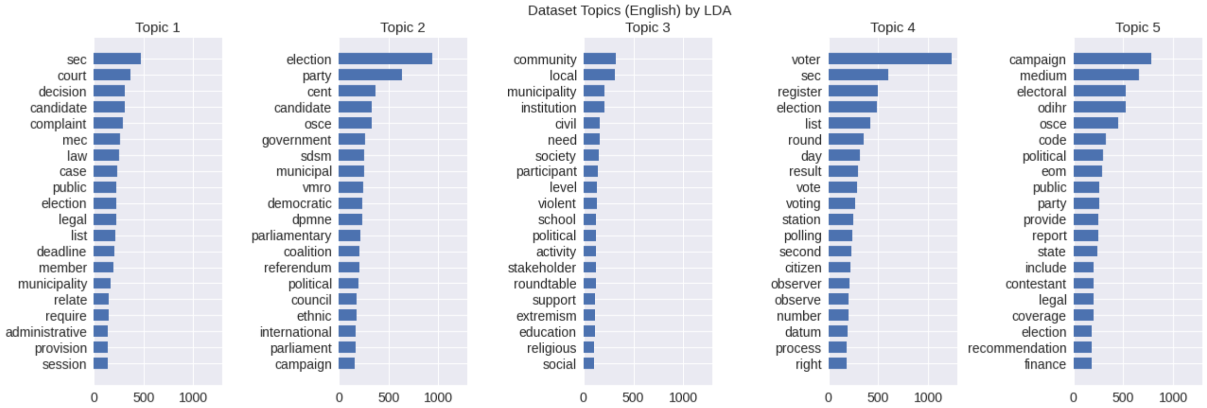


Figure 3: A representation of topics in the corpus.

sionally translated and available in most Western Balkan languages. After that, we will focus on the general multi-lingual resources from official media sources. We will focus on the latter mostly to diversify the topics and balance out representation.

Rosetta Balcanica is on an ongoing project, and the development of new resources continues. Although it is an ongoing process, we have identified releases and milestones (Figure 4) in a roadmap that maps to the inclusion of specific language resources and the dataset development tools (e.g. resource retrieval and pre-processing automation, quality assurance, dataset registration, etc.).

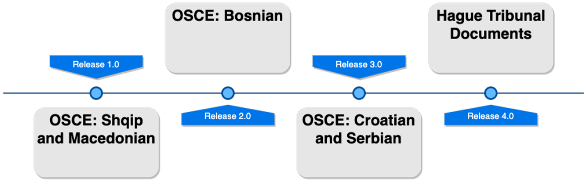


Figure 4: Roadmap of releases for 2021-2023 for Rosetta Balcanica.

We mentioned earlier that we attempted to automate the dataset collection/retrieval and pre-processing step. We found that the complexities and error-proneness of automation attempts, at that time, were slowing us down in the process of dataset development more than they were helping us. It is evident, though, that this process needs to be automated in order to be scalable and easily re-usable, and it is on our roadmap to automate the process. In fact, we are in the process of improving the automation process, and we expect that the automation scripts will be available by the time this paper is published (mid-2022).

7 Conclusion

There are three takeaways from the *Rosetta Balcanica* effort of that are, we presume, of interest to the natural language processing community:

1. Similar organizations, with a perhaps different mission, are likely sources of the similar "golden set" materials that can be used for the development of similar parallel datasets for low-resource languages. We recommend exploring similar resources for other low-resource languages. These could be UNESCO, UNICEF, United Nations, and other international organizations.

2. The workflow presented in this paper is a practice that we recommend as well. The approach we have taken, where we curate a single, raw corpus in parallel languages and then use it to create a training library or modality-specific dataset (Hugging Face Hub) is an approach that makes the dataset readily available to a broad community that uses state-of-the-art NLP methods (neural machine translation, etc.). This approach also scales well because the original, raw source can be used for the development of other library and modality-specific datasets.

3. While we found automated retrieval and preparation of data sources to be challenging and error-prone, we still intend to pursue this route in the future, and we encourage other similar efforts to attempt the same.

8 Acknowledgements

This manuscript has been in part co-authored by UT-Battelle, LLC under Contract No. DE-AC05-

00OR22725 with the U.S. Department of Energy.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Marta Banón, Miquel Espla-Gomis, Mikel L Forcada, Cristian García-Romero, Taja Kuzman, Nikola Ljubešić, Rik van Noord, Leopoldo Pla Sempere, Gema Ramírez-Sánchez, Peter Rupnik, et al. 2022. Macocu: Massive collection and curation of monolingual and bilingual data: focus on under-resourced languages. In *23rd Annual Conference of the European Association for Machine Translation, EAMT 2022*, pages 303–304. European Association for Machine Translation.

Ines Cebović and Marko Tadić. 2016. Building the macedonian-croatian parallel corpus. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4241–4244.

Ahmed El-Kishky, Vishrav Chaudhary, Francisco Guzmán, and Philipp Koehn. 2019. Ccaligned: A massive collection of cross-lingual web-document pairs. *arXiv preprint arXiv:1911.06154*.

Victor A Friedman. 2011. The balkan languages and balkan linguistics. *Annual Review of Anthropology*, 40:275–291.

David J Galbreath. 2007. *The organization for security and co-operation in Europe (OSCE)*. Routledge.

Anoop Kunchukuttan and Pushpak Bhattacharyya. 2021. Machine translation and transliteration involving related and low-resource languages.

Nikola Ljubesic, Nives Mikelic, and Damir Boras. 2007. Language indentification: How to distinguish similar languages? In *2007 29th International Conference on Information Technology Interfaces*, pages 541–546. IEEE.

Tamara Lukic, Rastislav Stojsavljevic, Branislav Durdev, Imre Nagy, and Bojan Dercan. 2012. Depopulation in the western balkan countries. *European Journal of Geography*, 3(2):6–23.

Wolfgang Lutz and Nicholas Gailey. 2020. Depopulation as a policy challenge in the context of global demographic trends.

Ulfeta A Marovac, Aldina R Avdić, and Nikola Lj Milošević. 2023. A survey of resources and methods for natural language processing of serbian language. *arXiv preprint arXiv:2304.05468*.

Olivier Masson and Anna Morpurgo Davies. 2016. Macedonian language. In *Oxford Research Encyclopedia of Classics*.

Diellza Nagavci Mati, Mentor Hamiti, Arsim Susuri, Besnik Selimi, and Jaumin Ajdari. 2021. Building dictionaries for low resource languages: Challenges of unsupervised learning. *Annals of Emerging Technologies in Computing (AETiC)*, 5(3):52–58.

Nima Pourdamghani and Kevin Knight. 2017. Deciphering related languages. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2513–2518.

Prokopis Prokopidis, Vassilis Papavassiliou, and Stelios Piperidis. 2016. Parallel global voices: a collection of multilingual corpora with citizen media stories. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 900–905.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in opus. In *Lrec*, volume 2012, pages 2214–2218. Citeseer.

Antonio Toral, Miquel Esplá-Gomis, Filip Klubička, Nikola Ljubešić, Vassilis Papavassiliou, Prokopis Prokopidis, Raphael Rubino, and Andy Way. 2017. Crawl and crowd to bring machine translation to under-resourced languages. *Language resources and evaluation*, 51:1019–1051.

Francis M Tyers and Murat Serdar Alperen. 2010. South-east european times: A parallel corpus of balkan languages. In *Proceedings of the LREC workshop on exploitation of multilingual resources and tools for Central and (South-) Eastern European Languages*, pages 49–53. Citeseer.

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

World Bank Group and WIIW. 2018. Western balkans labor market trends 2018.