Scholar Commons
USF University of South Florida

University of South Florida
Scholar Commons

Graduate Theses and Dissertations

Graduate School

June 2021

Turkic Interlingua: A Case Study of Machine Translation in Low-resource Languages

Jamshidbek Mirzakhalov
University of South Florida

Follow this and additional works at: https://scholarcommons.usf.edu/etd

 Part of the Computer Sciences Commons

Scholar Commons Citation

Mirzakhalov, Jamshidbek, "Turkic Interlingua: A Case Study of Machine Translation in Low-resource Languages" (2021). *Graduate Theses and Dissertations.*
https://scholarcommons.usf.edu/etd/8829

This Thesis is brought to you for free and open access by the Graduate School at Scholar Commons. It has been accepted for inclusion in Graduate Theses and Dissertations by an authorized administrator of Scholar Commons. For more information, please contact scholarcommons@usf.edu.

Turkic Interlingua: A Case Study of Machine Translation in Low-resource Languages

by

Jamshidbek Mirzakhalov

A thesis submitted in partial fulfillment
of the requirements for the degree of
Master of Science in Computer Science
Department of Computer Science and Engineering
College of Engineering
University of South Florida

Co-Major Professor: Sriram Chellappan, Ph.D.
Co-Major Professor: John Licato, Ph.D.
Marvin Andujar, Ph.D.

Date of Approval:
May 21, 2021

Keywords: multilingual, NLP, neural networks, participatory research

Copyright © 2021, Jamshidbek Mirzakhalov

Acknowledgments

I would like to thank the entire Turkic Interlingua (TIL) community including the researchers, annotators, translators, industry partners and every individual who has contributed in making this project possible. I also thank Dr. Sriram Chellappan and Dr. John Licato for their valuable advice throughout the project and beyond.

Table of Contents

List of Tables

List of Figures

Abstract

Machine Translation (MT) has the potential to bridge the gap between the developed world and the marginalized communities by making information more accessible in real-time. While there are over 7000 spoken languages in the world, only about a hundred have access to high-quality MT systems and even fewer enjoy the benefits of more advanced language technologies. Unfortunately, resource scarcity and the lack of digital infrastructure are only some of the many challenges associated with globalizing NLP. Many large-scale multilingual studies and datasets often get little to no feedback from native speakers or linguistic experts of the languages involved, leading to serious problems of data quality and potential biases. In this thesis, we present a case study of participatory research in 22 Turkic languages involving native speakers, language technologists, researchers, linguists, commercial entities, and more. Through this thesis, we compile and release the largest public corpus for MT in Turkic languages along with 26 bilingual baseline models. We outline the curation and release of public datasets, development of machine translation technologies, and their deployment in real-world scenarios. In addition, we discuss the lessons learned through this case study, its applications, and limitations, as well as implications for future projects.

Chapter 1: Introduction

Language technologies, particularly machine translation (MT), have the potential to break down communication barriers between societies and make information ubiquitously accessible for all. Recent advances in deep learning have drastically increased the potential in building systems that can be used in practice. However, as we scale our systems in terms of size and data resources, we also risk marginalizing many vulnerable populations that lack sufficient data or computational resources, consequently depriving them of the benefits of technological innovations in the space (Nekoto et al., 2020). In order to address these problems and illustrate the fall-back of state-of-the-art methods in MT in low-resource languages (Joshi et al., 2019), we make the first attempt of studying the practical performance of currently prominent MT methods in a very challenging case of the Turkic language family, consisting of a large number of extremely low-resource and morphologically-rich languages. In this thesis, we present a large-scale case study of MT through the practices of participatory research. First, we describe the current state of Natural Language Processing (NLP) including the recent advances in multilingual embedding representations and their performances in a variety of linguistic tasks. We then discuss the case of Turkic languages, their typological features and socio-economical shortcomings that play a role in hindering the development of language technologies such as MT. Second, we describe the concept of “participatory research”, its history of inception and practical

applications in dealing with low-resource NLP. Third, we introduce a community, Turkic Interlingua (TIL), based on the principles of participatory research with a mission of building language technologies and conducting academic research in the context of Turkic languages. Lastly, we describe at length the past and current projects at TIL that highlights the different aspects of low-resource NLP and the way participatory research can overcome these challenges. As a part of these projects, our main contributions are: a) a large-scale audit of 205 language-specific corpora in multilingual datasets, b) the compilation and release of the largest public corpus for MT in Turkic languages, c) training and release of 26 bilingual baselines and d) the practical feasibility demonstration of participatory research in low-resource NLP settings. We summarize our findings, lessons learned from this large, international and interdisciplinary community effort, and discuss future directions and collaborations.

1.1 State of Natural Language Processing

NLP is a subfield of computer science that combines concepts from computer science, linguistics, and artificial intelligence. It is sometimes used interchangeably with computational linguistics and it may be referred to as such in this work. NLP is concerned with the capabilities of the system to contextualize and understand natural language data which then can be used to extract useful information, summarize and classify documents, translate between languages etc. Early concepts of NLP were first formulated by a British computer scientist Alan Turing in his famous *Turing Test* (Turing, 1950) which is a test an artificial bot (chatbot in this case) would have to pass to prove its intelligence. Initial approaches to NLP (Symbolic NLP) involved a program following a set of complex hand-written rules and commands to parse through data and this

required extensive amounts of labor and domain knowledge (Searle & others, 1980). In the late 90s, as the available computation power increased, the rise of Statistical NLP took place. The gradual displacement of handwritten rules were followed by the statistical approaches underpinned by corpus linguistics. Statistical Machine Translation (SMT), for example, was a paradigm where translations were generated based on the complex parameters learned through the analysis of large bilingual text corpora (P. Brown et al., 1988). Since the 2010s, representation learning and deep neural networks have become mainstream after outperforming all previous approaches in almost all NLP tasks (T. B. Brown et al., 2020; Devlin et al., 2019; Mikolov et al., 2013; Pennington et al., 2014; Radford et al., 2019; Xue et al., 2021). This paradigm, also known as Neural NLP, is partly due to the increased computational power and the abundance of unlabeled and labeled corpora. This "data hungry" nature of the deep neural networks have led the researchers focusing on the low-resource scenarios where concepts such as pre-trained language models, cross-lingual transfer learning, multi-task learning have yielded state-of-the-art results in many NLP tasks and benchmarks.

1.2 Turkic Languages and Machine Translation

Turkic languages consist of over 35 languages spoken natively across Europe and Asia by almost 200 million people. Of the languages, 20 are official languages of a state or a sub-region while the rest remain as minority languages. Modern Turkic languages are written in several scripts such as Latin, Cyrillic and Perso-Arabic and it is common to see the same language written in two or more scripts (Róna-Tas, 2015). Multi-script languages and their diverse orthographies make it challenging to convert between different orthographic systems which eventually poses problems during data

collection . The languages exhibit elaborate morphology, possess a relatively free word order, and are very similar in their structure and grammar. These typological features make the MT and, more broadly, NLP, very challenging for these languages (Tantuğ et al., 2008; Bender, 2011; Tsarfaty et al., 2013, 2020).

In a recent work exploring the diversity and inclusion of the NLP community, Joshi et al., 2020 proposes a resource taxonomy where they categorize languages into 6 different groups based on the available amount of labeled and unlabeled data. The taxonomy outlines the definitions of the six categories in Table 1.

Table 1: Resource taxonomy and the definition of categories

Class	Category name	Definition
0	The Left-Behinds	These languages have been and are still ignored in the aspect of language technologies. With exceptionally limited resources, it will be a monumentous, probably impossible effort to lift them up in the digital space. Unsupervised pre-training methods only make the ‘poor poorer’, since there is virtually no unlabeled data to use.
1	The Scraping-Bys	With some amount of unlabeled data, there is a possibility that they could be in a better position in the ‘race’ in a matter of years. However, this task will take a solid, organized movement that increases awareness about these languages, and also sparks a strong effort to collect labelled datasets for them, seeing as they have almost none.
2	The Hopefuls	With light at the end of the tunnel, these languages still fight on with their gasping breath. A small set of labeled datasets has been collected for these languages, meaning that there are researchers and language support communities which strive to keep them alive in the digital world. Promising NLP tools can be created for these languages a few years down the line.
3	The Rising Stars	Unsupervised pre-training has been an energy boost for these languages. With a strong web presence, there is a thriving cultural community online for them. However, they have been let down by insufficient efforts in labeled data collection. With the right steps, these languages can be very well off if they continue to ride the ‘pre-training’ wave.
4	The Underdogs	Powerful and capable, these languages pack serious amounts of resource ‘firepower’. They have a large amount of unlabeled data, comparable to those possessed by the winners, and are only challenged by a lesser amount of labeled data. With dedicated NLP communities conducting research on these languages, they have the potential to become winners and enjoy the fruits of ‘digital superiority’.

Table 1 (Continued)

5	The Winners	Running strong and fast, these languages have been in the lead for quite a while now, some longer than others. With a dominant online presence, there have been massive industrial and government investments in the development of resources and technologies for these languages. They are the quintessential rich-resource languages, reaping benefits from each state-of-the art NLP breakthrough.

As shown in Table 2, based on the taxonomy, we identify that out of 22 Turkic languages that our work has focused on, 19 of them belong in categories 0-2 and only 1 is in category 4. This analysis showcases the unfortunate state of the resources and research exploration in the content of Turkic languages.

Table 2: Number of L1 speakers for each Turkic language, its corresponding language code and the associated category from the resource taxonomy[1]

Language Name	**Codes**	**Speakers (L1)**	**Category**
Turkish	tr, tur	85.0M	The Underdogs (4)
Uzbek	uz, uzb	27.0M	The Rising Star (3)
Azerbaijani	az, aze	23.0M	The Scraping-Bys (1)
Kazakh	kk, kaz	13.2M	The Rising Star (3)
Uyghur	ug, uig	10.0M	The Scraping-Bys (1)
Turkmen	tk, tuk	6.70M	The Scraping-Bys (1)
Tatar	tt, tat	5.20M	The Scraping-Bys (1)
Kyrgyz	ky, kir	4.30M	The Scraping-Bys (1)
Bashkir	ba, bak	1.40M	The Scraping-Bys (1)
Chuvash	cv, chv	1.04M	The Scraping-Bys (1)
Karakalpak	kaa	583K	The Scraping-Bys (1)
Crimean Tatar	crh	540K	The Scraping-Bys (1)
Sakha (Yakut)	sah	450K	The Scraping-Bys (1)
Kumyk	kum	450K	**The Left-Behinds (0)**
Karachay-Balkar	krc	310K	The Scraping-Bys (1)
Tuvan	tyv	280K	The Scraping-Bys (1)

[1] https://www.ethnologue.com/

Table 2 (Continued)

Urum	uum	190K	**The Left-Behinds (0)**
Gagauz	gag	148K	The Scraping-Bys (1)
Salar	slr	70K	**The Left-Behinds (0)**
Altai	alt	56K	**The Left-Behinds (0)**
Khakas	kjh	43K	**The Left-Behinds (0)**
Shor	cjs	3K	**The Left-Behinds (0)**

In the last few years, there has been an increased interest in multilingual parallel datasets usually scoped to the language families or linguistic regions (Choudhary & Jha, 2011; Esplà-Gomis et al., 2019; Nekoto et al., 2020; Nomoto et al., 2018; Post et al., 2012)**.** Notably, (Tiedemann, 2020) published a large-scale multilingual parallel corpus covering over 500 languages out of which 14 belong to the Turkic family. While this is a significant improvement in the inclusivity of the languages, the limited size and domain of the test sets in this dataset are not sufficient to serve as a standalone benchmark for the Turkic languages. (Khusainov et al., 2020) compiled a Russian-Turkic parallel corpus covering 6 Turkic languages and reported bilingual baselines using several mainstream NMT approaches. However, the authors do not release the dataset, test sets or the models to the public which prevents their work from being used as a public benchmark. A public rule-based MT system, Apertium (Washington et al., 2019), supports multiple Turkic languages within their platform. Since the scalability of rule-based MT systems still remains an open question, direct comparison of our work would not be possible.

1.3 Participatory Research

Participatory design methods, whereby citizen scientists actively participate in the process of scientific knowledge production without having formal scientific training, can be dated back to 1970s (Schuler & Namioka, 1993; Spinuzzi, 2005) “when workers in Scandinavia worked together collaboratively to design the technologies that they would use in corporate settings” (Sloane et al., 2020). The idea of participatory design has been defined and demonstrated as a way of engaging in research with ethics and values by involving the communities who benefit from this line of work. Nekoto et al., 2020 uses the participatory research ideas to involve various stakeholders in the process of developing MT systems. They showcase their work in a case study of African languages and provide the first baselines for many of the commonly spoken languages in the continent. The authors describe their methodology as one where “*the agents in the MT process originate from the countries where the low-resourced languages are spoken*”. This facilitates the involvement of stakeholders such as native speakers, researchers, software developers, translators, content creators, curators and more. As a more general ideology, it can be paraphrased as “*a way to ensure that everyone who should be in the room is in the room*”. In the following sections, we attempt to answer the research question of whether or not participatory research can facilitate MT development and research in low-resource scenarios with a case study on Turkic languages. We employ the essential design principles of participatory research and carry out multiple research projects to evaluate these methods.

Chapter 2: Turkic Interlingua (TIL)

Inspired by the potential of participatory research in combating the deficiencies of NLP research and development in low-resource languages, we founded a community with principles grounded in the concepts of openness, collaboration, mentorship and local initiatives. It was first conceptualized in September of 2020 with the goal of creating a collaborative space for researchers working on NLP for Turkic languages. Soon the amount of interest from diverse groups of individuals and entities lead to the realization that there is a big need for an open space where collaboration from these diverse communities is not only useful but necessary. These groups include native speakers of these languages, software engineers, professional translators, commercial entities, university labs, researchers, and more, all of whom share the passion for their language and are ready to engage in the development of language technologies for their native tongues. The term “interlingua” refers to the concept in computer science that defines an artificial language devised for machine translation that serves as an intermediary representation between languages. “Turkic Interlingua " refers to a specific instance of that term in the case of Turkic languages. The official domain of the community has been set up at https://turkininterlingua.org, where the public can learn about the community and new members can join through active projects.

The official objective of the community has been defined as “a community of researchers, Machine Learning (ML) engineers, language enthusiasts and community

leaders whose mission is to develop language technologies (from spell checkers to translation models), collect diverse datasets, and explore linguistic phenomena through the lens of academic research in Turkic languages".

2.1 Demographics of TIL

Currently, the community involves over 100 members across two workspaces on Slack and Telegram. A recent survey shows (see Figure 1) the demographics of the participants based on their occupations. A large majority of the members of the community are undergraduate students with an interest in research and language technologies, while the next biggest group is the graduate students. Engagement and outreach for the community is usually through the social media channels on Twitter and Facebook.

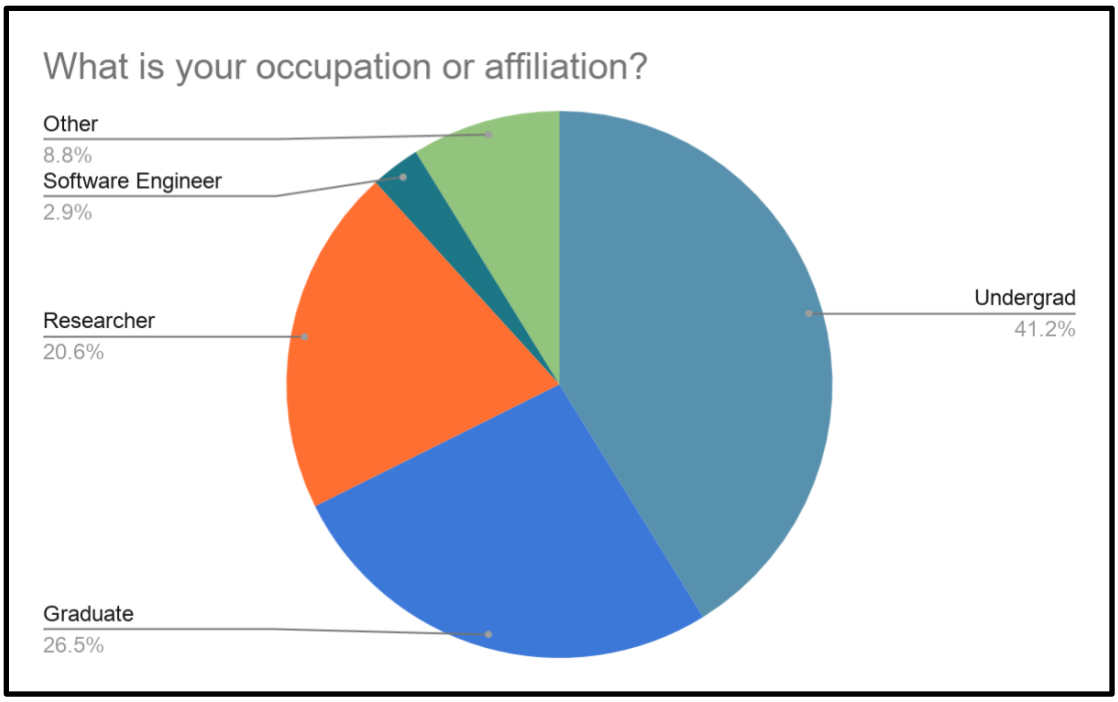

Figure 1. Demographics survey for TIL from March 2021

Chapter 3: Projects

On the basis of participatory research, the community has launched and completed several projects some of which have been published in peer-reviewed journals and conferences. Since the community is large and self-directed, it would be infeasible to cover all the active and past projects. Instead, we focus on the works that we have personally collaborated with and lead.

3.1 Quality at a Glance: An Audit of Web-Crawled Multilingual Datasets

The last few years have seen the proliferation of large, web-mined text datasets that are available in hundreds of languages. However, no study so far has examined the quality of these datasets in detail and whether they even contain the claimed material. Although it may sound surprising that the datasets may not contain what the authors claim they do, there are a few systemic reasons why it is in fact very likely. First, most of the researchers and research labs working on these corpora do not necessarily know or speak these languages, so the feasibility of manual audit is very little, to begin with. Second, the lack of native speakers or the high cost of hiring professionals usually leads to the over-reliance on automatic filtering and language identification tools which significantly underperform for most of the low-resource languages (Caswell et al., 2020). Lastly, before being released to the public, these datasets usually evaluate their data against the state-of-the-art multilingual benchmarks. However, these benchmarks only

include a few dozen high-resource languages which then results in overly confident high scores overall.

Turkic languages often face similar problems when it comes to finding high-quality datasets that can be used to train MT models. To examine the extent of this problem, we conduct the first large-scale analysis of public web-crawled multilingual datasets in a large collaborative study involving industry labs, grassroots NLP communities, and academic institutions. We manually audit random subsets from a total of 205 language-specific corpora within five main multilingual datasets: CCAligned (El-Kishky et al., 2020), ParaCrawl (Bañón et al., 2020; Esplà-Gomis et al., 2019), WikiMatrix (Schwenk et al., 2019), Oscar (Suárez et al., 2019), and mC4 (Xue et al., 2021). Results show an alarming trend: at least 15 corpora have no usable text and a large fraction of corpora have less than 50 percent in-language content. The heavy tail of this phenomenon is mainly carried by the low-resource languages. 7 Turkic languages were also evaluated as a part of the study both in the parallel and monolingual corpora.

3.1.1 Experimental Setup

To audit the quality of these datasets, we involved 51 volunteers from the NLP communities, TIL for example, which covered over 70 languages. For each language corpora in each dataset, we randomly sample 100 lines which can range from single words to short paragraphs. To measure the error rates, we create a simple error taxonomy as shown in Table 3. We then use this taxonomy to label the corpora.

Table 3: Description and examples from the annotation taxonomy used to quantify the quality of the datasets

Correct Codes	
CC: *Correct translation, natural sentence*	
en The Constitution of South Africa	nso Molaotheo wa Rephabliki ya Afrika Borwa
en Transforming your swimming pool into a pond	de Umbau Ihres Swimmingpools zum Teich
CB: *Correct translation, Boilerplate or low quality*	
en Reference number: 13634	ln Motango ya référence: 13634
en Latest Smell Stop Articles	fil Pinakabagong mga Artikulo Smell Stop
CS: *Correct translation, Short*	
en movies, dad	it cinema, papà
en Halloween - without me	ay Hallowen – janiw nayampejj
Error Codes	
X: *Incorrect translation, but both correct languages*	
en A map of the arrondissements of Paris	kg Paris kele mbanza ya kimfumu ya Fwalansa.
en Ask a question	tr Soru sor Kullanıma göre seçim
WL: *Source OR target wrong language, but both still linguistic content*	
en The ISO3 language code is zho	zza Táim eadra bracach mar bhionns na frogannaidhe.
en Der Werwolf — sprach der gute Mann,	de des Weswolfs, Genitiv sodann,
NL: *Not a language: at least one of source and target are not linguistic content*	
en EntryScan 4 _	tn TSA PM704 _
en organic peanut butter	ckb ������

3.1.2 Related Work

It is well established that web-mined texts tend to be noisy (Junczys-Dowmunt, 2018), especially in highly multilingual settings. Previous research found that web crawled data has much lower quality for the low-resource languages when used with the segment-level language identification tools (Caswell et al., 2020). This is likely the closest work to our study as they examine the quality of a large multilingual private corpus for low-resource languages. The authors also conduct experiments analysing the quality of in-language content in the OSCAR dataset.

3.1.3 Results and Discussion

The study analyzed more than 70 languages across 205 corpora, however in this thesis, we only focus on the results in the context of Turkic languages. Unfortunately yet unsurprisingly, these large 5 multilingual datasets include fewer than 10 Turkic

languages in total and many of which are almost unusable in research and especially production.

Table 4 shows results from the CCAligned dataset, which is a parallel dataset compiled from 68 Common Crawl snapshots. All sentences and documents in this dataset were aligned using automatic alignment methods such as language identification tools (Joulin et al., 2016, 2017) and LASER's cross-lingual embeddings (Artetxe & Schwenk, 2019)**.** The dataset was evaluated in an MT task on 6 European languages from the TED corpus (Qi et al., 2018) where it performed better than other contemporary corpora. The quality of data for Turkic languages (Azerbaijani, Kyrgyz, Kazakh and Turkish) is extremely low even in a relatively higher resource setting with Turkish.

Table 4: Audit results for a sample of 100 sentences from CCAligned for all Turkic languages present in the corpus

Language codes	**Language names**	**Error rate**	**Total number of sentences**
en-az_IR	English-Azerbaijani (Arabic script)	93.1%	158
en-ky_KG	English-Kyrgyz	55.88%	240657
en-kk_KZ	English-Kazakh	31.68%	689651
en-tr_TR	English-Turkish	55.00%	20282339

Table 5 shows the results from the WikiMatrix which is a parallel dataset with over 135 million parallel sentences from Wikipedia. This dataset was also processed using FastText LandID and LASER's cross-lingual embeddings. Even though only 2 out of 6 Turkic languages were evaluated from this dataset, the results are very alarming. English-Uighur corpus has an combined error rate of 84% and English-Kazakh corpus stands at 95% which makes these datasets unusable in any scenario.

Table 5: Audit results for a sample of 100 sentences from WikiMatrix for all Turkic languages present in the corpus

Language codes	Language names	Error rate	Total number of sentences
en-ug	English-Uighur	84.16%	22012
en-kk	English-Kazakh	95.00%	109074

Results from the OSCAR dataset which is a set of monolingual corpora are shown in Table 6. Data for OSCAR is also a compilation of Common Crawl snapshots and follows a similar filtering process. It uses the FastText LangID on a line-level and evaluates their datasets on POS tagging and dependency parsing tasks. Results from OSCAR are more optimistic, at least for the Turkic languages.

Table 6: Audit results for a sample of 100 sentences from OSCAR for all Turkic languages present in the corpus

Language codes	Language names	Error rate	Total number of sentences
tyv	Tuvinian	3.85%	26
uz	Uzbek	2.00%	34244
kk	Kazakh	0.00%	2719851

3.1.4 Conclusion

This collaborative work has, for the first time, evaluated the quality of multilingual corpora on such a large scale. Importantly, it has made it clear that the inclusion of low-resource languages in large multilingual datasets come at the expense of their quality, which makes several of the language-specific corpora unsuitable for research and especially production. Furthermore, it has also shed more light on the availability of resources for many of the Turkic languages which are absent in all of the datasets audited through this work. Future work will focus on the exploration of participatory

research in compiling datasets in low-resource settings that are scalable, ethical and of high-quality.

3.2 TIL Corpus

In the age of big data and deep neural networks, having access to large amounts of high-quality data is essential for building robust systems and models. As Section 4.1 highlights, there is a crisis of data quality which affects the low-resource languages the most and creates a skewed picture of progress as a field when not analyzed further. In the case of Turkic languages, fewer than 10 languages are included in the largest five multilingual datasets and even then their quality makes them extremely difficult and dangerous to use. Taking Machine Translation (MT) as the initial goal, we demonstrate the feasibility of participatory research in compiling and developing multilingual corpora that is of high-quality. In an effort to put together the largest public corpus for MT in Turkic languages, we organize an effort consisting of 40+ community members and several commercial entities. The result is a parallel corpus with over 75 million sentences covering 22 Turkic languages as well as human translated evaluation set in 8 Turkic languages. To our knowledge, this is the largest and most comprehensive public corpus ever released for MT research in Turkic languages. This section will cover the details about the data collection, quality control and limitations.

3.2.1 Data Collection

First, we collect and compile all public datasets containing parallel sentences between any Turkic language as well as English and Russian. English and Russian were chosen as pivot languages because of their prevalence and the data availability. We selected 3 most comprehensive multilingual datasets: The Tatoeba corpus

(Tiedemann, 2020), JW300 corpus (Agić & Vulić, 2019, p. 300), and GoURMET (Birch et al., 2019). The Tatoeba corpus consisted of 58 language pairs of interest, JW300 and GoURMET had 59 and 2 respectively. It is important to note that the Tatoeba corpus in itself is a compilation of many public resources for MT and may have overlapping data with JW300 and GoURMET. We deduplicate the data before using it in our experiments.

Second, we compile a list of sources[2][3] of data which include unreleased datasets from commercial entities, websites with multilingual support and other private resources. This was a large effort that involved more than 20 individuals and multiple commercial entities in the process. In total, we obtain 400+ language directions of interest and millions of high-quality parallel sentences in the process.

Third, we recruit a group of 20 volunteers who are proficient in English/Russian and one of the Turkic languages. We then translate a well-established test set from WMT 2020 News Translation Task[4] into 8 Turkic languages. This allowed us to obtain high-quality test sets for 56 language directions each containing between 300-1000 sentences. Although the size of the test set remains limited, this is the first comprehensive and high-quality evaluation set for all of the 8 languages involved. We refer to this test set as X-WMT and the data sizes are shown in Table 7.

Table 7: X-WMT Test sets in number of parallel sentences. Bolded entries indicate the original translation direction

	en	**ru**	**tr**	**uz**	**ky**	**kk**	**az**	**ba**	**kaa**	**sah**
en	-									
ru	**1000**	-								
tr	**800**	800	-							

[2] https://www.bible.is/
[3] https://www.ted.com/participate/translate/our-languages
[4] http://www.statmt.org/wmt20/

Table 7 (Continued)

uz	**800**	800	700	-						
ky	500	**500**	400	500	-					
kk	700	700	500	**700**	500	-				
az	**600**	600	500	600	500	500	-			
ba	600	**600**	600	600	500	500	600	-		
kaa	300	**300**	300	300	300	300	300	300	-	
sah	300	300	300	**300**	300	300	300	300	300	-

3.2.2 Ethics and Limitations of the Data Collection

Overall, the process of data collection and compilation took between 4 months involving a total of more than 40 individuals. The resulting dataset consisted of 75M+ parallel sentences and evaluation sets available for 300+ language directions. Although it is important to note that the dataset still remains very imbalanced in terms of resource distribution. Almost 40M of the parallel sentences are within the English-Turkish corpus, while 8 languages only have 8.5K aligned sentences per direction on average.

Previous work has emphasized the importance of data validity and the ethics of data sharing (Bender et al., 2021; Biderman & Scheirer, 2020). As the datasets grow in size, it is increasingly more difficult to ensure their integrity and quality, especially in web-mined text. Unfortunately, our dataset is likely not immune to that quality issue. Here we outline some open issues and limitations of our datasets to encourage researchers to proceed with caution in studying these languages or using them in production.

One of the main limitations of the corpus is domain diversity. For a majority of the languages, their only source of parallel data is from religious texts and books. While this

is helpful and can be utilized in a cross-lingual transfer learning setting, it limits the exploration of phenomena in these languages and makes them unsuitable for production in most of the cases.

The TIL Corpus is a multi-centric dataset where languages are aligned with English and Russian as well as within the family. The way multi-centric datasets are usually constructed is through cross-alignment. To explain this phenomenon, we can take an example of the book Bible. When a translator is translating the book into a new language, Karakalpak, let's say, they use the source language that is common, for example, the English version. Once the translation is over, Karakalpak-English translation is considered to be the original translation direction, but hypothetically it is possible to construct datasets for Karakalpak to any language that English is paired with as well. This results in alignments that are not from the original direction and may possibly have words, phrases or even sentences lost in translation. This compounding error of lost translation may be rampant in the TIL Corpus since the majority of the 400 language directions are not the original translation direction.

Nevertheless, we believe the quality of TIL Corpus is controlled well through the participatory research as each source of data is manually examined and the work of volunteer translators are cross-validated with each other. We believe this corpus will become a valuable resource for the NLP community to engage with Turkic languages.

3.3 TIL Machine Translation

The TIL corpus allows us not only to train MT models for more than 400 language directions involving 22 Turkic languages, but also serves as a valuable resource for researchers and linguists to conduct large-scale analyses and

comparisons. As a practical demonstration of participatory research in action, we conduct a large-scale study of MT using the TIL Corpus. Through the involvement of 15+ individuals, we train MT models for 26 language pairs to serve as the first bilingual baselines. The study reveals several interesting results in terms of domain differences, the effect of the language scripts as well as the evaluation metrics.

3.3.1 Experimental Setup

We train 26 bilingual baselines in 3 different resource categories: high (>5M sentence pairs), medium (100K-5M), and low(<100K). The selection of these pairs was based on multiple factors such as the availability of native speakers in the community, training/testing sets, and the other comparable commercial MT systems (e.g. Google Translate).

All models are Transformers (Vaswani et al., 2017) *(transformer-base)* whose exact configuration depends on the amount of data available for training. Models for low-resource pairs use 256-dimensional word embeddings and hidden layers. Models for mid-resource pairs use the embedding and hidden layer size of 512. The models for high-resource pairs use the same embedding and hidden layer sizes for the encoder, but for the decoder both dimensions are increased to 1024. All models are trained with the Adam optimizer (Kingma & Ba, 2015) over cross-entropy loss with a maximum learning rate of $3 * 10^{-4}$and a minimum of $1 * 10^{-8}$, which warms up for the first 4800 training steps and then decays after reaching the maximum. We use a training batch size of 4096. We use perplexity as our early stopping metric with a patience of 5 epochs. We set a dropout (Srivastava et al., 2014) probability of 0.3 in both the encoder and the decoder. We apply a byte pair encoding (BPE) (Dong et al., 2015; Sennrich et

al., 2015) with a joint vocabulary size of 4K and 32K for low- and mid/high-resource scenarios respectively.

All models use the Joey NMT (Kreutzer et al., 2019) implementation and Apex[5] where possible to speed up training. Models were trained on preemptible GPUs freely available on Google Colab[6].

3.3.2 Evaluation

Evaluation of MT systems often rely on automatic evaluation metrics such as BLEU (Papineni et al., 2002)**,** ChrF (Popović, 2015, 2017) and more recent ones such as BLEURT (Sellam et al., 2020), Comet (Rei et al., 2020). Generally, recent approaches that rely on contextual embeddings outperform more traditional metrics like BLEU and ChrF. However, these approaches often fall short in their language coverage. This is mainly due to the pretraining process of these metrics that require large amounts of monolingual data that some of the low-resource languages might lack. To ensure the coverage of all languages in our study, we employ two more language-agnostic evaluation metrics: BLEU and ChrF. We use the SacreBLEU (Post, 2018) implementation of the BLEU metric and original code for the ChrF provided through the NLTK library.

As part of the evaluation, we use three different test sets in three unique domains: religious, conversational and news. For the religious domain, we use several chapters from the Bible corpus (as described in 3.2.1) which can cover more than 300 language directions. For the conversation domain, we use the overlapping talks from

[5] https://github.com/NVIDIA/apex

[6] https://colab.research.google.com/

the TedTalks dataset[7] which yields test sets for 22 language directions. Finally, we employ 20 volunteers to translate a news dataset used in the WMT 20 News Translation task (X-WMT). X-WMT is our test set in the news domain based on the professionally translated test sets in English-Russian from the WMT 2020 shared task. This set contains approximately 1,000 sentences curated both from English and Russian centric news sources. Through the engagement of native speakers and professional translators, we have partially translated this test set into 7 Turkic languages (Uzbek, Turkish, Kazakh, Kyrgyz, Azerbaijani, Karakalpak, and Sakha).

3.3.3 Results and Discussion

Table 8 highlights the results of the models on the different test sets. First, it is interesting to note that the high-resource languages, despite the millions of parallel sentences, perform modestly on the Bible and TedTalks test sets. Our hypothesis is that the domain of the data for the Turkish-English and Turkish-Russian is very different from these test sets. This hypothesis is also supported by the fact that these models perform a lot higher on the X-WMT which is more in-domain. Another possible explanation is the suboptimal model size and hyperparameter settings which were not tuned extensively due to the computational limitations. In the mid-resource setting, the variance between the results is extremely high as pairs such as *ru-uz and uz-ru* with 1.22M sentence pairs underperform in almost all test domains as compared to other pairs with less data. While this behavior needs to be investigated further with more experiments and analysis, we hypothesize that this is due to the fact that *ru-uz* data

[7] https://www.ted.com/participate/translate/our-languages

comes mostly from the legislative domain (from law documents) and test set domains are very distant from that.

Table 8: Bilingual baselines for 26 language pairs using the TIL Corpus

Pair	Train size	Test size	Bible		Test size	Ted Talks		Test size	X-WMT	
			BLEU	ChrF		BLEU	ChrF		BLEU	ChrF
en-tr	39.9m	416	7.15	0.30	5.2k	12.32	0.43	800	19.87	0.51
ru-tr	16.8m	455	7.44	0.33	5.1k	8.64	0.38	800	8.81	0.41
ru-uz	1.22M	684	6.01	0.41	2.7K	4.51	0.76	800	5.95	0.39
uz-ru	1.22M	684	9.84	0.51	2.7K	7.57	0.73	800	7.45	0.37
en-az	784K	455	10.56	0.24	3.3K	10.58	0.29	600	8.88	0.41
az-en	784K	455	21.17	0.45	3.3K	17.01	0.17	600	12.14	0.42
en-ky	733K	451	6.47	0.32	-	-	-	500	3.18	0.19
ky-en	733K	451	13.08	0.43	-	-	-	500	4.30	0.40
tr-az	634K	606	13.78	0.65	3.6K	20.50	0.40	500	9.68	0.33
az-tr	634K	606	11.66	0.71	3.6K	24.20	0.95	500	11.53	0.49
en-kk	601K	453	3.62	0.61	3.6K	6.31	0.29	700	6.99	0.38
kk-en	601K	453	11.22	0.27	3.6K	9.78	0.30	700	9.75	0.46
en-uz	555K	465	5.23	0.40	3.2K	5.89	0.20	800	6.60	0.42
uz-en	555K	465	16.20	0.63	3.2K	11.61	0.18	800	12.32	0.48
tr-uz	161K	486	6.50	0.14	2.9K	4.28	0.20	700	1.58	0.23
uz-tr	161K	486	7.40	0.32	2.9K	3.92	0.26	700	1.73	0.22
kk-ky	6.4K	696	2.39	0.33	-	-	-	500	0.14	0.09
ly-kk	6.4K	696	2.53	0.24	-	-	-	500	0.11	0.13
en-krc	6.5K	374	5.57	0.25	-	-	-	-	-	-
krc-en	6.5K	374	11.57	0.22	-	-	-	-	-	-
kk-tt	7.7K	678	4.13	0.22	-	-	-	-	-	-
tt-kk	7.7K	678	3.75	0.17	-	-	-	-	-	-
ru-sah	8K	759	2.48	0.27	-	-	-	300	0.08	0.20
sah-ru	8K	759	2.44	0.23	-	-	-	300	0.31	0.16
uz-kaa	8.9K	772	9.90	0.71	-	-	-	300	5.39	0.41
kaa-uz	8.9K	772	9.58	0.60	-	-	-	300	5.24	0.44

Another notable aspect is the importance of scripts in the performance of the models. Language pairs with more than one script consistently under-perform the ones where both the source and target language use the same script. In fact, the best 6 models on the X-WMT test sets all have Latin scripts in both the source and target language. This is likely due to the shared vocabulary of the models since different scripts would result in disjoint vocabularies, which would have a negative effect on performance by preventing knowledge transfer (Aji et al., 2020; Amrhein & Sennrich, 2020).

3.3.4 Conclusion

In this study, we present the first-ever bilingual baselines for 26 language pairs involving Turkic languages. In a participatory research setting, we demonstrate that it is feasible and scalable to bootstrap the development of MT technologies using the TIL Corpus. Our objective is to continue developing more resources and technology for MT in Turkic languages. Such improvements include studies of methods for cross-lingual transfer, extending the coverage of our corpus to more languages and domains, and increasing the size of the test sets to provide more comprehensive benchmarks.

References

Agić, Ž., & Vulić, I. (2019). JW300: A Wide-Coverage Parallel Corpus for Low-Resource Languages. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 3204–3210. https://doi.org/10.18653/v1/P19-1310

Aji, A. F., Bogoychev, N., Heafield, K., & Sennrich, R. (2020). In Neural Machine Translation, What Does Transfer Learning Transfer? Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 7701–7710.

Amrhein, C., & Sennrich, R. (2020). On Romanization for Model Transfer Between Scripts in Neural Machine Translation. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings, 2461–2469.

Artetxe, M., & Schwenk, H. (2019). Massively Multilingual Sentence Embeddings for Zero-Shot Cross-Lingual Transfer and Beyond. Transactions of the Association for Computational Linguistics, 7, 597–610. https://doi.org/10.1162/tacl_a_00288

Bañón, M., Chen, P., Haddow, B., Heafield, K., Hoang, H., Esplà-Gomis, M., Forcada, M. L., Kamran, A., Kirefu, F., Koehn, P., Ortiz Rojas, S., Pla Sempere, L., Ramírez-Sánchez, G., Sarrías, E., Strelec, M., Thompson, B., Waites, W., Wiggins, D., & Zaragoza, J. (2020). ParaCrawl: Web-Scale Acquisition of Parallel Corpora. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 4555–4567. https://doi.org/10.18653/v1/2020.acl-main.417

Bender, E. M. (2011). On achieving and evaluating language-independence in NLP. Linguistic Issues in Language Technology, 6(3), 1–26.

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? . Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, 610–623. https://doi.org/10.1145/3442188.3445922

Biderman, S., & Scheirer, W. J. (2020). Pitfalls in Machine Learning Research: Reexamining the Development Cycle. CoRR, abs/2011.02832. https://arxiv.org/abs/2011.02832

Birch, A., Haddow, B., Tito, I., Barone, A. V. M., Bawden, R., Sánchez-Martínez, F., Forcada, M. L., Esplà-Gomis, M., Sánchez-Cartagena, V., Pérez-Ortiz, J. A., Aziz, W., Secker, A., & van der Kreeft, P. (2019). Global Under-Resourced Media Translation (GoURMET). Proceedings of Machine Translation Summit XVII Volume 2: Translator, Project and User Tracks, 122–122. https://www.aclweb.org/anthology/W19-6723

Brown, P., Cocke, J., Della Pietra, S., Della Pietra, V., Jelinek, F., Mercer, R., & Roossin, P. (1988). A Statistical Approach to Language Translation. Coling Budapest 1988 Volume 1: International Conference on Computational Linguistics. https://www.aclweb.org/anthology/C88-1016

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., & others. (2020). Language models are few-shot learners. ArXiv Preprint ArXiv:2005.14165.

Caswell, I., Breiner, T., van Esch, D., & Bapna, A. (2020). Language ID in the Wild: Unexpected Challenges on the Path to a Thousand-Language Web Text Corpus. Proceedings of the 28th International Conference on Computational Linguistics, 6588–6608. https://doi.org/10.18653/v1/2020.coling-main.579

Choudhary, N., & Jha, G. N. (2011). Creating multilingual parallel corpora in indian languages. Language and Technology Conference, 527–537.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), 4171–4186. https://doi.org/10.18653/v1/N19-1423

Dong, D., Wu, H., He, W., Yu, D., & Wang, H. (2015). Multi-Task Learning for Multiple Language Translation. Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), 1723–1732. https://doi.org/10.3115/v1/P15-1166

El-Kishky, A., Chaudhary, V., Guzmán, F., & Koehn, P. (2020). CCAligned: A Massive Collection of Cross-Lingual Web-Document Pairs. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), 5960–5969. https://doi.org/10.18653/v1/2020.emnlp-main.480

Esplà-Gomis, M., Forcada, M. L., Ramírez-Sánchez, G., & Hoang, H. (2019). ParaCrawl: Web-scale parallel corpora for the languages of the EU. Proceedings of Machine Translation Summit XVII Volume 2: Translator, Project and User Tracks, 118–119.

Joshi, P., Barnes, C., Santy, S., Khanuja, S., Shah, S., Srinivasan, A., Bhattamishra, S., Sitaram, S., Choudhury, M., & Bali, K. (2019). Unsung challenges of building and deploying language technologies for low resource language communities. ArXiv Preprint ArXiv:1912.03457.

Joshi, P., Santy, S., Budhiraja, A., Bali, K., & Choudhury, M. (2020). The State and Fate of Linguistic Diversity and Inclusion in the NLP World. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 6282–6293. https://doi.org/10.18653/v1/2020.acl-main.560

Joulin, A., Grave, E., Bojanowski, P., Douze, M., Jégou, H., & Mikolov, T. (2016). FastText.zip: Compressing text classification models. ArXiv Preprint ArXiv:1612.03651.

Joulin, A., Grave, E., Bojanowski, P., & Mikolov, T. (2017). Bag of Tricks for Efficient Text Classification. Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers, 427–431. https://www.aclweb.org/anthology/E17-2068

Junczys-Dowmunt, M. (2018). Dual Conditional Cross-Entropy Filtering of Noisy Parallel Corpora. Proceedings of the Third Conference on Machine Translation: Shared Task Papers, 888–895. https://doi.org/10.18653/v1/W18-6478

Khusainov, A., Suleymanov, D., Gilmullin, R., Minsafina, A., Kubedinova, L., & Abdurakhmonova, N. (2020). First Results of the “TurkLang-7” Project: Creating Russian-Turkic Parallel Corpora and MT Systems.

Kingma, D. P., & Ba, J. L. (2015). Adam: A Method for Stochastic Optimization. ICLR 2015 : International Conference on Learning Representations 2015.

Kreutzer, J., Bastings, J., & Riezler, S. (2019). Joey NMT: A Minimalist NMT Toolkit for Novices. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations, 109–114. https://doi.org/10.18653/v1/D19-3019

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. ArXiv Preprint ArXiv:1301.3781.

Nekoto, W., Marivate, V., Matsila, T., Fasubaa, T., Fagbohungbe, T., Akinola, S. O., Muhammad, S., Kabongo Kabenamualu, S., Osei, S., Sackey, F., Niyongabo, R. A., Macharm, R., Ogayo, P., Ahia, O., Berhe, M. M., Adeyemi, M., Mokgesi-Selinga, M., Okegbemi, L., Martinus, L., … Bashir, A. (2020). Participatory Research for Low-resourced Machine Translation: A Case Study in African Languages. Findings of the Association for Computational Linguistics: EMNLP 2020, 2144–2160. https://doi.org/10.18653/v1/2020.findings-emnlp.195

Nomoto, H., Okano, K., Moeljadi, D., & Sawada, H. (2018). Tufs asian language parallel corpus (talpco). Proceedings of the Twenty-Fourth Annual Meeting of the Association for Natural Language Processing, 436–439.

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). Bleu: A Method for Automatic Evaluation of Machine Translation. Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, 311–318. https://doi.org/10.3115/1073083.1073135

Pennington, J., Socher, R., & Manning, C. (2014). GloVe: Global Vectors for Word Representation. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 1532–1543. https://doi.org/10.3115/v1/D14-1162

Popović, M. (2017). chrF++: Words helping character n-grams. Proceedings of the Second Conference on Machine Translation, 612–618. https://doi.org/10.18653/v1/W17-4770

Popović, M. (2015). ChrF: character n-gram F-score for automatic MT evaluation. Proceedings of the Tenth Workshop on Statistical Machine Translation, 392–395. https://doi.org/10.18653/v1/W15-3049

Post, M. (2018). A Call for Clarity in Reporting BLEU Scores. Proceedings of the Third Conference on Machine Translation: Research Papers, 186–191. https://doi.org/10.18653/v1/W18-6319

Post, M., Callison-Burch, C., & Osborne, M. (2012). Constructing parallel corpora for six indian languages via crowdsourcing. Proceedings of the Seventh Workshop on Statistical Machine Translation, 401–409.

Qi, Y., Sachan, D., Felix, M., Padmanabhan, S., & Neubig, G. (2018). When and Why Are Pre-Trained Word Embeddings Useful for Neural Machine Translation? Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers), 529–535. https://doi.org/10.18653/v1/N18-2084

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. OpenAI Blog, 1(8), 9.

Rei, R., Stewart, C., Farinha, A. C., & Lavie, A. (2020). COMET: A neural framework for MT evaluation. ArXiv Preprint ArXiv:2009.09025.

Róna-Tas, A. (2015). Turkic Writing Systems. In L. Johanson & É. Á. C. Johanson (Eds.), The Turkic Languages (pp. 126–137). Routledge.

Schuler, D., & Namioka, A. (1993). Participatory design: Principles and practices. CRC Press.

Schwenk, H., Chaudhary, V., Sun, S., Gong, H., & Guzmán, F. (2019). WikiMatrix: Mining 135M Parallel Sentences in 1620 Language Pairs from Wikipedia.

Searle, J. R. & others. (1980). Minds, brains, and programs. The Turing Test: Verbal Behaviour as the Hallmark of Intelligence, 201–224.

Sellam, T., Das, D., & Parikh, A. (2020). BLEURT: Learning Robust Metrics for Text Generation. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 7881–7892. https://doi.org/10.18653/v1/2020.acl-main.704

Sennrich, R., Haddow, B., & Birch, A. (2015). Neural machine translation of rare words with subword units. ArXiv Preprint ArXiv:1508.07909.

Sloane, M., Moss, E., Awomolo, O., & Forlano, L. (2020). Participation is not a design fix for machine learning. ArXiv Preprint ArXiv:2007.02423.

Spinuzzi, C. (2005). The methodology of participatory design. Technical Communication, 52(2), 163–174.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. Journal of Machine Learning Research, 15(1), 1929–1958.

Suárez, P. J. O., Sagot, B., & Romary, L. (2019). Asynchronous pipelines for processing huge corpora on medium to low resource infrastructures (P. Bański, A. Barbaresi, H. Biber, E. Breiteneder, S. Clematide, M. Kupietz, H. Lüngen, & C. Iliadi, Eds.; pp. 9–16). Leibniz-Institut für Deutsche Sprache. https://doi.org/10.14618/ids-pub-9021

Tantuğ, A. C., Oflazer, K., & El-Kahlout, I. D. (2008). BLEU+: A Tool for Fine-Grained BLEU Computation. LREC.

Tiedemann, J. (2020). The Tatoeba Translation Challenge – Realistic Data Sets for Low Resource and Multilingual MT. Proceedings of the Fifth Conference on Machine Translation, 1174–1182. https://www.aclweb.org/anthology/2020.wmt-1.139

Tsarfaty, R., Bareket, D., Klein, S., & Seker, A. (2020). From SPMRL to NMRL: What Did We Learn (and Unlearn) in a Decade of Parsing Morphologically-Rich Languages (MRLs)? Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 7396–7408. https://doi.org/10.18653/v1/2020.acl-main.660

Tsarfaty, R., Seddah, D., Kübler, S., & Nivre, J. (2013). Parsing Morphologically Rich Languages: Introduction to the Special Issue. Computational Linguistics, 39(1), 15–22. https://doi.org/10.1162/COLI_a_00133

TURING, A. M. (1950). I.—COMPUTING MACHINERY AND INTELLIGENCE. Mind, LIX(236), 433–460. https://doi.org/10.1093/mind/LIX.236.433

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention Is All You Need. CoRR, abs/1706.03762. http://arxiv.org/abs/1706.03762

Washington, J. N., Salimzianov, I., Tyers, F. M., Gökırmak, M., Ivanova, S., & Kuyrukçu, O. (2019). Free/open-source technologies for Turkic languages developed in the Apertium project. Proceedings of the International Conference on Turkic Language Processing (TURKLANG 2019).

Xue, L., Constant, N., Roberts, A., Kale, M., Al-Rfou, R., Siddhant, A., Barua, A., & Raffel, C. (2021). mT5: A massively multilingual pre-trained text-to-text transformer.