University of Hamburg

Masters Thesis

Leveraging Morphological and Lexical Features in Synthetic Data Generation for Dialect-Specific Machine Translation

Author: Christian Schuler (6449321)

Supervisor: Dr. Sina Ahmadi
University of Zurich
Supervisor: Dr. Seid Muhie Yimam
University of Hamburg
Examiner: Prof. Dr. Chris Biemann
University of Hamburg

A thesis submitted in fulfillment of the requirements for the Master of Science Informatik

of the

Language Technology
Faculty of Informatics
Fachbereich Informatik

July 1, 2024

UNIVERSITY OF HAMBURG

Abstract

Faculty of Informatics
Fachbereich Informatik

Master of Science Informatik

Leveraging Morphological and Lexical Features in Synthetic Data Generation for Dialect-Specific Machine Translation

by Christian Schuler

This work attempts to advance neural machine translation of dialects from low-resource languages in a meaningful way by utilizing synthetic text data, created via incorporation of linguistic information.

Natural language processing, Neural Machine Translation, Low-resourced language, Dialectal varieties, Linguistic information

List of Abbreviations

Chapter 1

Introduction

1.1 Natural Language Processing and Linguistics

natural language processing (NLP) encompasses a vast field of theories and applications. It is concerned with all forms of language, be it written texts, audible speech or visual sign languages, and how they can be processed by computers in differing degrees of sophistication. عليكم سلام ئەردەلانى

Prior work in NLP has resulted in more and more complex applications and tools that enable astounding investigations of text. These include many tasks such as sentiment analysis, conversational agents (CA), and machine translation (MT), with the latter being the focus of this work.

Linguistics is the study of languages, language families, and their history. It sheds light on language properties and characteristics, regarding sounds, grammar, and meaning. Linguistic investigations already had a long tradition long before the first computer was even built and insights gained in this field serve as the backbone of language technology. Despite the rapid progress thanks to deep learning, such as large language models like GPT, linguistics still remains an essential component in analyzing the performance of models and studying the languages of the world.

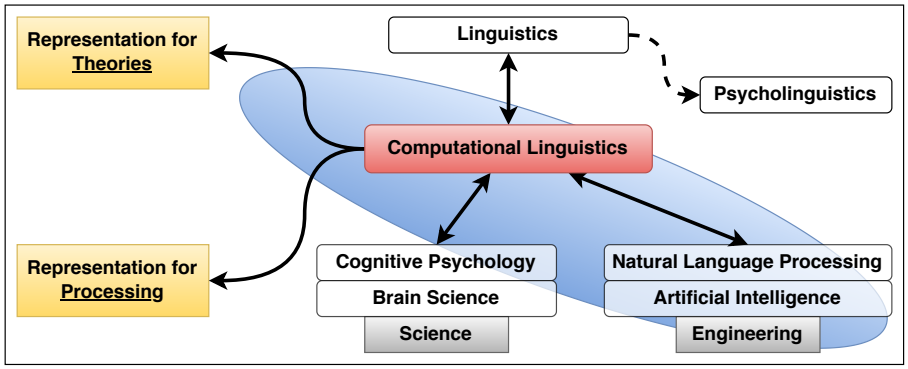


Figure 1.1: A schematic view of the research disciplines surrounding computational linguistics. The bottom disciplines are concerned with the processing of language, while the top discipline (linguistics), is concerned with language rules (Tsujii, 2021).

Both, natural language processing (NLP) and computational linguistics (CL) address natural language, doing this from an algorithmic and a linguistic perspective respectively. The simplified schematic view of Tsujii (2021), paints a rough picture of the interplay of the research disciplines revolving around CL (seen Figure 1.1). This schema places the more theoretical computational linguistics as a sub-field of linguistics and the more engineering-oriented natural language processing can be found

related, but separated from it. Even though, tightly connected and sometimes overlapping, this is not generally valid and these terms can oftentimes be found to be conflated.

1.2 Low-Resource Languages

Of the 7168 languages[1] worldwide only 10-15 of them can be considered economically important and have a strong digital presence online (Bali et al., 2019). Recent NLP research has not only witnessed the blossoming of large language models (LLMs), but also a considerable shift towards engaging with low-resource languages. Without special attention, these languages are endangered to soon perish (soon meaning in just a few generations) and with them a great chunk of their respective cultural heritage (Kornai, 2013). Although English along with a few dozen other languages make up a majority of the internet and receive a lot of attention and effort in research, many languages are neglected resulting in an uneven resource hierarchy (Moseley and Nicolas, 2010) (see Figure 1.2).

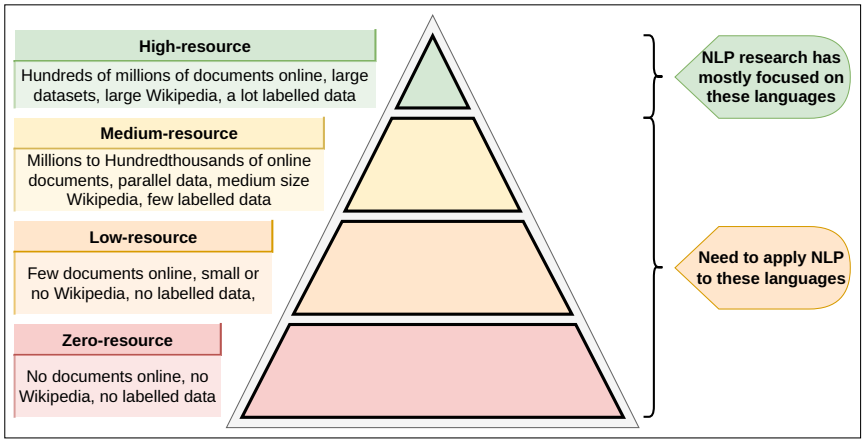


FIGURE 1.2: A conceptual view of the NLP resource hierarchy, with only very few languages at the top and most of the world's languages at the bottom. Modified from[2].

1.3 Machine Translation

The investigation of using software to translate text or speech from a source language into a target language falls under machine translation (MT). MT has been considered the flagship of NLP due to its founding history in the 1940s with its true origins found in the Arabic cryptography of the ninth century (Dupont, 2017). It was Al-Kindi who developed techniques for systematic language translation, including cryptoanalysis and frequency analysis besides probability and statistics. Multiple approaches have emerged from prior research, which, in general, depend on more and more data to be applicable. Starting with rule-based and dictionary-based machine translations (Tripathi, 2010), science moved to statistical approaches (Koehn, 2009), which use statistical methods on bilingual text corpora and subsequently to neural approaches (Koehn, 2020) based on deep learning and showing rapid progress in recent years.

[1] https://www.ethnologue.com/
[2] https://www.ruder.io/unsupervised-cross-lingual-learning/

By advancing machine translation via utilizing recurrent neural networks, the area of neural machine translation (NMT) has been successful in generating state-of-the-art results for many languages (Koehn, 2020). NMT can be considered to be the state-of-the-art of machine translation and has seen numerous approaches and methods for fine-tuning models and improving results. These efforts include the improvement of translation's accuracy and acceptance, the reduction of required time and resources, but also enabling an easier access for humans from around the world. Sufficient text data of adequate quality is a strict necessity for training models for translation and for evaluating their performance. This is where low-resource languages and their associated data scarcity results in them falling short, which is best displayed by a major lack of provided solutions for their speakers.

This work strives to alleviate the scarcity-based issues of low-resource languages by exploring different data collection methods.

1.4 Objectives & Research Questions

Low-resource languages and especially their dialects lack the required data to train sophisticated models to do machine translation. Alleviating this issue by creating adequate data synthetically could benefit many language communities world-wide.

This work attempts to advance neural machine translation of dialect variations (especially from low-resource Languages) in a meaningful way by utilizing synthetic text data, created via incorporation of linguistic information. An additional benefit of this approach lies in the cost-effectiveness compared to manual data annotations done by experts, who have to be acquired for each target language and are often very time- and money-consuming.

In order to generate data that is reasonably useful, linguistic rules have to be identified, codified and then incorporated into the data creation process to have the emerging dialectal variations in concordance with real data, as produced by native speakers, to be used in downstream tasks and to improve performance of machine translation systems.

The linguistic feature, dubbed "*negative concord*" and used in Ziems et al. (2022) for a very similar purpose, will serve as an illustration. This feature involves two negative morphemes to convey a single negation (Martin and Wolfram, 1998) and results in the Standard American English sentence "He doesn't have a camera" to look like "He don't have no camera" in African American Vernacular English. This particular transformation is said to be sensitive to the verb-object dependency structure and requires the object to be an indefinite noun (Green, 2002). By covering enough linguistic features that together define a language variety, the already available text data from the standard variety can be transformed and then be used in downstream tasks and applications, like in this work machine translation.

This work aims to answer the following research questions:

RQ1: What is the performance of the current state-of-the-art models in translating dialects?

RQ2: Can we incorporate linguistic information in MT to synthetically generate sentences in language variants so that dialects of various (and especially low-resource) languages can be processed more efficiently?

RQ3: When using synthetic data, what roles do state-of-the-art approaches in fine-tuning, transfer learning and adapters play in improving the performance of MT systems to process (particularly low-resource) language variations effectively?

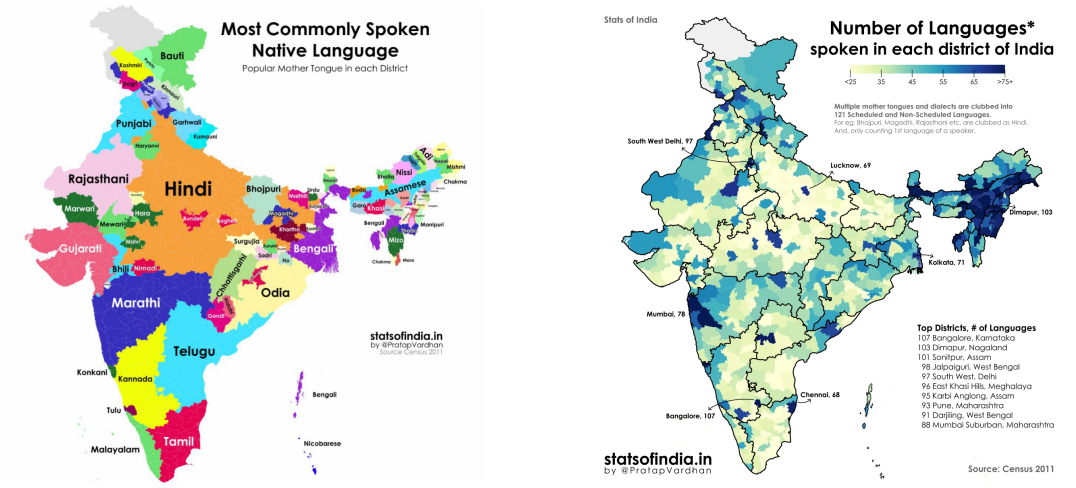
RQ4: What are requirements for deriving tools and processes that can be applied to vastly different languages from various language families?

Chapter 2

Motivation

2.1 Languages and Dialects

Languages come in different flavors, called dialects, that can be considered to be leaves on a language branch. These dialects are often unknown outside their local sphere and find little recognition in the wider, global, population and research alike. That Hindi is the main language of India and that everyone, who grows up in India, speaks Hindi, is an often encountered misconception. India is a cauldron for a plethora of languages. Many different languages can be found as the most commonly spoken native language of a region and many regions host an astounding number of languages, as can be seen in Figure 2.1.



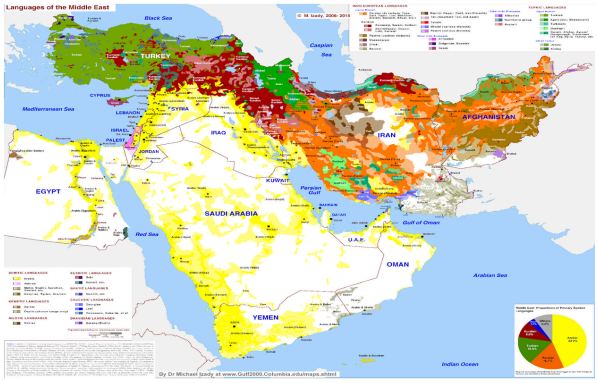
(A) Diversity of languages

(B) Vast numbers of languages

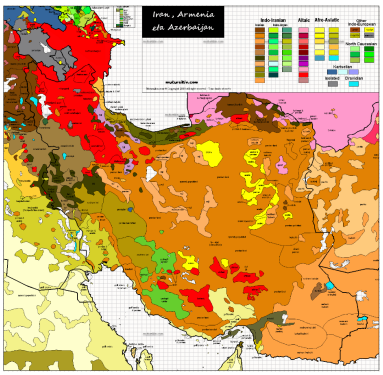
FIGURE 2.1: India. Credit: statsofindia.in Source: Indian census 2011

Some languages and dialects are not only geographically separated but also display a very splintered or fractured image on the map. Neither straight lines nor heat maps can do justice to the nature of how languages (and the humans who speak them) distribute on this planet. Various degrees of granularity can be found in text descriptions but also in maps, that provide information about the geolocations related to languages. A lot of the maps indicate very complex borders, as in Figure 2.2, and identifying true native speakers (e.g. for sending out field workers to acquire language data) can be complicated.

[2] https://gulf2000.columbia.edu/maps.shtml
[2] http://www.muturzikin.com/cartesasie/2.htm



(A) Languages of the Middle East in 2000. Source[1]



(B) Languages of Iran, Armenia and Azerbaijan from 2008. Source[2]

FIGURE 2.2: Languages and linguistic composition of Iran and the surrounding area.

2.1.1 Dialect Continua

In addition to the previously-discussed issues, comes the smoothness of many language and dialect transitions. Languages and their speakers do not exist in perfect isolation from each other; they interact and intermingle with each other. People being exposed to another language for long stretches of time might start using some of the words or affect the native speaker with whom they interact (Tavadze, 2019). Some languages are known to make heavy use of loanwords from other languages, which sometimes complicates language identification and processing (Matras, 2017; Matras, 2019) (e.g. some Kurdish dialects which incorporated Arabic, Farsi and even Turkish words). Once enough time has passed, a new language or dialect can grow out of this interaction, now positioned in between the previously dominant languages. This new language can then be considered to be closer to both of the other languages than they are to each other. In this way, it can happen, that speaker of the new language understand their neighbors, while these can not communicate with each other without problems (Salam Khalid, 2020). This and similar processes have resulted in many dialect continua (Salam Khalid, 2015). Figure 2.3 exemplifies this via a German sentence which gradually changes while moving through dialect regions.

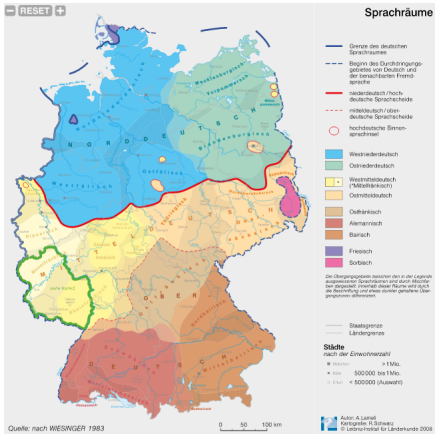
2.1.2 Mutual Understanding between Dialects

Speakers of many dialects that officially belong to the same language can not properly understand each other. In cases like Germany, this is less of an issue, since every citizen studies Standard German in school, which alleviates dialect-based communication problems. But languages and dialects that are spoken in regions in which there is no agreed-upon standard language have been observed to suffer from mutual unintelligibility Salam Khalid, 2020.

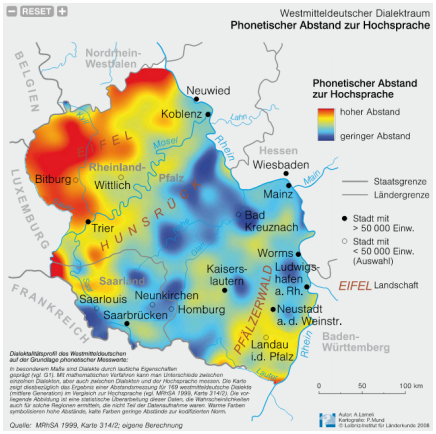
2.2 Data Scarcity

There are many reasons that can lead to a language or a dialect being considered to be low-resource. The most obvious one is a low number of native speakers, as for the Saterland Frisian which has approximately 2,000 speakers [3], but also political,

[3] https://en.wikipedia.org/wiki/Saterland_Frisian_language#cite_note-e21-1



(A) Areas of German dialects.



(B) Phonetic distance to Standard German.

Niederdeutsch	**ik**	maken	dorp	dat	pund	appel
Niederfränkisch	**ich**	**maken**	dorp	dat	pund	appel
Ripuarisch	ich	**machen**	**dorp**	dat	pund	appel
Moselfränkisch	ich	machen	**dorf**	**dat**	pund	appel
Rheinfränkisch	ich	machen	dorf	**das**	**pund**	appel
Sächsisch	ich	machen	dorf	das	**pfund**	**appel**
Oberdeutsch	ich	machen	dorf	das	pfund	**apfel**
Standardsprache	**ich**	**machen**	**dorf**	**das**	**pfund**	**apfel**

(C) Phonetic shifts of the same German sentence.

FIGURE 2.3: Distribution and transitions of German dialects.
Source: (Lameli, 2008)

cultural, religious and economic factors can be at fault. This scarcity of data is a serious bottleneck for developing or even testing any NLP tools and applications. The trend of NLP research towards solutions built on various neural network architectures comes hand-in-hand with a requirement for vast amounts of language data, be it from text, image or speech.

Solving, or at least circumventing, this issue is the main motivation of the current work. To take part in globalization and modern culture, the native speakers of low-resource languages and dialects need their languages to be recognized by technology. Only once these languages have been elevated from their low-resource state, can they benefit from state-of-the-art NLP solutions. Neglecting these languages and missing the opportunity to attend to them now, might very well lead to the degradation of language-based cultural heritage(Crystal, 2000; Bird, 2020).

2.3 Machine Translation and Dialects

How badly even very popular and large translation systems fail to properly process text from less represented dialects can be seen in Table 2.1. While the translation from Standard German is not perfect, it still comes very close and conveys the true meaning of the original sentence. For the two German variations Saxony and Danube Bavarian, the translations are close to useless and almost consist of simply copying the input text.

[4] https://translate.google.com/?sl=de&tl=en&op=translate&hl=en (accessed in July 2023)

Language (variety)	Text
Saxony	*Eema' ham sisch dor Nordwind und de Sonne geschdridden, währ vunn deen beeden dor Schdärgre is, als ä Wandror, där nen wahrm Manddl anhadde, däs Wägs gohm.*
"English"	*Eema 'had sisch dor north wind and de sun, while vunn they ended dor Schdärgre, when ä Wandror, dan had a real manddl, the wags gohm.*
Danube Bavarian	*Amoi håbn si die Sunn und da Nurdwind gstrittn wea von de beidn woi da Sterkare warat, wia pletzlich a Wåndara mit aan woamen Måntl vurbeikemma is.*
"English"	*Amoi håbn si the Sunn and da Nurdwind gstretn wea von de both woi da Sterkare warat, like suddenly a Wåndara with a woamen Måntl vurbeikemma is.*
Standard German	*Einst stritten sich Nordwind und Sonne, wer von ihnen beiden wohl der Stärkere wäre, als ein Wanderer, der in einen warmen Mantel gehüllt war, des Weges daherkam.*
"English"	Once upon a time, the North Wind and the Sun were arguing about which of them was the stronger, when a wanderer wrapped in a warm cloak came along the path.

TABLE 2.1: The same text in different variations of German as found in (Alam, Ahmadi, and Anastasopoulos, 2023) and their translations into English according to Google Translate[4]

2.4 Language Classification Issues

As seemingly custom by now, the exact notation/names of these dialects are shrouded in mystery and can probably only be found spoken of in the ancient legends of old.

2.4.1 Bengali

2.4.2 The Language Bengali:

"Bengali generally known by its endonym Bangla, is an Indo-Aryan language native to the Bengal region of South Asia.
With approximately 234 million native speakers and another 39 million as second language speakers as of 2017, Bengali is the sixth most spoken native language and the seventh most spoken language by the total number of speakers in the world.
Bengali is the fifth most spoken Indo-European language." [5]

2.4.3 Language Variety Classification:

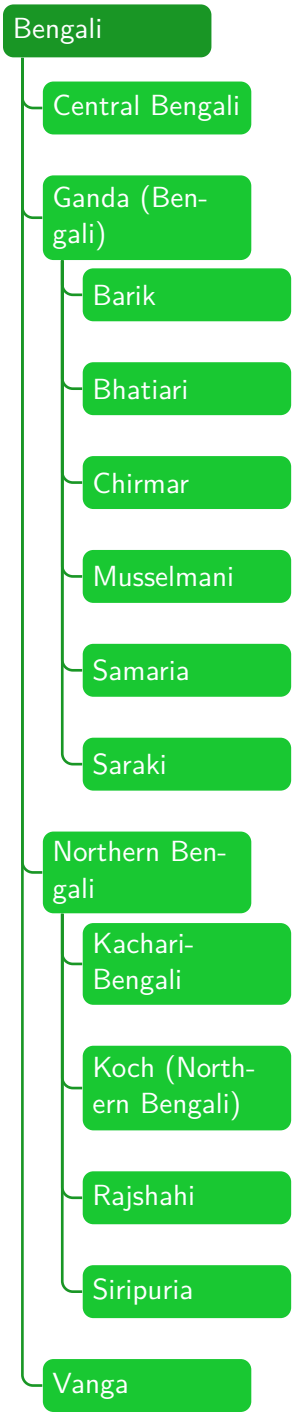
The varieties are named according to major cities where the data was collected.

[5] https://en.wikipedia.org/wiki/Bengali_language

> "Anecdotally, Bangladesh witnesses a linguistic transition approximately every 10 miles. This work specifically focuses on five prominent dialects from five locales of Bangladesh:
>
> - Jessore
> - Khulna
> - Kushtia
> - Barisal
> - Dhaka
>
> The selection of these dialects was strategic, encompassing regions both close to the origin of standard Bengali (Jessore, Kushtia) and those situated farther away."

Taken from Alam, Ahmadi, and Anastasopoulos, 2023.

Bengali varieties included in the used data.



Bengali variety classification according to Glottolog [a].

[a] https://glottolog.org/resource/languoid/id/beng1280

While the linguistic map of Bangladesh 2.5 only lists 6 languages of Bangladesh (Shendu, Tangchangya, Pankhu, Khumi Chin, Khasi & Pnar & War, and Kok Borok), the original source of this map lists 42 living languages of Bangladesh (refer to 2.2.

Assamese	(Asambe, Asami, Ahomiyo)
A'tong	(Attong)
Bengali	(Banga-Bhasa, Bangala, Bangla)
Bihari	(Urdu)
Bishnupriya	(Bishnupria, Bishnupuriya, Bisna Puriya)
Burmese	(Bama, Bamachaka, Myen)
Chak	(Sak, Tsak)
Chakma	(Sangma, Sakma, Takam)
Chin, Asho	(Khyang, Khyeng, Qin, Sho, Shoa)
Chin, Bawm	(Bawm, Bawn, Bawng, Bom)
Chin, Falam	(Falam, Fallam, Halam, Hallam Chin)
Chin, Haka	(Baungshe, Haka, Lai)
Chin, Khumi	(Kami, Khami, Khumi, Khuni, Khweymi, Kumi)
Chittagonian	(Chatgaiyan Buli, Chatgaya, Chittagonian Bengali)
Garo	(Garrow, Mande, Mandi)
Hajong	(Hajang)
Indian Sign Language	
Khasi	(Cossyah, Kahasi, Khasie, Khasiyas, Khassee, Khuchia, Kyi)
Koch	(Koc, Kocch, Koce, Kochboli, Konch)
Koda	
Kok Borok	(Debbarma, Tipura, Tripura, Tripuri)
Kurux	(Kurukh, Oraoan, Uraon)
Marma	("Mogh")
Megam	(Migam, Negam)
Meitei	(Kathe, Kathi, Manipuri, Meetei, Meiteiron, Meithe, Meithei, Mitei, Mithe, Ponna)
Mizo	(Hualngo, Lei, Lusai, Lushai, Lushei, Sailau, Whelngo)
Mru	(Maru, Mrung, Murung)
Mundari	(Colh, Horo, Mandari, Mondari, Munari, Munda)
Pangkhua	(Pangkhu, Pankho, Panko, Pankhu)
Pnar	
Rakhine	(Rakhain, Rakkhaine, Mogh)
Rangpuri	(Bahe Bangla, Anchalit Bangla, Kamta, Polia)
Riang	(Kau Bru, Reang)
Rohingya	(Rohinga, Rohinja)
Sadri, Oraon	
Santali	(Har, Hor, Sandal, Sangtal, Santal, Santhali, Satar, Sonthal)
Sauria Paharia	(Malto, Paharia)
Sylheti	(Sileti, Siloti, Sylhetti, Sylhetti Bangla, Syloti, Syloty, Srihattia)
Tangchangya	(Tanchangya)
Tippera	(Kok Borok, Tipperah, Tippurah, Tipra, Tipura, Triperah, Tripura)
Usoi	(Kau Brung, Tippera, Tripura, Unshoi, Unsuiy, Ushoi)
War-Jaintia	

TABLE 2.2: Living languages of Bangladesh according to http://www.muturzikin.com/cartesasiesudest/7.htm

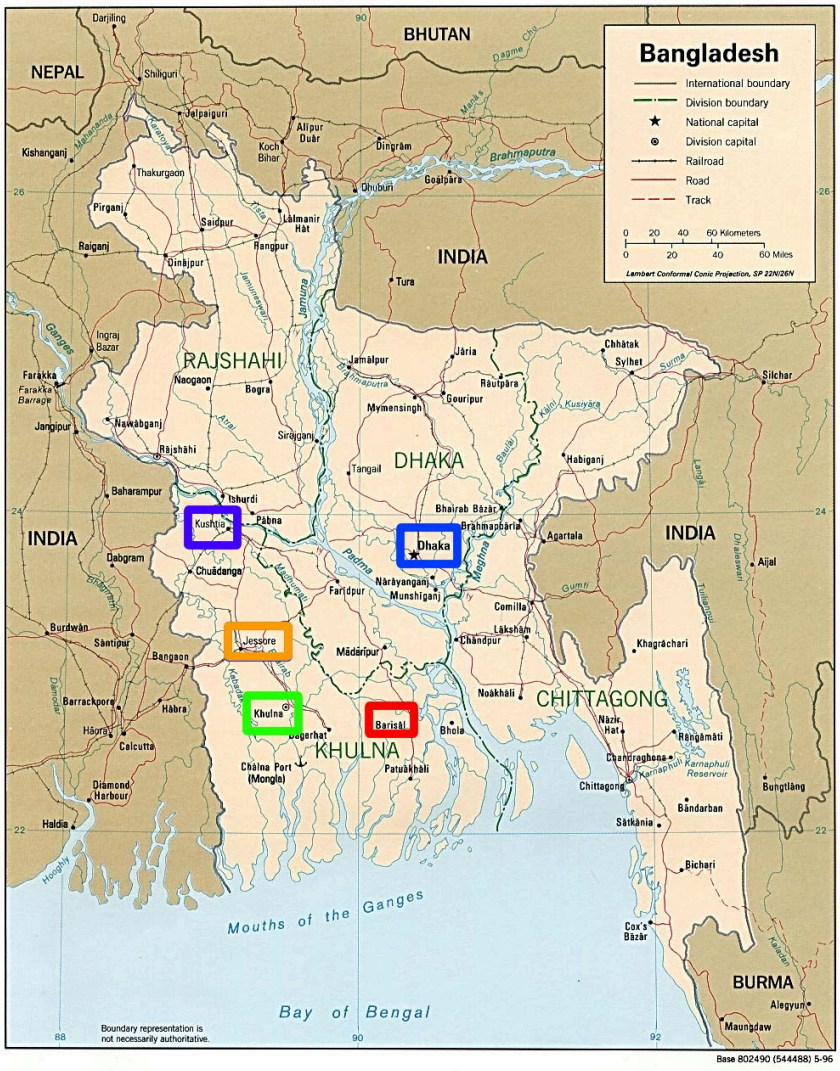


FIGURE 2.4: Locations of collected data, marked Red: Barisal, Blue: Dhakaiya (Dhaka), Orange: Jessore, Green: Khulna, Purple: Kushtia. Based on a map of Bangladesh from https://www.worldofmaps.net/en/asia/maps-of-bangladesh/map-of-bangladesh-political-map.htm

2.4.4 Central Kurdish

2.4.5 The Language Central Kurdish:

"Sorani Kurdish or Central Kurdish, also called Sorani, is a Kurdish dialect or a language that is spoken in Iraq, mainly in Iraqi Kurdistan, as well as the provinces of Kurdistan, Kermanshah, and West Azerbaijan in western Iran.
Sorani is one of the two official languages of Iraq, along with Arabic, and is in administrative documents simply referred to as Kurdish." [6]

[6] https://en.wikipedia.org/wiki/Sorani

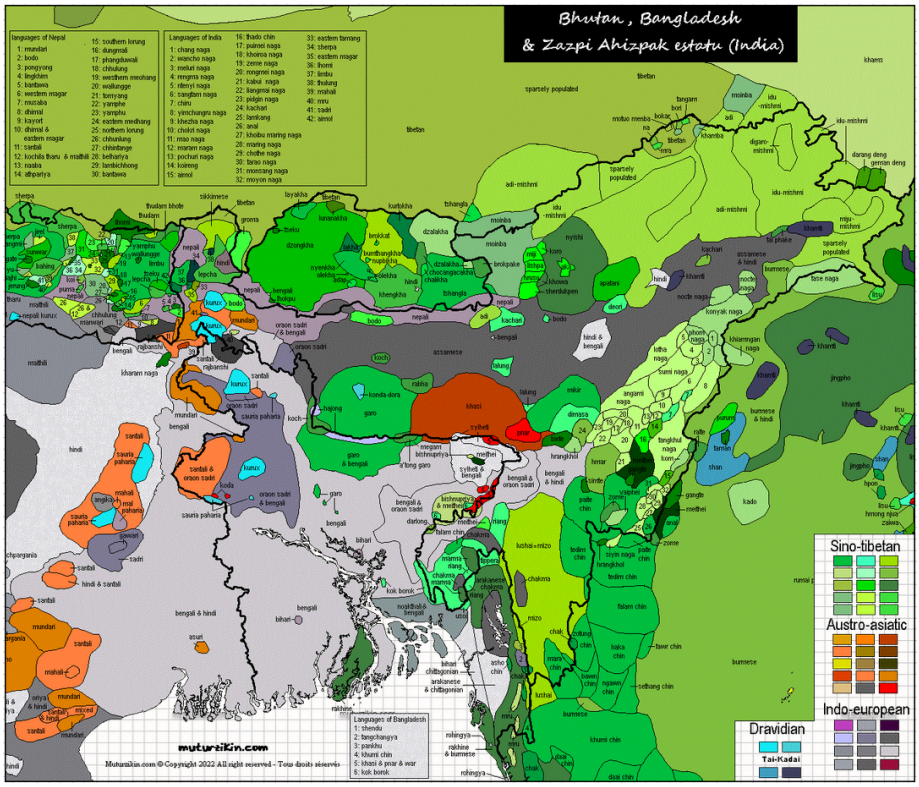


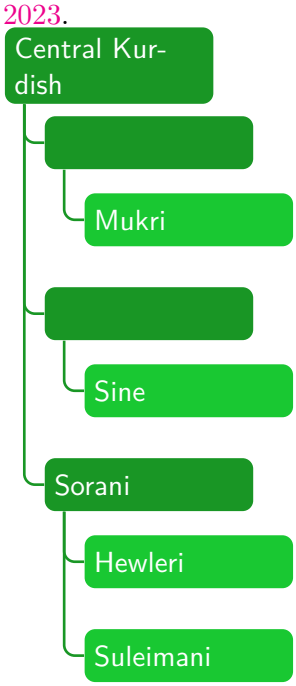
FIGURE 2.5: Linguistic map of Bangladesh, Bhutan & Seven Sister States (India).
Source http://www.muturzikin.com/cartesasiesudest/7.htm

2.4.6 Language Variety Classification:

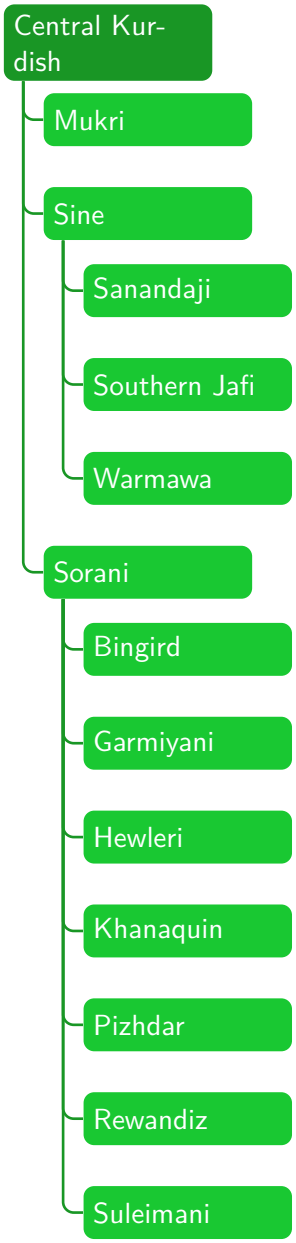
The varieties are often named according to major cities which can be considered to be their epicenter.

> "Kurdish is known as a dialect continuum and is mainly classified into Northern, Central, and Southern dialects and is closely related to Zaza-Gorani languages, Laki and Lori (Ahmadi et al., 2023). In this project, we focus on the varieties of Central Kurdish, also known as Sorani, which are mainly spoken in Kurdistan, Iran, and Iraq.
> Although more extensive studies on Kurdish dialectology are needed to describe Central Kurdish varieties, the following local names are generally and broadly used to refer to the dialects of Central Kurdish spoken in regions of the cities specified in parentheses:
>
> - Babanî (Sulaymaniyah, Iraq) (McCarus, 1956),
> - Ardalanî (Sanandaj, Iran),
> - Cafî (Javanrud, Iran),
> - Mukriyanî or Mukrî (Mahabad, Iran) (De Chiara, 2018), and
> - Hewlêrî (Erbil, Iraq).
>
> Among these, the variant of Sulaymaniyah is the most studied one, which is also widely used as a standard variant of Central Kurdish in the press and media (Thackston, 2006). According to various linguistic analyses of fieldwork data, Matras (2019) classifies Central Kurdish varieties into Northern and Southern Sorani, with their epicenters being based on the dialects of Erbil (Hewlêr in Kurdish) and Sulaymaniyah (Silêmanî in Kurdish). Based on this classification, Babanî, Ardalanî, and Cafî belong to Southern Sorani, while Mukriyanî and Hewlêrî belong to Northern Sorani. Similarly, we believe that the selected varieties can further elucidate the distinctiveness of the varieties and the classification quantitatively."

Taken from Alam, Ahmadi, and Anastasopoulos, 2023.



Central Kurdish varieties included in the used data.



Central Kurdish variety classification according to Glottolog [a].

[a] https://glottolog.org/resource/languoid/id/cent1972

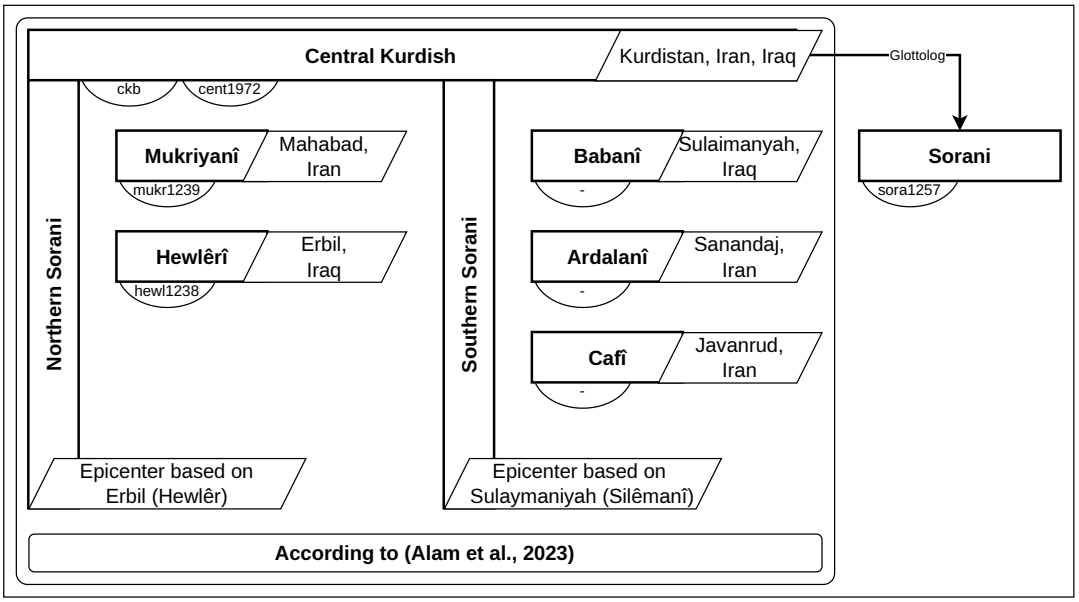


FIGURE 2.6: Central Kurdish variety classification according to (Alam, Ahmadi, and Anastasopoulos, 2023), with Sulaymaniyah widely used as a standard variant.

2.4.7 German

2.4.8 The Language German:

"German is a West Germanic language mainly spoken in Western Europe and Central Europe. It is the most widely spoken and official or co-official language in Germany, Austria, Switzerland, Liechtenstein, and the Italian province of South Tyrol. It is also an official language of Luxembourg and Belgium, as well as a recognized national language in Namibia.
Outside Germany, it is also spoken by German communities in France (Alsace), Czech Republic (North Bohemia), Poland (Upper Silesia), Slovakia (Košice Region, Spiš, and Hauerland), and Hungary (Sopron)." [7]

2.4.9 Language Variety Classification:

This would be way out of scope, considering the amount of data being used. German was mainly included for easy sanity checking by the author.

2.4.10 Northern Kurdish

2.4.11 The Language Northern Kurdish:

"Kurmanji, also termed Northern Kurdish, is the northernmost of the Kurdish languages, spoken predominantly in southeast Turkey, northwest and northeast Iran, northern Iraq, northern Syria and the Caucasus and Khorasan regions.
It is the most widely spoken form of Kurdish." [8]

Of the Kurdish language, the Kurmanji dialect group has the highest number of speakers and it received some attention in various research projects- something that can not be said for it's sub-dialect Kobani.

[7] https://en.wikipedia.org/wiki/German_language
[8] https://en.wikipedia.org/wiki/Kurmanji

2.4.12 Language Variety Classification:

This would also be out of scope for the current proof of concept experimentation. Working together with a native speaker of the Kobani dialect (a sub dialect not included in Glottolog, but with approximate position in the lanuage tree being: Kurdish → Northern Kurdish → Western Kurmanji → Southwest Kurmanji → Aleppo Kurmanji → Rojavayi → Kobani). I am currently collecting linguistic features that distinguish Kobani from the standard Kurmanji, in order to generate Kobani test data.

2.4.13 Tigrinya

2.4.14 The Language Tigrinya:

"Tigrinya is an Ethiopian Semitic language commonly spoken in Eritrea and in northern Ethiopia's Tigray Region by the Tigrinya and Tigrayan peoples.
It is also spoken by the global diaspora of these regions." [9]

2.4.15 Language Variety Classification:

Glottolog [10] distinguishes between Northern Tigrinya and Southern Tigrinya, which corresponds to Eritrea being north of Ethiopia.

> "Tigrinya is an Ethio-Semitic language predominantly spoken in Eritrea and by the Tigrayan people in the Tigray Region of northern Ethiopia.
> Within Tigrinya, two major varieties exist the Eritrean dialect and the Ethiopian dialect. "

Taken from Alam, Ahmadi, and Anastasopoulos, 2023.

[9] https://en.wikipedia.org/wiki/Tigrinya_language
[10] https://glottolog.org/resource/languoid/id/sout3326

2.5 Language Differences

Chapter 3

Research Questions

3.1 RQ1

Text

3.2 RQ2

Text

3.3 RQ3

Text

Chapter 4

Theoretical Background

4.1 Natural Language Processing and Linguistic Research

4.2 Natural Language Processing:

The term Natural Language Processing (NLP) encompasses a vast field of theories and applications. It is concerned with all forms of language, be it written texts, audible speech or visual sign languages, and how they can be processed by computers in differing degrees of sophistication. There are tokenization methods that separate text into smaller units like sentences, words, or single phonemes. In their more simple forms, they just look for empty spaces to distinguish between words and look for punctuation in order to identify sentences. These simple rules quickly fail in the face of ambiguity, since a full stop (dot) can indicate the end of a sentence, but also be part of a number (e.g. 3.1415) or belong to a noun (e.g. Mr.). Different writing systems complicate the separation into words, such as Chinese, which uses a logographic alphabet. Prior work in NLP has resulted in more and more complex applications and tools which enable astounding investigations of text. These include Named Entity Recognition (NER), Information Retrieval (IR), Sentiment Analysis, Questioning & Answering (Q&A), Conversational Agents (CA), Machine Translation (MT) and many more. The remainder of this work focuses on Machine Translation.

4.3 Linguistic Research / Linguistic Rules:

Linguistic research is the study of languages, language families and their history. It contains the investigation of language properties and characteristics, regarding sounds, grammar and meaning. A number of concepts which build the foundation of almost everything that happens in NLP have been systematically investigated and formulated by past linguists.

4.4 Machine Translation

The investigation of using software to translate text or speech from a source language into a target language falls under Machine Translation (MT), which can be considered to be a sub-field of computational linguistics. Multiple approaches have emerged from prior research, which, in general, depend on more and more data to be applicable. Starting with rule-based and dictionary-based machine translations, science moved to statistical approaches, which use statistical methods on bilingual text corpora and subsequently to neural approaches based on deep learning and showing a rapid process in recent years. Current advances in the field of Artificial Intelligence (AI) make it challenging to predict the future of Machine Translation.

4.5 Neural Machine Translation:

By advancing Machine Translation via utilizing recurrent neural networks, the area of Neural Machine Translation (NMT) has been successful in generating state-of-the-art results for many languages. Koehn, 2020 NMT can be considered to be the state-of-the-art of Machine Translation and has seen numerous approaches and methods for fine-tuning models and improving results. These efforts include the improvement of translation's accuracy and acceptance, the reduction of required time and resources, but also enabling an easier access for humans from around the world. At times including only two, sometimes over a hundred different languages, these efforts are most often found in relation to English, as it is the currently dominating language on the internet and therefore provides the largest trove of text data. Sufficient text data of adequate quality is a strict necessity for training models for translation and for evaluating their performance. This is where the prior discussed nature of Low-Resource languages and associated data scarcity turns into such a crippling hindrance, displayed by a major lack of provided solutions for their speakers.

How to alleviate issues of Low-Resource language and dialect translation is what this work strives to accomplish by exploring methods to solve problems of data scarcity.

Chapter 5

Related Work

This section provides an overview of the up to now collected related work that is being explored closer at the time of this writing. The related work is presented in an initial categorization, closely aligning with the aims of this work.

5.1 Approaches to Machine Translation of Low-Resource Languages

There have been numerous works following different approaches, all having their benefits and downsides. It might be possible to combine some of these, while others will only be guiding or informing this work from the sidelines. While Table 5.1 shows a vast amount of related work to potentially draw from, some finer foci of which literature to include are already emerging, becoming more clear in Phase 1 of this work.

5.2 Strategies to Enhance Translation Quality in Low-Resource Situations

One such promising strategy is proposed by Reimers and Gurevych, 2020, where they introduce Multilingual Knowledge Distillation as a method to make monolingual sentence embeddings multilingual with aligned vector spaces between languages. They demonstrate a successful transfer of properties from the source language vector space (English) to various target languages and it is said that their model can be extended to multiple languages in the same training process. Additionally, their work provides an overview of a plethora of datasets and experiment setups involving various models, from which this work might benefit or at least draw inspiration.

5.3 Synthetic Text Data Generation

The value of synthetic data has been recognized by many works in the past and, not neglecting their risks and limitations, enables research to benefit greatly in many ways and forms. From earlier work (Foster and Andersen, 2009) up to more recent approaches including various data augmentations (Xie et al., 2017; Gao et al., 2019; Xia et al., 2019; Duan et al., 2020; Sánchez-Cartagena et al., 2021), to word representations in form of robust word vectors (Malykh, Logacheva, and Khakhulin, 2018) and robust embeddings (Doval, Vilares, and Gómez-Rodríguez, 2020).

Synthetic data has been successfully applied in Neural machine translation (Artetxe et al., 2018; Ngo et al., 2022; Bogoychev and Sennrich, 2020) but also utilized for lexical normalization (Dekker and van der Goot, 2020), script normalization (Ahmadi and Anastasopoulos, 2023) and text normalization for Ligurian (Lusito, Ferrante, and Maillard, 2022).

5.4 Linguistic Features, Dialectal Variations and Translation

A comprehensive overview of the history of modern artificial neural networks and their use in linguistic generalization can be found in Baroni (2019), with additional information regarding their compositionality.

The importance of handling dialectal variations has been stressed very recently in Demszky et al. (2021) where the authors based the learning of dialect feature detection for English variations on a small set of minimal pairs. They show that this approach can circumvent the need for large-scale annotated corpora, which are unavailable for many dialects (Demszky et al., 2021).

Recently, a series of publications originating in Stanford University, has revolved around this issue. Starting in 2022, Ziems et al., 2022 released VALUE, a challenging variant of GLUE in order to understand disparities in current models and to facilitate more dialect-competent NLU systems. VALUE expands on established benchmarks that contain only Standard American English (SAE), namely General Language Understanding Evaluation (GLUE)[1] (Wang et al., 2018) and SuperGLUE (Wang et al., 2019). In their initial release they construct rules for 11 linguistic features of African American Vernacular English (AAVE), and recruit fluent AAVE speakers to validate each feature transformation via linguistic acceptability judgments in a participatory design manner.

Subsequently, Ziems et al. (2022) lamented the lack of a systematic study on cross-dialectal model performance and aimed to fill this gap by expanding on VALUE (Ziems et al., 2022) by introducing Multi-VALUE (Ziems et al., 2023). Multi-Value expanded VALUE from 1 English language variant to 50 different dialects and from 11 linguistic features to a total of 189 unique linguistic features, ranging from 29 to 136 per language variety.

Eventually, DADA was released by Liu et al., a work building on Multi-VALUE (Ziems et al., 2023) and using five of their dialects as representatives. Namely: Appalachian English (AppE), Chicano English (ChcE), Colloquial Singapore English (CollSgE), Indian English (IndE), and African American Vernacular English (AAVE). In this work, they propose to adapt a trained language model via feature adapters, each corresponding to a linguistic feature. They train nearly 200 feature adapters and demonstrate their method on the 5 mentioned English dialects. This dynamic approach enables the reuse of the same feature adapter in various language variants, which is in accordance to he notion of flexible borders between dialects (see Section 7).

A rationale for these kind of approaches can also be found in the work of, Held et al. in the same year Held, Ziems, and Yang, 2023, where they argue that current approaches to improving dialect robustness, at least in English, have only focused on a single task at a time, which makes then less scaleable and severely limits their impact since language technology applications are increasingly more diverse and pervasive. This work aims to fill the gap of underexplored training to enable task-agnostic zero-shot transfer. In their work they use perturbations from Ziems et al., 2023 and 4 dialect variants of GLUE Wang et al., 2018 to empirically show the effectiveness of their proposed method.

All of this showcases that a current trend in research is to attempt the utilization of High-Resource Languages for their low-resource counterparts, especially, their dialectal variations. Even though the summarized related works from this section

[1] https://gluebenchmark.com/

mainly focus on English variants, they can still guide similar work on various other languages in similar settings.

5.5 Bilingual Lexicon Induction

NLP has seen the development of various methods to cross (or at least narrow) the gap between languages and to enable high quality machine translations. One such method, and possible focus of this work, is known as bilingual lexicon induction (BLI). The beginnings of BLI can be found in (Vulić and Moens, 2013; Vulić and Moens, 2015; Gouws, Bengio, and Corrado, 2016). (TODO: Search for earlier works)

Czarnowska et al., 2019 describe BLI as well established choice for evaluation of cross-lingual word embedding models.

In order to give an idea, of how a BLI setup can look like, Czarnowska et al., 2019's experiment description shown in the following extract:

"Given a list of N source language word forms $x_1, ..., x_N$, the goal is to determine the most appropriate translation t_i, for each query form x_i. In the context of cross-lingual embeddings, this is commonly accomplished by finding a target language word that is most similar to x_i in the shared semantic space, where words' similarity is usually computed using a cosine between their embeddings. The resulting set of (x_i, t_i) pairs is then compared to the gold standard and evaluated using the precision at k (P@k) metric, where k is typically set to 1, 5 or 10.[2] Throughout our evaluation we use P@1, which is equivalent to accuracy."

In recent work methods were presented for unsupervised BLI for data-imbalanced closely related language pairs (Bafna et al., 2023) and the use of BLI to improve the translation of out of vocabulary words in Low-Resource MT (Waldendorf et al., 2022).

Additionally, BLI has come into contact with large language models (LLMs) in order to utilize their often astounding capabilities. LLMs are now used for developing bilingual lexicons (Li, Korhonen, and Vulić, 2023), even going so far as to apply this approach to low-resource language varieties, such as German dialects (Artemova and Plank, 2023).

Approaches in research that span across many languages can bring forth massively multilingual data sets with the added benefit of providing a direct link between two languages and thereby avoiding the necessity to use English as a pivot language between them. This is especially advantageous, due to the English language being morphologically poor (TODO: Find good source comparing the morphology of languages), making it less suitable for the analysis of morphological generalization (Czarnowska et al., 2019). One such data set could be found in CogNet (Batsuren, Bella, and Giunchiglia, 2022).

[2]Precision at k represents how many times the correct translation of a source word is returned as one of its k nearest neighbours in the target language.

5.6 Machine Translation Evaluation

Numerous metrics and benchmarks have surfaced for evaluation of machine translation systems and their performance. At the time of this writing, it is still unclear, which of those would be most suited for the nature of this work- but nonetheless, the following displays a short overview of the options that will be explored further.

Beginning with the "classics", and often well established metrics:

- Bilingual evaluation understudy (BLUE) (Papineni et al., 2002), (Post, 2018)
- BLEURT (Sellam, Das, and Parikh, 2020)
- Metric for Evaluation of Translation with Explicit ORdering (METEOR) (Lavie and Agarwal, 2007)
- chrF (Popović, 2015)

To more recent methods:

- Crosslingual Optimized Metric for Evaluation of Translation (COMET) (Rei et al., 2020)
- General Language Understanding Evaluation (GLUE)[3] (Wang et al., 2018)
- SuperGLUE (Wang et al., 2019)
- IGLUE (Bugliarello et al., 2022a)
- Cross-lingual TRansfer Evaluation of Multilingual Encoders (XTREME) (Hu et al., 2020)
- XTREME-R (Ruder et al., 2021)
- BERTScore (Zhang* et al., 2019)
- PALI (Ahmadi, Agarwal, and Anastasopoulos, 2023)

And especially promising, the recently published approaches for evaluating dialect specific machine translation:

- VALUE (Ziems et al., 2022)
- Multi-VALUE (Ziems et al., 2023)
- FRMT (Riley et al., 2023)
- CODET (Alam, Ahmadi, and Anastasopoulos, 2023)

Finally, depending on the scope of this work, it might be feasible to include human-evaluation by native speakers of the targeted language varieties. This is often considered to be the ideal gold standard for machine translation evaluation, but can not be the main method of this work, due to resource and time constraints.

A selection of promising evaluation methods and frameworks identified in an initial review are shown in Figure 7.6.

[3] https://gluebenchmark.com/

TABLE 5.1: Publication list of probably most helpful items from initial literature review.

Topic/Focus	Reference	Title
Strategies to Enhance Translation Quality in Low-Resource Situations		
Back-Translation	(Sennrich, Haddow, and Birch, 2016)	Improving Neural Machine Translation Models with Monolingual Data
	(Edunov et al., 2018)	Understanding Back-Translation at Scale
	(Dou, Anastasopoulos, and Neubig, 2020)	Iterative Back-Translation, using TF-IDF to select relevant sentences) Dynamic Data Selection and Weighting for Iterative Back-Translation
Joint Training	(Zhang et al., 2018)	Joint Training for Neural Machine Translation Models with Monolingual Data
Adapters	(Bapna and Firat, 2019)	Simple, Scalable Adaptation for Neural Machine Translation
	(Pfeiffer et al., 2020)	MAD-X: An Adapter-Based Framework for Multi-Task Cross-Lingual Transfer[4]
Fine-Tuning Adapters	(Ansell et al., 2023a)	Composable Sparse Fine-Tuning for Cross-Lingual Transfer with a variant of the Lottery Ticket Hypothesis[5]
	(Cooper Stickland, Li, and Ghazvininejad, 2021)	Recipes for Adapting Pre-trained Monolingual and Multilingual Models to Machine Translation
Denoising Adapters	(Üstün et al., 2021)	Multilingual Unsupervised Neural Machine Translation with Denoising Adapters
Cross-Lingual Transfer	(Ansell et al., 2023b)	Distilling Efficient Language-Specific Models for Cross-Lingual Transfer
Sim. to Üstün	(Garcia et al., 2021)	Harnessing Multilinuality in Unsupervised Machine Translation for Rare Languages
Zero-Shot	(Lauscher et al., 2020)	From Zero to Hero: On the Limitations of Zero-Shot Language Transfer with Multilingual Transformers
	(Parović et al., 2022)	BAD-X: Bilingual Adapters Improve Zero-Shot Cross-Lingual Transfer[6]
Compressed Models	(Ansell et al., 2023b)	Distilling Efficient Language-Specific Models for Cross-Lingual Transfer
Massively Multilingual	(Team et al., 2022)	No Language Left Behind: Scaling Human-Centered Machine Translation[7][8]
Linguistically-grounded	(Casas et al., 2021)	Linguistic knowledge-based vocabularies for Neural Machine Translation
Transformer	(Vaswani et al., 2017)	Attention Is All You Need

Continued on next page

[4] https://adapterhub.ml/
[5] https://github.com/cambridgeltl/composable-sft
[6] https://github.com/parovicm/BADX
[7] https://github.com/facebookresearch/fairseq/tree/nllb
[8] https://huggingface.co/docs/transformers/model_doc/nllb

Table 5.1 – continued from previous page

Topic/Focus	Reference	Title
Neural Machine Translation	(Bandyopadhyay, 2023)	Factored Neural Machine Translation on Low Resource Languages in the COVID-19 crisis
Monolingual Data	(Karakanta, Dehdari, and Van Genabith, 2018)	Neural machine translation for low-resource languages without parallel corpora
	(Reimers and Gurevych, 2020)	Making Monolingual Sentence Embeddings Multilingual using Knowledge Distillation[9]
	(de Vries et al., 2021)	Adapting Monolingual Models: Data can be Scarce when Language Similarity is High
Synthetic Text Data Generation		
Script Normalization	(Ahmadi and Anastasopoulos, 2023)	Script Normalization for Unconventional Writing of Under-Resourced Languages in Bilingual Communities[10]
Lexical Normalization	(Dekker and van der Goot, 2020)	Synthetic Data for English Lexical Normalization: How close Can We Get to Manually Annotated Data?
Text Normalization	(Lusito, Ferrante, and Maillard, 2022)	Text normalization for endangered languages: the case of Ligurian
Grammatical Err. Det.	(Foster and Andersen, 2009)	GenERRate: Generating Errors for Use in Grammatical Error Detection
Word Embeddings	(Doval, Vilares, and Gómez-Rodríguez, 2020)	Towards robust word embeddings for noisy texts
	(Malykh, Logacheva, and Khakhulin, 2018)	Robust Word Vectors: Context-Informed Embeddings for Noisy Texts
Neural Machine Translation	(Bogoychev and Sennrich, 2020)	Domain, Translationese and Noise in Synthetic Data for Neural Machine Translation
Artificial translation units	(Ngo et al., 2022)	An Efficient Method for Generating Synthetic Data for Low-Resource Machine Translation
Swapping	(Artetxe et al., 2018)	Unsupervised Neural Machine Translation
Dropping	(Xia et al., 2019)	Generalized Data Augmentation for Low-Resource Translation
Replacing	(Gao et al., 2019)	Soft Contextual Data Augmentation for Neural Machine Translation
Dependency Parsing	(Xie et al., 2017)	Data Noising as Smoothing in Neural Network Language Models
Reversing sentences	(Duan et al., 2020)	Syntax-aware Data Augmentation for Neural Machine Translation

Continued on next page

[9] https://github.com/UKPLab/sentence-transformers
[10] https://github.com/sinaahmadi/ScriptNormalization

Table 5.1 – continued from previous page

Topic/Focus	Reference	Title
Mix-Source	(Sánchez-Cartagena et al., 2021)	Rethinking Data Augmentation for Low-Resource Neural Machine Translation: A Multi-Task Learning Approach
Copy source sentences	(Ha, Niehues, and Waibel, 2016)	Toward Multilingual Neural Machine Translation with Universal Encoder and Decoder
Zero-Shot	(Ye et al., 2022)	ZEROGEN: Efficient Zero-shot Learning via Dataset Generation
Bilingual Lexicon Induction		
BLI and Large Language Models	(Artemova and Plank, 2023)	Low-resource Bilingual Dialect Lexicon Induction with Large Language Models
	(Li, Korhonen, and Vulić, 2023)	On Bilingual Lexicon Induction with Large Language Models
Low-Resource Bilingual Lexicon Induction	(Waldendorf et al., 2022)	Improving Translation of Out Of Vocabulary Words using Bilingual Lexicon Induction in Low-Resource Machine Translation
	(Bafna et al., 2023)	A Simple Method for Unsupervised Bilingual Lexicon Induction for Data-Imbalanced, Closely Related Language Pairs
Morphological Generalization	(Czarnowska et al., 2019)	Don't Forget the Long Tail! A Comprehensive Analysis of Morphological Generalization in Bilingual Lexicon Induction
Cross-Lingual Word Embeddings	(Vulić and Moens, 2013)	Cross-Lingual Semantic Similarity of Words as the Similarity of Their Semantic Word Responses
	(Vulić and Moens, 2015)	Bilingual Word Embeddings from Non-Parallel Document-Aligned Data Applied to Bilingual Lexicon Induction
	(Gouws, Bengio, and Corrado, 2016)	BilBOWA: Fast Bilingual Distributed Representations without Word Alignments
Work on Linguistic Features, Dialectal Variations and Translation		
Linguistic Features	(Baroni, 2019)	Linguistic generalization and compositionality in modern artificial neural networks
Dialect Features	(Demszky et al., 2021)	Learning to Recognize Dialect Features
	(Liu, Held, and Yang, 2023)	DADA: Dialect Adaptation via Dynamic Aggregation of Linguistic Rules
Benchmark	(Ziems et al., 2022)	VALUE: Understanding Dialect Disparity in NLU[11]
	(Ziems et al., 2023)	Multi-VALUE: A Framework for Cross-Dialectal English NLP[12]
	(Riley et al., 2023)	FRMT: A Benchmark for Few-Shot Region-Aware Machine Translation

Continued on next page

[11] https://github.com/salt-nlp/value
[12] http://value-nlp.org/

Table 5.1 – continued from previous page

Topic/Focus	Reference	Title
Dialect-Adapters	(Held, Ziems, and Yang, 2023)	TADA: Task-Agnostic Dialect Adapters for English[13]
About Evaluation		
Translationese	(Bizzoni et al., 2020)	How Human is Machine Translationese? Comparing Human and Machine Translations of Text and Speech
Script Identification	(Ahmadi, Agarwal, and Anastasopoulos, 2023)	PALI: A Language Identification Benchmark for Perso-Arabic Scripts
Text Generation	(Zhang* et al., 2019)	BERTScore: Evaluating Text Generation with BERT[14]
Language Understanding	(Wang et al., 2018)	GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding[15]
Transfer Learning	(Bugliarello et al., 2022b)	IGLUE: A Benchmark for Transfer Learning across Modalities, Tasks, and Languages[16]
Cross-lingual Generalization	(Hu et al., 2020)	XTREME: A Massively Multilingual Multi-task Benchmark for Evaluating Cross-lingual Generalization[17][18]
About Benchmarking	(Kiela et al., 2021)	Dynabench: Rethinking Benchmarking in NLP[19]
Machine Translation	(Papineni et al., 2002)	BLEU: a method for automatic evaluation of machine translation[20]
	(Post, 2018)	A Call for Clarity in Reporting BLEU Scores
	(Lavie and Agarwal, 2007)	METEOR: An Automatic Metric for MT Evaluation with High Levels of Correlation with Human Judgments[21]
	(Rei et al., 2020)	COMET: A Neural Framework for MT Evaluation[22]
	(Popović, 2015)	chrF: character n-gram F-score for automatic MT evaluation[23]
	(Snover et al., 2006)	A Study of Translation Edit Rate with Targeted Human Annotation[24]
	(Alam, Ahmadi, and Anastasopoulos, 2023)	CoDET: A Benchmark for Contrastive Dialectal Evaluation of Machine Translation
	(Ruder et al., 2021)	XTREME-R: Towards More Chellenging and Nuanced Multilingual Evaluation

[13] Soon: https://github.com/boschresearch/ACL23-TADA
[14] https://github.com/Tiiiger/bert_score
[15] https://gluebenchmark.com/
[16] https://github.com/e-bug/iglue
[17] https://sites.research.google/xtreme
[18] https://github.com/google-research/xtreme
[19] https://dynabench.org/
[20] https://github.com/bangoc123/BLEU
[21] http://www.cs.cmu.edu/~alavie/METEOR/
[22] https://github.com/Unbabel/COMET
[23] https://github.com/m-popovic/chrF
[24] https://github.com/jhclark/tercom

Chapter 6

Related Work & State of the Art

Chapter 7

Methodology

7.1 The Importance of Dialects

To build NLP systems for not only the dominant languages, but also for less digitally present languages and language variations, required to account for the numerous differences between them. To overlook a dialect is the same as overlooking an entire language community, which, for the NLP landscape to be fair, should not happen. Demszky et al. (2021) put emphasis on the notion of flexible boarders between dialects. They declare dialects not to be monolithic entities, but rather to have distinctions which can be measured by the presence, absence, and frequency of numerous linguistic features found in speech but also in text (see Figure 7.1). These linguistic features can be shared by multiple dialects (see Figure 7.2) and encountering one such feature in a text does not necessarily guarantee the text to be of a specific dialect.

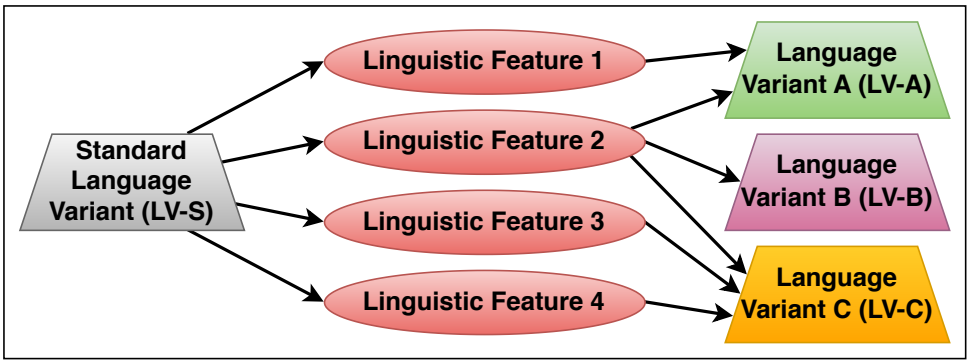


Figure 7.1: The standard variant of a language, modified by certain linguistic features, takes on the form of a language variation- or dialect.

7.2 Methods Suitable for Sparse Data & Resources

It has to be assumed, that for many languages, there will not suddenly be a surplus of data available in the near future. To make matters worse, many researchers that apply themselves to low-resource languages have to struggle with very limited resources in the sense of computational power and infrastructure, often taking place close to the language community in question. This makes the elaborate training of large models impractical, which is another guiding factor in the selection of techniques and methodology for this work.

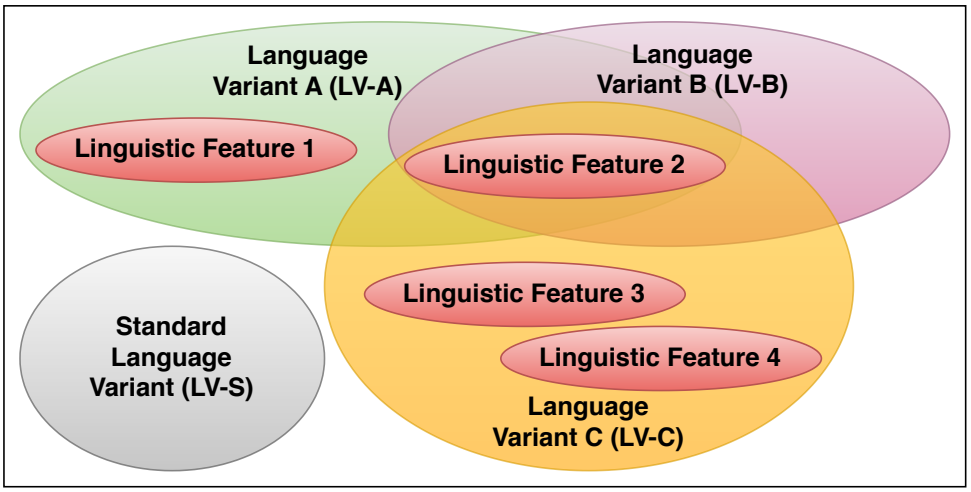


FIGURE 7.2: Dialects are not monolithic i.e., being discretely distinguishable classes with clearly defined boarders, but can have fluent transitions and overlap between each other.

7.3 Synthetic Data Generation & Dynamic Model Adaptation via Linguistic Features

The idea is to work on a selected number of dialects from low-resource languages for which at least some data in the form of the standard language exists and is openly available. For each of these dialects, a set of linguistic rules will have to be identified, either from prior research or as part of this work, that codify their creation based on the standard variant of the corresponding language.

Figure 7.3 shows how the set of linguistic features of a language variety might be used to generate synthetic data based on text data from the languages' standard variant (similar to Ziems et al., 2023), while the same features, but separately processed, can enable the training of feature-specific adapters for the use in training language models (similar to Liu, Held, and Yang, 2023).

7.4 Machine Translation on the Basis of Synthetic Data

Linguistic rules will enable the creation of text data which aligns with the dialectal variant of said language. Using this trove of novel data, a translation model will be trained with the aim to display a higher dialectal robustness than previously available for said language variant (see Figure 7.4). Prior to training, it is crucial to evaluate the quality of the synthetic data. One way to do this is via acceptance checking from native speakers, similar as done by Ziems et al. (2022) to validate this phase of the work.

7.5 Target Languages considered for this Work

An overview of the currently considered target languages to be included in this work can be seen in Figure 7.5.

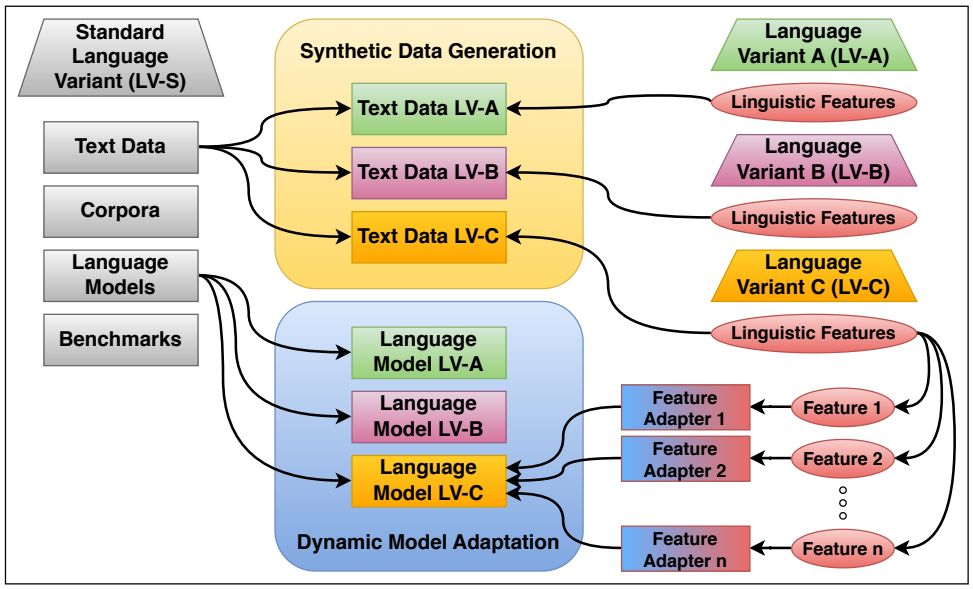


FIGURE 7.3: Utilization of the resources that a language standard variation brings with it in order to benefit the low-resource variants or dialects by generating synthetic text data (yellow) or deriving pretrained language models (blue) based on the variant's characteristic linguistic features (red).

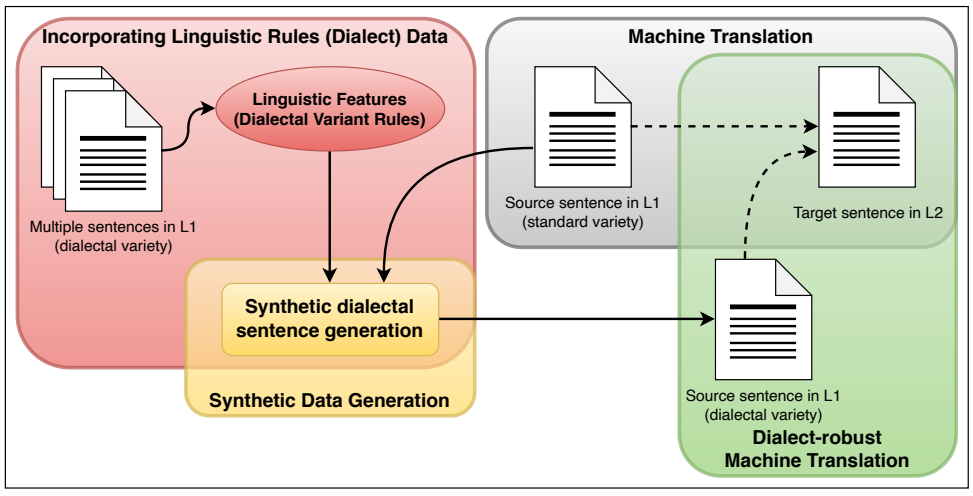


FIGURE 7.4: Concept of this work for arriving at dialect-robust machine translation for low-resource language variants via synthetic data generation.

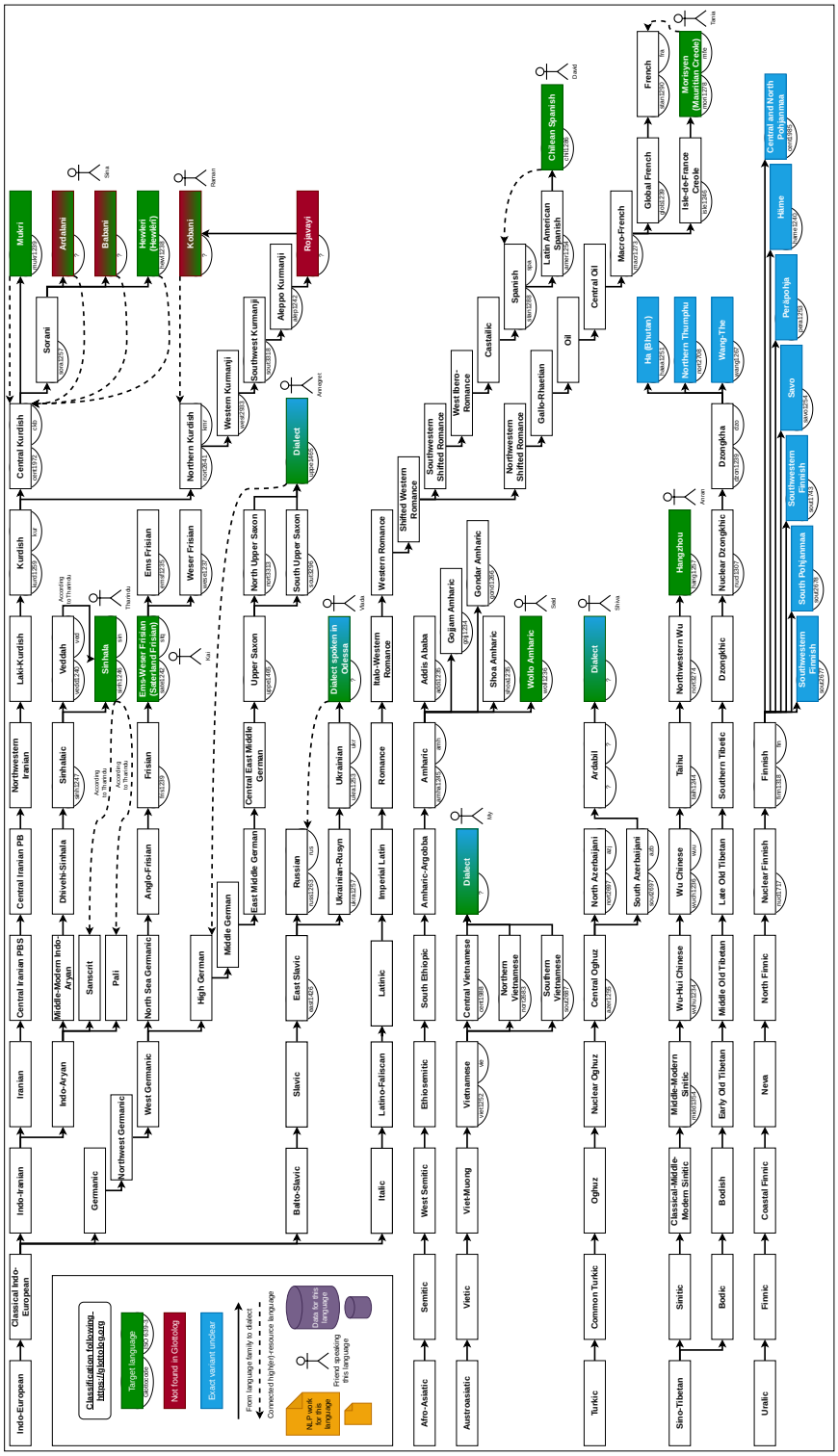


FIGURE 7.5: Considered languages and dialects with their position in the corresponding family tree. Considerations are currently mainly based on the author's access to native speakers from the language varieties. In a next step the availability of language data will have to be considered to reach a final language selection.

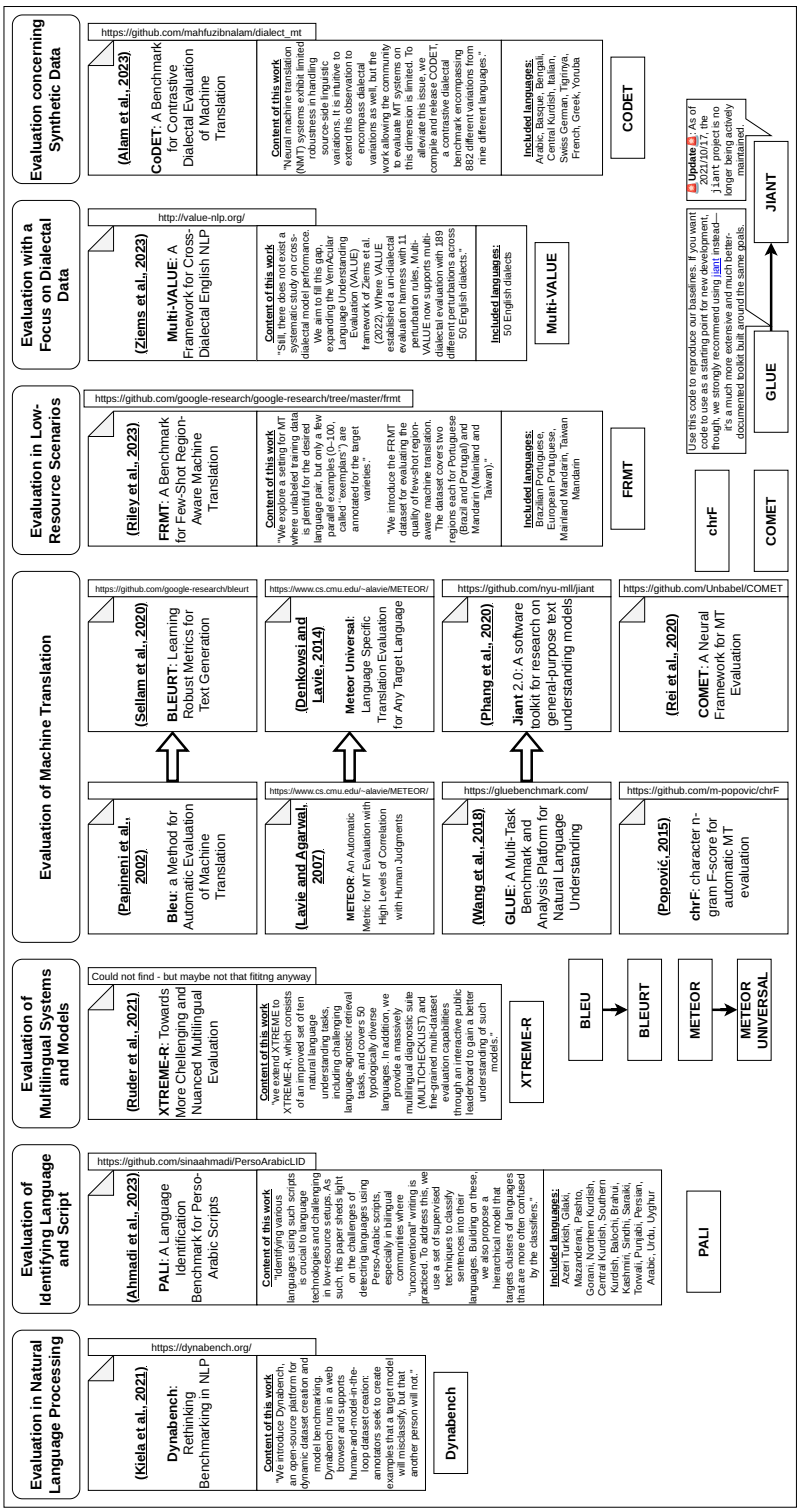


FIGURE 7.6: Evaluation methods and frameworks considered for this work.

Chapter 8

Data

8.1 Available Data by Language

German

List with 59 prefixes and 63 suffixes of German [1].

Alemannic Bavarian Northern Bavarian Central Bavarian Southern Bavarian Franconian Saxon Swiss German

Sanandaji Mukriyani

Northern Kurdish Kobani

Copula Example Kurmanji:

Kobani:

8.2 Available Data

No	Corpus	Language	Size
1	Tanzil	Kurdish–English	92.354 texts
2	Ted	Kurdish–English	2.358 texts
3	KurdNet	Kurdish–English	4.663 texts
4	Auta	Kurdish–English	100.000 texts
Total			199.375 texts

Table 8.1: A Transformer-based Neural Network Machine Translation Model for the Kurdish Sorani Dialect Badawi, 2023
Size of each Kurdish–English corpus.
Publicly available are only 100.000 En-Ku pairs.

[1] https://www.prosperosisle.org/spip.php?article1001

Chapter 9

Exploration & Experiment

The main part of this work is separated into the following two phases:

9.1 Exploratory Experiments

Multiple factors can be identified which directly affect the scope and success of this work and which have to be explored for this work's efforts to bear fruit. The viewed literature provides a number of approaches and different methods that could be beneficial to combine with each other in order to end up with an improved translation model for Low-Resource languages. In the plethora of Low-Resource languages, there are some for which ambitious researchers have already provided at least small amounts of data which might be applicable in this work's context. Which and to what degree will have to be identified prior to the main experiments.

This is also the point in time, at which the contents of the methodology chapter will be validated or, at least partially, be revised.

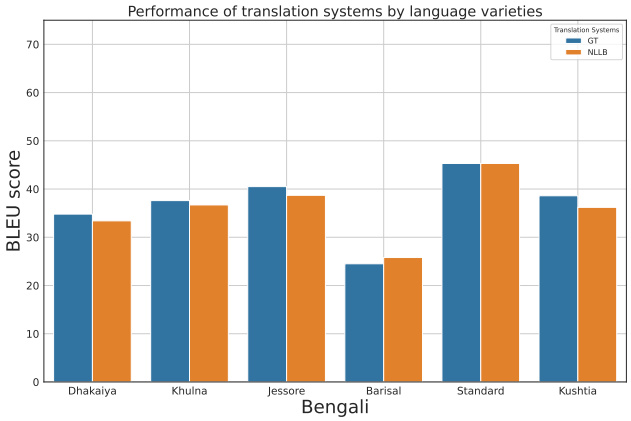
9.2 Main Experiments

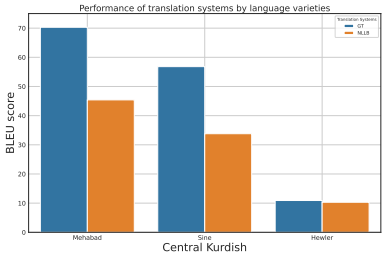
Once the most promising course of action and its corresponding compartments have been identified, the actual "main experiments" can take place. This includes not only the incorporation of linguistic rules to create synthetic data and a subsequent training of translation models, but also a meticulous investigation of how these models perform. If possible, directly compared to state-of-the-art applications from related work, but definitely based on established metrics. Which metrics, will depend on the insights gained from the exploratory experiments and suitability to the eventually resulting models.

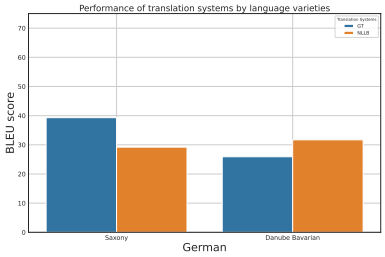
Chapter 10

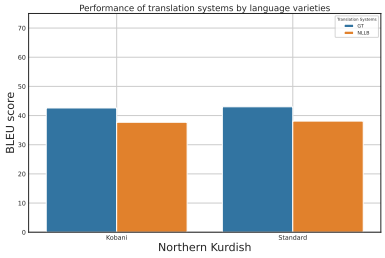
Proof of Concept

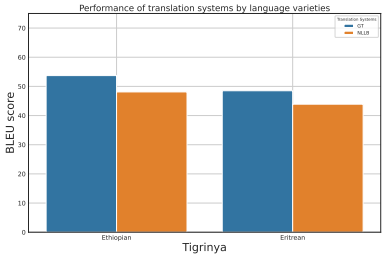
10.1 Summary of Experiment Results











These exploratory experiments utilized the translation systems Google Translate[1] (via the translate-shell[2]) and NLLB (Team et al., 2022), various data sets (primarily CODET (Alam, Ahmadi, and Anastasopoulos, 2023)), and the implementation of BLEU from Fairseq[3]. The translation of text data from language varieties that can

[1] https://translate.google.com/
[2] https://github.com/soimort/translate-shell
[3] https://github.com/facebookresearch/fairseq

be considered to be low-resourced, suffers performance-wise as is shown in Figure 10.1 (and in more detail further below).

10.2 Selected Data sets and the Origin of their Data

10.2.1 Origin of Bengali data

Taken from Alam, Ahmadi, and Anastasopoulos, 2023:

> "Our approach involved initially gathering 200 standard Bengali sentences from the Bengali-English translation dataset presented in Hasan et al. (2020), a high-quality dataset comprising 2.75 million parallel sentence pairs.
> From this dataset, we selected short sentences comprising 6 to 7 words, facilitating ease of translation for the language speakers. Initially, there were 200,000 sentences to choose from, and we randomly selected 200 sentences for our dataset.
> Our initial step involved recruiting proficient annotators fluent in the standard and in one of the dialects. Subsequently, we requested these annotators to provide their respective dialectal renditions of specific sentences.
> Given that dialects primarily exist in spoken form without standardized orthography, we instructed the annotators to transcribe the sentences in Bengali script based on the acoustic signals they perceived.
> This process is called dialectal writing (Nigmatulina, Kew, and Samardzic, 2020), which entails creating phonemic transcriptions that closely align grapheme labels with the acoustic signals, despite their inherent inconsistency.
> This approach, in our view, mimics what speakers of the varieties would do should they attempt to write them."

10.2.2 Origin of Central Kurdish data

Taken from Alam, Ahmadi, and Anastasopoulos, 2023:

> "Given that there are no corpora documenting varieties of Central Kurdish, we resort to movies where speakers of these varieties play a role.
> To that end, we transcribe movies in Babanî, Ardalanî, and Mukriyanî.
> Since none of these movies are available in other varieties, we perform a dialect translation by a native speaker of Ardalanî and Mukriyanî by randomly selecting and translating 300 sentences in Babanî transcriptions.
> To mitigate the impact of orthography on the dialect, we normalize and standardize the sentences based on the common orthography of Kurdish using KLPT (Ahmadi, 2020)."

10.2.3 Origin of German data

A short fable in three language varieties, provided by the authors of Alam, Ahmadi, and Anastasopoulos (2023) online[4].

10.2.4 Origin of Northern Kurdish data

The text data for Northern Kurdish (Kurmanji) I collected from various data sets: To acquire the Kobani data, a set of transformations where applied on the Kurmanji

- (Ahmadi et al., 2023) [5]
- (Ahmadi et al., 2022) [6]
- (Ahmadi, 2020) [7]
- (Fatihkurt, 2020) [8]
- (Esmaili & Salavati, 2013) [9]
- (Haig, 2001) [10]

data.

[4] https://github.com/mahfuzibnalam/dialect_mt/tree/main/German

10.2.5 Origin of Tigrinya data

Taken from Alam, Ahmadi, and Anastasopoulos, 2023:

> "To explore and compare these two, we leverage the dataset available from TICO-19 (Anastasopoulos et al., 2020).
> The TICO-19 dataset emerged as a translation initiative during the COVID-19 pandemic, aiming to enhance society's readiness to respond to the ongoing crisis through the utilization of translation technologies effectively.
> This dataset specifically focuses on the COVID-19 domain, containing translations of the same content in multiple languages.
> The same 3071 English sentences were professionally translated into both varieties of Tigrinya, making it ideal for our purposes."

10.3 Formal (Exploratory) Evaluation

Notes regarding notation

- "Standard" variety is the dominantly used language variety.
- "Reference" variety identifies the (in the used data sets provided) translation, which should be considered to be on "Gold Standard" level.
- In cases where a "Reference" exists, all language varieties are compared against it for calculating BLEU scores. (Northern Kurdish, Bengali, Tigrinya)
- If there was no such reference data, the provided "Standard" variety was used to generate translations, against which the remaining language varieties where tested. (Central Kurdish, German)

10.3.1 Setup & Pipeline

The current setup includes the two translation systems Google Translate and NLLB for translating the text data into English. The evaluation is done with Fairseq's implementation of BLEU. At the current time, many parts of the pipeline displayed in Figure 10.1 still have to be done manually and should be automated in the comming weeks.

10.3.2 Initial Results

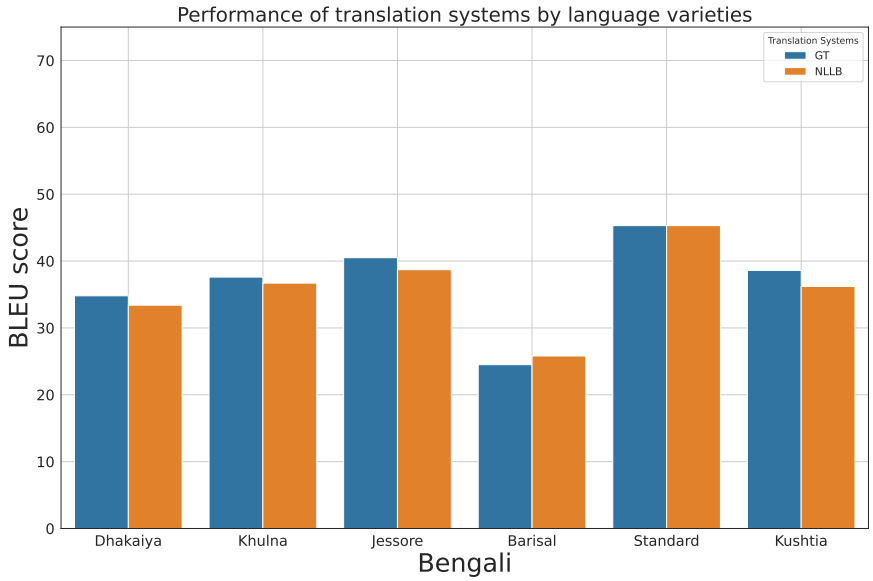
Translating the available text data resulted in new text documents for which the respective word counts can be found in Table 10.1 and further evaluation follow below.

Language	Variety	Words (Source)	Words (GT)	Words (NLLB)
Northern Kurdish	Reference	34178	34178	34178
Northern Kurdish	Standard	34856	36563	33670
Northern Kurdish	Kobani	34856	36673	33765
Bengali	Reference	1503	1503	1503
Bengali	Standard	1297	1426	1400
Bengali	Barisal	1321	1511	1641
Bengali	Dhakaiya	1284	1456	1415
Bengali	Jessore	1295	1414	1425
Bengali	Khulna	1300	1406	1426
Bengali	Kushtia	1292	1430	1411
Central Kurdish	Standard	1855	2199	2145
Central Kurdish	Mahabad	1855	2235	2392
Central Kurdish	Sine	1880	2075	2492
Central Kurdish	Hewler	1824	2166	2362
German	Standard	108	115	108
German	Danube Bavarian	110	113	115
German	Saxony	108	105	119
Tigrinya	Reference	70570	70570	70570
Tigrinya	Eritrean	66416	67329	68095
Tigrinya	Ethiopian	65165	65880	66046

Table 10.1: Word counts of used data per language and their varieties respectively.

10.3.3 Bengali

BLEU scores for Bengali



The CODET project [11] has Bengali text data that was collected from various

[11] https://github.com/mahfuzibnalam/dialect_mt

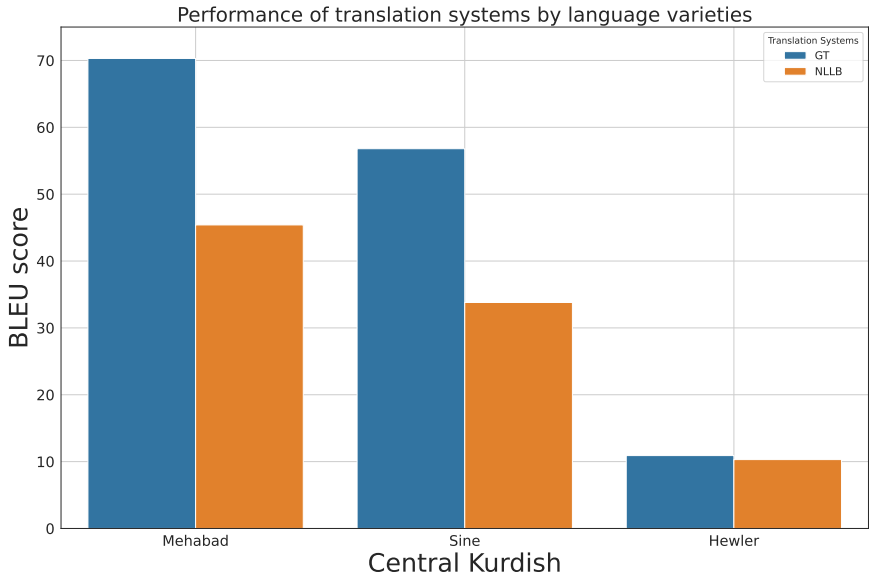
locations of Bangladesh: "This work specifically focuses on five prominent dialects from five locales of Bangladesh: Jessore, Khulna, Kushtia, Barisal, and Dhaka. The selection of these dialects was strategic, encompassing regions both close to the origin of standard Bengali (Jessore, Kushtia) and those situated farther away." (Alam, Ahmadi, and Anastasopoulos, 2023)

Based on this assessment, it could be expected that current translation systems are better able to handle the data from Jessore and Kushtia, but showing worse performance for Khulna, Barisal and Dhaka. Examining the results (see Figures 10.3.3), confirms this expectation.

Noteworthy might be the stark difference in the varieties' word counts compared to the translation, which is about 200 words (more than 13%) (refer to Table 10.1).

10.3.4 Central Kurdish

BLEU scores for Central Kurdish



In Alam, Ahmadi, and Anastasopoulos (2023) the authors describe the variant of Sine (also called Sulaimanî or Babanî) to be "the most studied one, which is also widely used as a standard variant of Central Kurdish in the press and media" (Alam, Ahmadi, and Anastasopoulos, 2023). In these initial experiments though, this variant only got to be second best performing (for Central Kurdish) and this with quite some distance to the first place (see Figure 10.3.4).

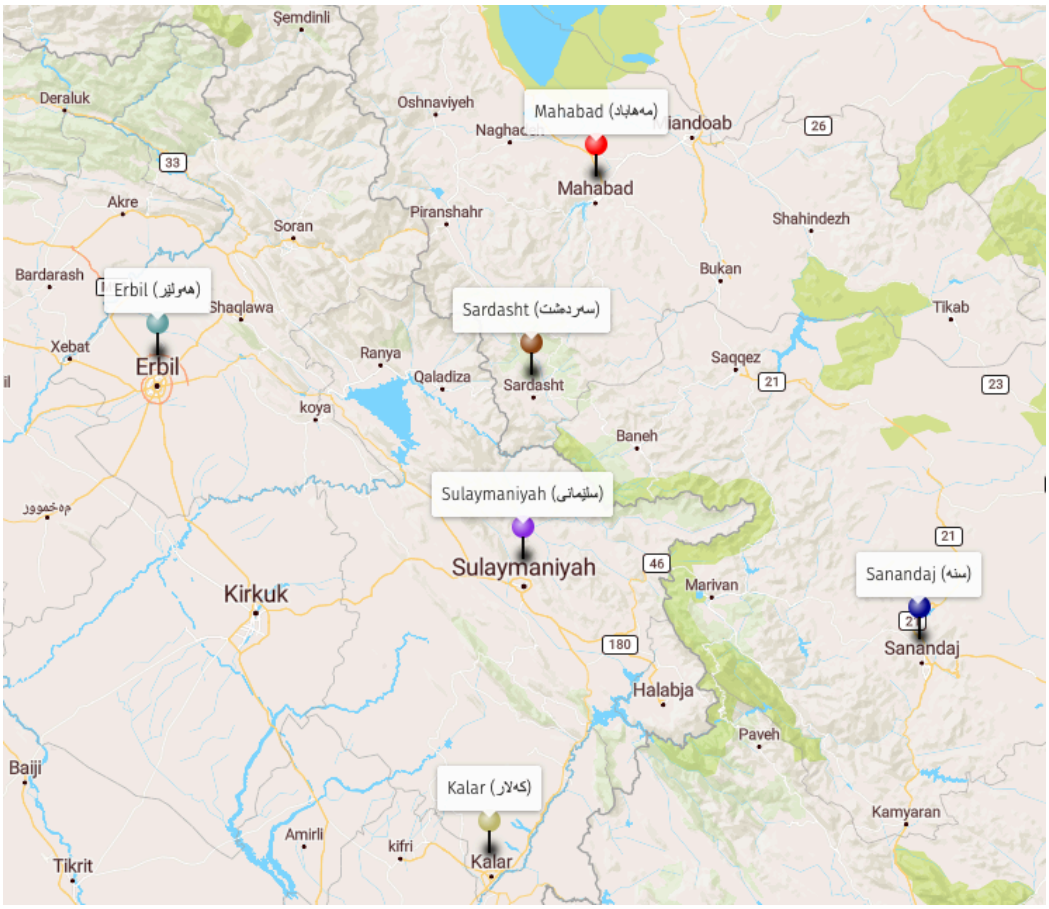
The best performing language variety for Central Kurdish is currently Mahabad (also called Mukrî or Mukriyanî) which has simply been described as one of many dialects, what makes this result a bit surprising.

Finally, there is the variety Hewler (also called Hewlêrî or Erbil), which shows the most perplexing results, with the BLEU scores being as low as they are. Hewler, next to Sine, has been described to be the epicenter of Central Kurdish language varieties: "According to various linguistic analyses of fieldwork data, Matras (2019) classifies Central Kurdish varieties into Northern and Southern Sorani, with their epicenters being based on the dialects of Erbil (Hewlêr in Kurdish) and Sulaymaniyah (Silêmanî in Kurdish)." (Alam, Ahmadi, and Anastasopoulos, 2023) A possible explanation could be, if there exists a massive imbalance between Northern and Southern Sorani (Central Kurdish) (the inverse of the scenario displayed by the, geographically closely located, Azerbaijani language, where the Southern variety is disproportionately lower-resourced than the Northern variety).

TODO: Figure out the true reason why Hewler performs so much worse than the other two varieties.

While the word count of the Hewler variety is comparatively low (refer to Table 10.1), the length of the generated translations is pretty close to the ones of the other varieties.

Central Kurdish text data locations from CORDI.

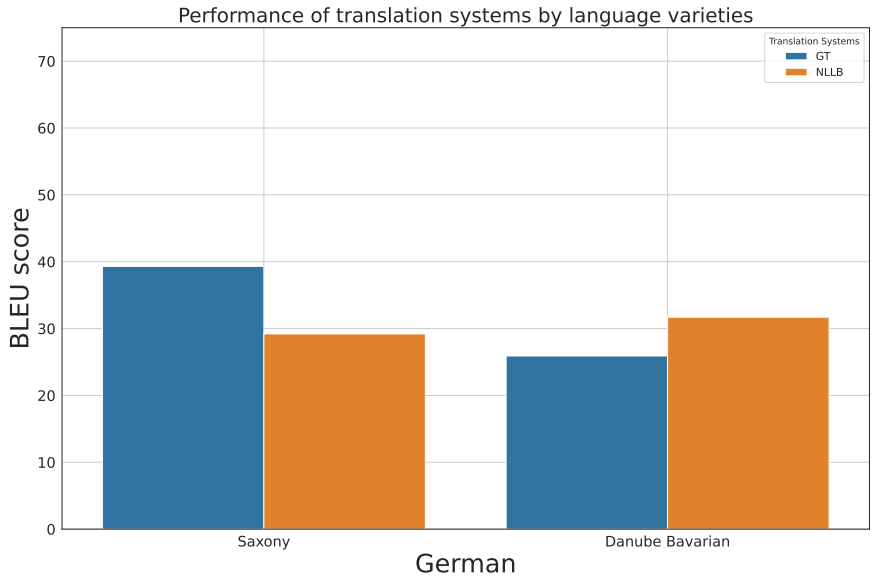


10.3.5 German

TODO: Figure out what the reason might be, that one translations systems handle these two varieties so differently well.

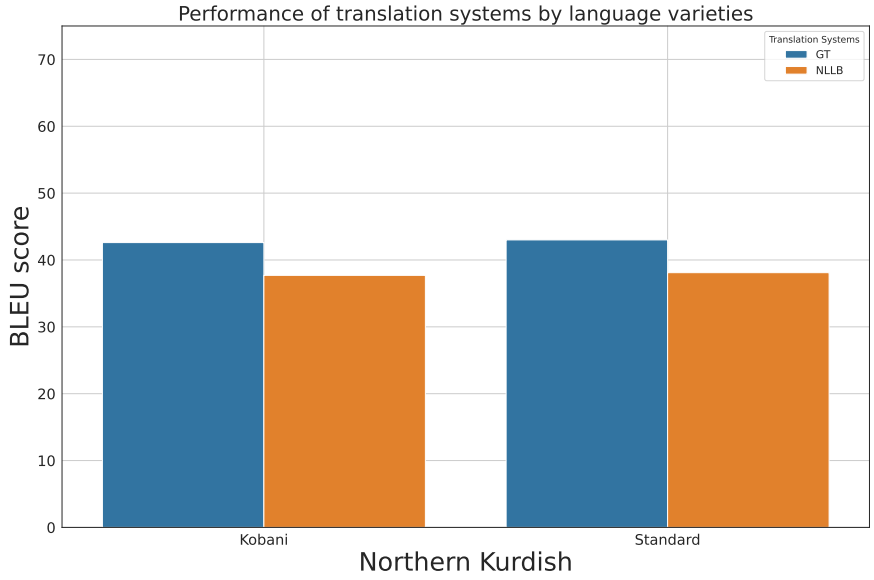
For exact word counts of the language varieties, refer to Table 10.1.

BLEU scores for German



10.3.6 Northern Kurdish

BLEU scores for Northern Kurdish



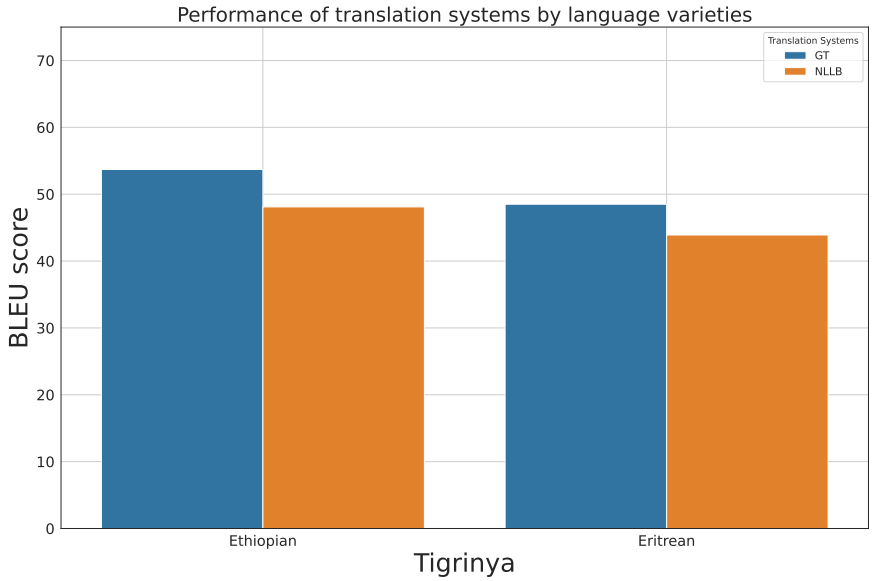
These similar results very much fit the expectations, due to the Kobani dialect not only being quite close to the standard Kurmanji, but also based on the set of linguistic rules used for data transformation, still being developed and relatively limited.

The used data of the Kobani variety has the exact same number of words as the standard variant (refer to Table 10.1), due to the way it has been created. Nevertheless, the translation systems created translations of slightly varying length, indicating

that these two language varieties are different enough difference to affect the translation system's processes.

10.3.7 Tigrinya

BLEU scores for Tigrinya



TODO: Investigate the details of the digital presence for both language varieties respectively.

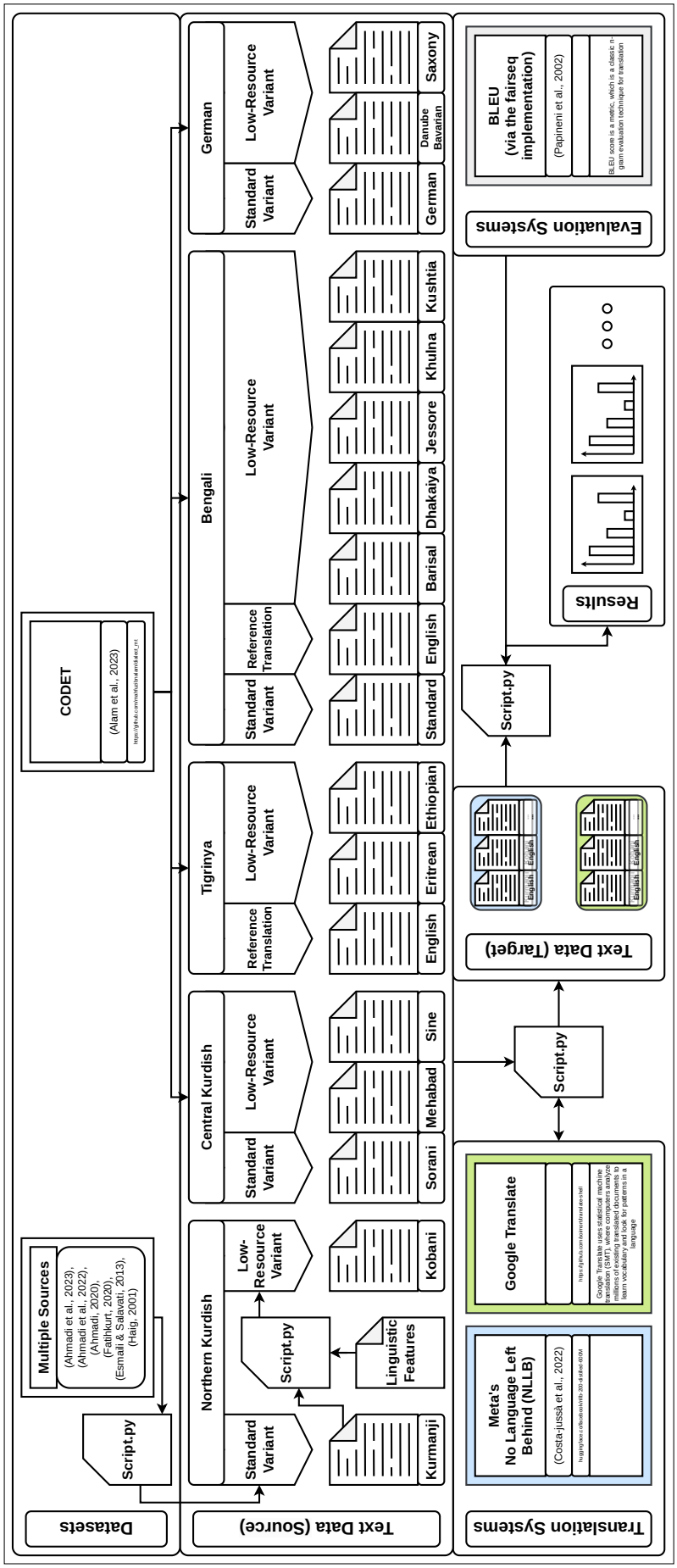


Figure 10.1: Exploratory experiment pipeline schema.

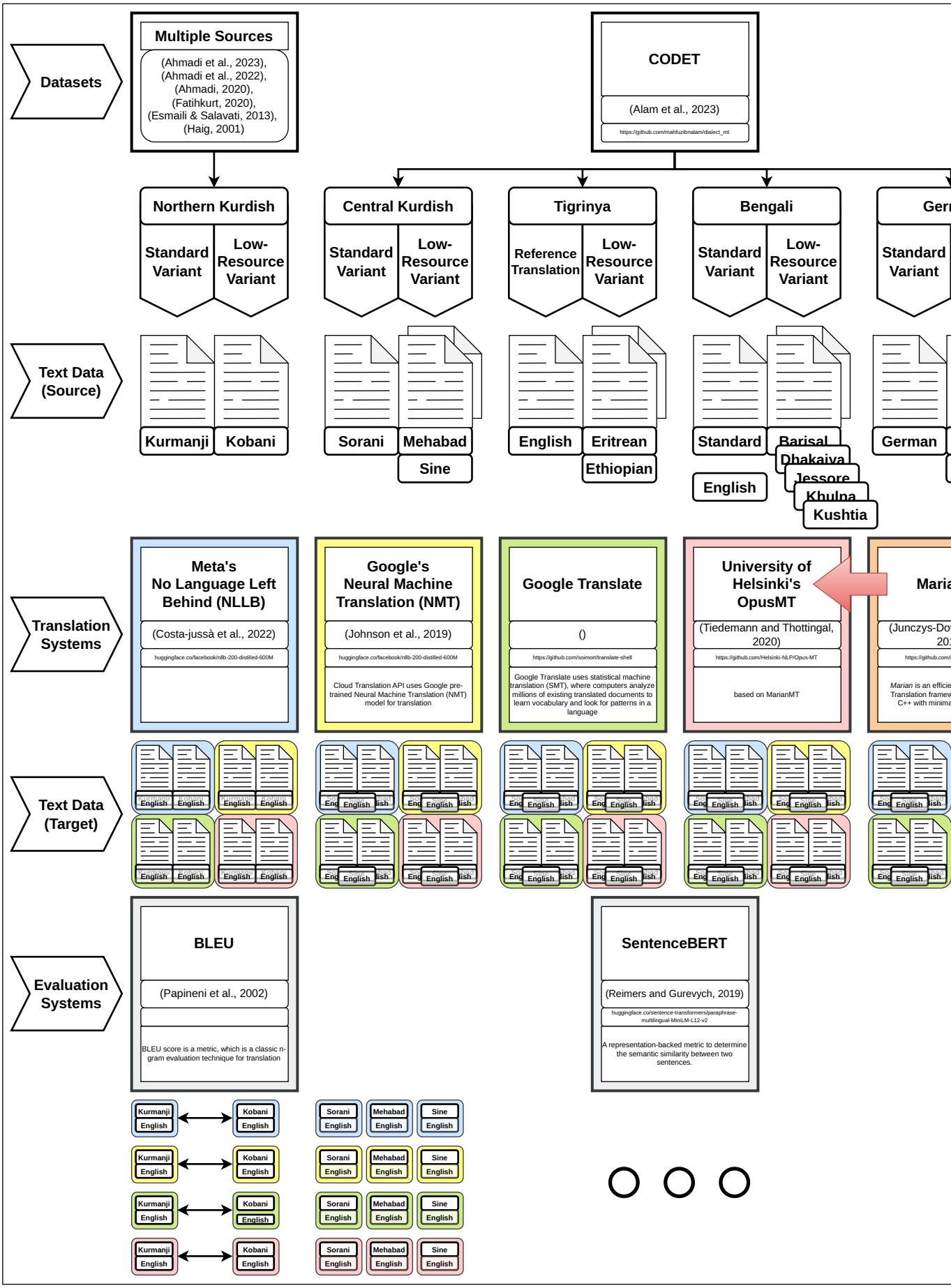


FIGURE 10.2: Overview of the short-term plans for the experiment's pipeline components.

Chapter 11

Experiment

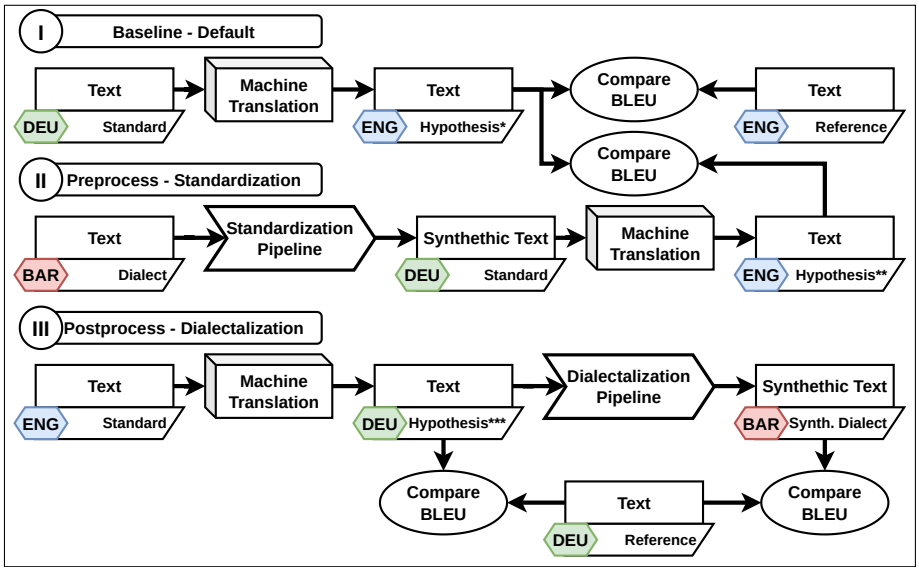


Figure 11.1: Experimental setup of three approaches.

If not otherwise specified, the default parameters of each tool/framework were applied.

German Data	Standard Baseline	→ Variety Postprocess	Variety Baseline	→ Standard Preprocess	#Bidict
Bar-01	"Gold-stand-in"	81.0	48.5	49.7	
Bar-02	"Gold-stand-in"	80.3	46.4	48.3	
Als-01	"Gold-stand-in"	93.4	40.7	39.0	
Als-02	"Gold-stand-in"	90.0	41.3	39.1	

Table 11.1: Initial machine translation of German varieties via NLLB based on the data from DialectBLI.

Argo[1] NLLB[2] Aper[3] Translate Shell[4] Goog[5] Aspe[6] Spel[7] Huns[8] Bing[9] Yand[10]

Model	Src	Trg	BLEU	chrF2	TER
Data: NLLB Seed Devtest					
Argo	Eng	Deu	31.75	60.49	53.76
NLLB	Eng	Deu	34.24	61.58	52.00
Aper	Eng	Deu	0	0	0
Via Translate Shell					
Goog	Eng	Deu	43.39	68.87	43.70
Aspe	Eng	Deu	1.25	21.72	102.49
Spel	Eng	Deu	1.25	21.72	102.49
Huns	Eng	Deu	0	0	0
Bing	Eng	Deu	0	0	0
Yand	Eng	Deu	0	0	0
The scores are shockingly low: I looked into the translated texts and they sound perfectly fine! They just do not use the exact same words that were used in the test-sentences.					

Table 11.2: Machine translation from English to German.

(A): Console output as comment following this line. Something about a "TracerWarning" and "codegen" failing? (B): Console output as comment following this line. Something about the text lines not being de-tokenized which might hurt the scores?

[1] Argos Translate: https://github.com/argosopentech/argos-translate
[2] No Language Left Behind: https://github.com/facebookresearch/fairseq/tree/nllb
[3] Apertium: https://www.apertium.org/index.eng.html#?dir=eng-epo&q=
[4] https://github.com/soimort/translate-shell
[5] Translate Shell (Google Translate): https://github.com/soimort/translate-shell
[6] GNU Aspell: http://aspell.net/
[7] Spell: Is this related to ASpell and/or HunSpell ???
[8] Hunspell: http://hunspell.github.io/
[9] Bing Translator: https://www.bing.com/translator
[10] Yandex.Translate: https://translate.yandex.com/

Model	Src	Trg	BLEU	chrF2	TER
Data: NLLB Seed Devtest					
Argo	Deu	Eng	36.30	63.56	48.30
NLLB	Deu	Eng	42.20	66.72	43.14
Aper	Deu	Eng	0	0	0
Via Translate Shell					
Goog	Deu	Eng	48.18	71.21	37.70
Aspe	Deu	Eng	0	0	0
Spell	Deu	Eng	0	0	0
Huns	Deu	Eng	0	0	0
Bing	Deu	Eng	0	0	0
Yand	Deu	Eng	0	0	0
A lot of the systems/models do not want to work: Maybe for another time then.					

Table 11.3: Machine translation from German to English.

Src	Trg	Method	Checkpoints	BLEU	chrF2	TER
Train: 26141 Dev: 1089 Test: 179						
Data: DialectBLI, Cleaned; Sockeye: Default parameter						
Bar	Deu	Default	9 (6 hrs)	65.0		
Deu	Bar	Default	9 (6 hrs)	56.7		
Data: DialectBLI, Cleaned; Sockeye: Smaller setup						
Bar	Deu	Default	16 (60 min)	65.9		
Deu	Bar	Default	8 (30 min)	62.0		
Main difference is halfing the model size: Quite the speed improvement!						
Data: OpusTools, Cleaned; Sockeye: Smaller setup						
Bar	Deu	Default	1 (15 min)	56.1	75.5	30.6
Deu	Bar	Default	2 (20 min)	50.0	74.8	34.1
More training data decreased the performance- Cleaning not yet as good as hoped?						
Data: DialectBLI, Cleaned, Pivoted-NLLB-German; Sockeye: Smaller setup						
Bar	”Eng”	Hyperparameters: 256	15 (10 min)	43.4	61.0	46.7
Bar	”Eng”	Hyperparameters: 512	15 (10 min)	44.5	61.5	45.8
Bar	”Eng”	Hyperparameters: 1024	5 (5 min)	39.5	56.8	49.5
Translating into ”English” (above) worked better than into Bavarian (below). But why does the ”larger” (1024) Bar-Eng model get worse, while the larger Eng-Bar still improves compared to the smaller ones?						
”Eng”	Bar	Hyperparameters: 256	8 (5 min)	35.0	53.4	53.7
”Eng”	Bar	Hyperparameters: 512	6 (5 min)	36.5	55.5	51.7
”Eng”	Bar	Hyperparameters: 1024	5 (10 min)	38.0	55.7	50.4
256, 512, and 1024 stand for the applied Sockeye hyperparameters ”–transformer-model-size”, ”–transformer-feed-forward-num-hidden”, and ”–num-embed”. The smaller settings took a moment longer, but each of them was blazingly fast! And the evaluation scores look very good too... What am I missing here? Only noticed peculiarities worth following up: (A) and (B).						
Data: OpusTools, Cleaned, Perturbed; Sockeye: Smaller setup						
Bar	Deu					
Deu	Bar					
Data: OpusTools, Cleaned; Sockeye: (Her & Kruschwitz) setup						
Evaluation here on validation due to setup						
Kmr	Deu	5-fold-CV	1 (5 min)	30.34	53.53	
Kmr	Deu	5-fold-CV	1 (5 min)	28.67	52.51	
Kmr	Deu	5-fold-CV	1 (5 min)	30.78	52.92	
Kmr	Deu	5-fold-CV	1 (5 min)	30.36	52.14	
Kmr	Deu	5-fold-CV	1 (5 min)	29.60	52.57	
How come ”Validation-perplexity has not improved for 3 checkpoints” right off the bat?						

Table 11.4: Training my own models in Sockeye. Shown are the number of sentences in Train-, Dev-, Test-Sets.

Chapter 12

Results

12.1 Variables and Parameter

12.1.1 Language Variety (Granularity Level)

The degree of data scarcity vastly differs between language varieties (refer to recent alam paper) which in turn limits the quality and amount of data and linguistic rules that can be derived from it. The choice of language variety can be expected to have the greatest impact on the performance of the entire pipeline / this approach in it's entirety.

Coarse as granularity level describes the (currently used) data as it was provided from various sources and labeled on language level (German, Kurdish, English, ...) and sometimes dialect level (Alemannic, Bavarian, Kurmanji, Sorani, ...). As to be expected, this data contains a lot of noise and can be compared to a pot of soup with many cooks, all demanding to use their most favorite spices.

Fine denotes a (planned) granularity level in which the available data is separated by sub-dialects. For example the text data taken from wikidumps come from articles which sometimes contain dialect tag (up to three level deep and usually describing the area of the speaker and not necessarily a commonly used language (variety) name), by which the data can be sorted. The vast majority of Wikipedia articles found in the wikidumps do not have such tags though, which would require an initial dialect-identification to take place.
Note: I manually collected wordlists for a set of sub-dialects and tried to automatically tag the remaining articles. Sadly the available wordlists (found on Wikipedia, scientific literature, and other websites) contain words which are less frequently found in Wikipedia articles, such as personal pronouns (I, You, She, He, ...), resulting in almost no labels designated to articles.
Note: I already sorted the articles from the Bavarian-Wikipedia by their officially assigned dialect-tags. I then created frequency dictionaries for the words contained in the set of articles for each sub-dialect. Next I plan to filter those (such as by dropping all words with less than 5 occurrences) and then use these to designate a set of dialect-tag-labels to each article. These will function as tag-candidates. The idea is to find a reasonable heuristic to decide upon the correct dialect-tag for each article. If for example, an article contains 200 words and 15 of those are found in the frequency dictionary of dialect A, 10 are found in dialect B and 5 are from dialect C, then to use the dialect-tag of A for this article might seem prudent.
BUT: Imagine now, that the 5 words from C are the 5 most frequently used words in the C-dialect, which **never** find any use in any of the other dialects. And the 15 words from A are the least frequent words from its dialect, but simply managed to

be included here, due to dialect A having many more articles to draw words from, than dialect C.
Ideally the number and rarity of words (inside the to-be-tagged article, but also inside the frequency dictionaries (and their corresponding source articles, weighted according to their distribution across all other dialect-tagged-articles to draw from) would be considered for weighting their importance on the matter of dialect-identification.

12.1.2 Data Quality

Naive denotes data that has been collected via means such as opustools which encompasses a range of different text corpora of varying degrees of quality. This data contains a lot of noise. Sentences that a ill-aligned, sentences that are of low quality, and even text from very different languages. (Show Examples: Corpora in which the aligned sentence is the same in both languages AND sometimes Chinese or Bengali characters labelled to be German or Bavarian text...)

Clean is the data that has gone through a rudimentary round of preprocessing such as detecting the language based on the script in which the characters are written and estimating the validity of sentence alignments by comparing their length (reasoning that a sentence more than x times the length of another sentence, can hardly be considered to be -well-aligned-.

Informed will either be all those data that has been labeled and evaluated by human native speaker of the corresponding language (for aligned text: translators).

12.1.3 Feature Validity

Guess is similar to the above -naive- as it is based on automatic functions and a data-driven approach which can be applied without access to native speakers, experts, or linguistic literature to draw from. As the name indicates, these rules are very basic and might be considered close to guessing the correct replacement of a word or sub-word unit. They a high in number (thousands) but also include single character replacements and removals without concern for the context inside the text.

Reason is an improved version of the rules from above in which multiple quality assuring measures are taken. Such as preventing the replacement of single characters with an empty string (without taking the context into account) which results in entire texts missing a set of characters. This is accomplished by including a context-window around the sub-word units during replacement rule creation (a reasonable approach for lex is still needed). Currently this window has a length of 1 in each direction. Therefore, the aligned word-pair *fochgebiet* $\rightarrow$ *fachgebiet* would not result in a rule replacing $o \rightarrow a$, but *foc* $\rightarrow$ *fac*.

Authentic is an approach of using replacement rules that have been derived from descriptions in scientific literature by expert linguists.

12.1.4 Perturbation Type

Lex denotes lexicographic replacements of entire words based on bilingual word lists.

Mor denotes morphological replacements of sub-word units based on rules derived by processing bilingual word lists.

All denotes the combination of both previous replacements by applying morphological ones after the lexicographic ones. During this process, the lexicographic replacement marks each replaced word, such that they are ignored during the morphological replacement to not distort properly replaced words as an aftereffect.

TABLE 12.1: **Baseline** Evaluation Metrics for **Bavarian and German without any operations** or modifications (How different are the Bavarian texts from their aligned German texts?)
→ Higher BLEU for clean data quality makes sense, due to one of the cleaning operations excluding sentence pairs if their lengths differ too much from each other (usually due to things such as sub-sentences missing or words translating into multi-word-expressions).

Data Quality	Feature Validity	Perturbation Type	Experiment	BLEU	chrF2	TER
naive	none	none	PERT	16.745	28.9858	97.9605
clean	none	none	PERT	22.0817	46.0659	68.867

TABLE 12.2: **Dialectalization** Evaluation Metrics for **German-Bavarian and their perturbations** (How much does the modified German text resemble the original aligned Bavarian text?)
→ The naive-guess combination being the worst fits perfectly, due to its chaotic nature (single characters get replaced or removed without concern for the context, which messes up most of the text contents.)
→ The strong similarities between lex and all (lex and then mor) perturbations could be due to a strong overlap of the corresponding replacement rules between lex (entire words) and mor (sub-word-units). The fact that words that have been replaced during the lex-perturbation are not being considered while searching for matching strings during the mor-perturbation drastically reduces the applicable replacements.
→ The fact that naive (very noisy) data ends up more similar to the targetted language variety by applying perturbations is puzzling. (Idea for a sanity-check: Look at a set of sentences and manually compare their perturbed states and see if they make sense?)

Data Quality	Feature Validity	Perturbation Type	Experiment	BLEU	chrF2	TER
naive	guess	lex	PERT	65.1959	81.9966	16.2399
naive	guess	mor	PERT	0.0109	8.2787	100.0043
naive	reason	lex	PERT	65.1963	82.0015	16.2399
naive	reason	mor	PERT	24.928	62.5813	48.9352
clean	guess	lex	PERT	61.0166	81.4545	18.0258
clean	guess	mor	PERT	0.0141	8.1328	100.0018
clean	guess	all	PERT	61.0162	81.4362	18.0258
clean	reason	lex	PERT	61.0164	81.4217	18.0258
clean	reason	mor	PERT	18.2865	61.422	54.113
clean	reason	all	PERT	61.0164	81.4	18.0258

TABLE 12.3: **Standardization** Evaluation Metrics for **Bavarian-German and their perturbations** (How much does the modified Bavarian text resemble the original aligned German text?)
→ Here we see an improvement from perturbations higher feature validity (reason) than on lower one (guess). The creation and application of replacement rules is done separately for each language direction, explaining the observed differences.
→ The perturbation direction of standardization (Bavarian into German) performing so much worse fits with the inherent noise of the data (no matter if naive or clean data quality) since it contains text data from all kinds of sub-dialects. These make it less likely to detect the words (and sub-word units) to be affected by the perturbation process, leaving a lot of the data untouched.
→ The other direction (German to Bavarian) starts on standardized text with regulated rules and conformity. Even if the replaced part does not always perfectly match the original (Bavarian) content, there should be many more found text passages for replacement, bringing the perturbed data closer to the desired outcome.
→ Idea: Calculate the same metrics between the perturbed and the original files to get a measure for how much the texts changed according to each experiment setup.

Data Quality	Feature Validity	Perturbation Type	Experiment	BLEU	chrF2	TER
naive	guess	lex	PERT	11.7641	27.2298	101.3146
naive	guess	mor	PERT	0.0063	11.9015	123.348
naive	guess	all	PERT	11.7586	27.224	101.3231
naive	reason	lex	PERT	11.7597	27.2123	101.325
naive	reason	mor	PERT	3.4873	24.3305	109.4471
naive	reason	all	PERT	11.763	27.2314	101.3171
clean	guess	lex	PERT	14.228	42.0906	75.2158
clean	guess	mor	PERT	0.0089	14.0903	106.5991
clean	guess	all	PERT	14.2266	42.0916	75.2094
clean	reason	lex	PERT	14.2286	42.1197	75.1975
clean	reason	mor	PERT	3.7324	38.1686	88.1964
clean	reason	all	PERT	14.2245	42.101	75.1912

TABLE 12.4: **Baseline** Evaluation Metrics for **Bavarian and their English translations** (How different are the English translations of Bavarian text compared to the English translations of their aligned German text?)
→ Better scores for clean than for naive make sense again for the same reason: The clean data can be expected to be more similar across both language varieties than the naive (very noisy) data.

Data Quality	Feature Validity	Perturbation Type	Experiment	BLEU	chrF2	TER
naive	none	none	NLLB	4.4749	14.3248	119.8208
clean	none	none	NLLB	26.0255	44.3625	69.7986

TABLE 12.5: **Dialectalization** Evaluation Metrics for **German-Bavarian and the translation of their perturbations** (How different are the English translations of perturbed German text (looking more like Bavarian text), compared to the English translations of their aligned Bavarian text?)
→ Clean data quality is far superior to the naive (very noisy) data quality for all (currently finished) types of perturbations.
→ Applying perturbations with a higher feature validity (reason) is comparable and slightly worse than the lower feature validity (guess), which at first glance seems unintuitive and will require a deeper investigation. Current assumption: This likely stems from the fact that the perturbations that happen on reason level might have a better performing effect on the score, but are far lower in number and are thus limited at the moment.
Number of replacement rules for comparison:
clean-guess-mor-prefix: 3728 = clean-reason-mor-prefix: 15
clean-guess-mor-suffix: 1368 = clean-reason-mor-suffix: 84
clean-guess-mor-infix: 3619 = clean-reason-mor-infix: 45
clean-guess-lex: 7748 = clean-reason-lex: 7748
→ Idea: Modify scripts to keep track of how many replacements are applied during each step to better compare the impact that each replacement has on the final score?

Data Quality	Feature Validity	Perturbation Type	Experiment	BLEU	chrF2	TER
naive	guess	lex	NLLB	38.2932	51.2928	65.3969
naive	guess	mor	NLLB	1.0201	9.3358	152.2364
naive	guess	all	NLLB	0.7376	8.2818	161.4027
clean	guess	lex	NLLB	72.0584	82.3938	20.7683
clean	guess	mor	NLLB	Running	Running	Running
clean	reason	lex	NLLB	71.7758	82.1987	21.088
clean	reason	mor	NLLB	38.8726	62.2132	49.385

Table 12.6: **Standardization** Evaluation Metrics for **Bavarian-German and the translations of their perturbations** (How different are the English translations of perturbed Bavarian text (looking more like German text), compared to the English translations of their aligned German text?)
→ Similar to the observations from the above perturbation tables, the clean data for the (translated) standardization direction of perturbation outperforms the guess (very noisy) data in the same setup.
→ It is also still the case, that dialectalization has an easier time to perturb the text data as it is desired compared to the standardization.

Data Quality	Feature Validity	Perturbation Type	Experiment	BLEU	chrF2	TER
naive	guess	lex	NLLB	4.322	19.6195	127.2325
naive	guess	mor	NLLB	0.4404	10.9578	181.0172
naive	guess	all	NLLB	0.424	12.2833	223.8087
clean	guess	lex	NLLB	19.1092	41.26	77.5735
clean	guess	mor	NLLB	0.3164	17.0362	226.5734
clean	reason	lex	NLLB	19.0075	41.2929	78.0534
clean	reason	mor	NLLB	14.0651	38.3677	83.5022

Table 12.7: **ALL IN ONE - PERT**

Data Quality	Feature Validity	Perturbation Type	Experiment	BLEU	chrF2	TER
naive	none	none	PERT	16.745	28.9858	97.9605
clean	none	none	PERT	22.0817	46.0659	68.867
naive	guess	lex	PERT	65.1959	81.9966	16.2399
naive	guess	mor	PERT	0.0109	8.2787	100.0043
naive	reason	lex	PERT	65.1963	82.0015	16.2399
naive	reason	mor	PERT	24.928	62.5813	48.9352
clean	guess	lex	PERT	61.0166	81.4545	18.0258
clean	guess	mor	PERT	0.0141	8.1328	100.0018
clean	guess	all	PERT	61.0162	81.4362	18.0258
clean	reason	lex	PERT	61.0164	81.4217	18.0258
clean	reason	mor	PERT	18.2865	61.422	54.113
clean	reason	all	PERT	61.0164	81.4	18.0258
naive	guess	lex	PERT	11.7641	27.2298	101.3146
naive	guess	mor	PERT	0.0063	11.9015	123.348
naive	guess	all	PERT	11.7586	27.224	101.3231
naive	reason	lex	PERT	11.7597	27.2123	101.325
naive	reason	mor	PERT	3.4873	24.3305	109.4471
naive	reason	all	PERT	11.763	27.2314	101.3171
clean	guess	lex	PERT	14.228	42.0906	75.2158
clean	guess	mor	PERT	0.0089	14.0903	106.5991
clean	guess	all	PERT	14.2266	42.0916	75.2094
clean	reason	lex	PERT	14.2286	42.1197	75.1975
clean	reason	mor	PERT	3.7324	38.1686	88.1964
clean	reason	all	PERT	14.2245	42.101	75.1912

Table 12.8: **ALL IN ONE - NLLB**

Data Quality	Feature Validity	Perturbation Type	Experiment	BLEU	chrF2	TER
naive	none	none	NLLB	4.4749	14.3248	119.8208
clean	none	none	NLLB	26.0255	44.3625	69.7986
naive	guess	lex	NLLB	4.322	19.6195	127.2325
naive	guess	mor	NLLB	0.4404	10.9578	181.0172
naive	guess	all	NLLB	0.424	12.2833	223.8087
clean	guess	lex	NLLB	19.1092	41.26	77.5735
clean	guess	mor	NLLB	0.3164	17.0362	226.5734
clean	reason	lex	NLLB	19.0075	41.2929	78.0534
clean	reason	mor	NLLB	14.0651	38.3677	83.5022
naive	none	none	PERT	16.745	28.9858	97.9605
clean	none	none	PERT	22.0817	46.0659	68.867
naive	guess	lex	PERT	65.1959	81.9966	16.2399
naive	guess	mor	PERT	0.0109	8.2787	100.0043
naive	reason	lex	PERT	65.1963	82.0015	16.2399
naive	reason	mor	PERT	24.928	62.5813	48.9352
clean	guess	lex	PERT	61.0166	81.4545	18.0258
clean	guess	mor	PERT	0.0141	8.1328	100.0018
clean	guess	all	PERT	61.0162	81.4362	18.0258
clean	reason	lex	PERT	61.0164	81.4217	18.0258
clean	reason	mor	PERT	18.2865	61.422	54.113
clean	reason	all	PERT	61.0164	81.4	18.0258
naive	guess	lex	PERT	11.7641	27.2298	101.3146
naive	guess	mor	PERT	0.0063	11.9015	123.348
naive	guess	all	PERT	11.7586	27.224	101.3231
naive	reason	lex	PERT	11.7597	27.2123	101.325
naive	reason	mor	PERT	3.4873	24.3305	109.4471
naive	reason	all	PERT	11.763	27.2314	101.3171
clean	guess	lex	PERT	14.228	42.0906	75.2158
clean	guess	mor	PERT	0.0089	14.0903	106.5991
clean	guess	all	PERT	14.2266	42.0916	75.2094
clean	reason	lex	PERT	14.2286	42.1197	75.1975
clean	reason	mor	PERT	3.7324	38.1686	88.1964
clean	reason	all	PERT	14.2245	42.101	75.1912

Chapter 13

Evaluation

13.1 Variables and Parameter

13.1.1 Language Variety

The degree of data scarcity vastly differs between language varieties (refer to recent alam paper) which in turn limits the quality and amount of data and linguistic rules that can be derived from it. The choice of language variety can be expected to have the greatest impact on the performance of the entire pipeline / this approach in it's entirety.

13.1.2 Data Quality

Naive denotes data that has been collected via means such as opustools which encompasses a range of different text corpora of varying degrees of quality. This data contains a lot of noise. Sentences that a ill-aligned, sentences that are of low quality, and even text from very different languages. (Show Examples: Corpora in which the aligned sentence is the same in both languages AND sometimes Chinese or Bengali characters labelled to be German or Bavarian text...)

Clean is the data that has gone through a rudimentary round of preprocessing such as detecting the language based on the script in which the characters are written and estimating the validity of sentence alignments by comparing their length (reasoning that a sentence more than x times the length of another sentence, can hardly be considered to be -well-aligned-.

Informed is all those data that has been labelled and evaluated by human native speaker of the corresponding language (for aligned text: translators)

13.1.3 Feature Validity

Guess is similar to the above -naive- as it is based on automatic functions and a data-driven approach which can be applied without access to native speakers, experts, or linguistic literature to draw from. As the name indicates, these rules are very basic and might be considered close to guessing the correct replacement of a word or sub-word unit.

Reason is an improved version of the rules from above in which multiple quality assuring measures are taken. Such as preventing the replacement of single characters with an empty string (without taking the context into account) which results in entire texts missing a set of characters.

Authentic is an approach of using replacement rules that have been derived from descriptions in scientific literature by expert linguists.

13.1.4 Perturbation Type

Lex denotes lexicographic replacements of entire words based on bilingual word lists.

Mor denotes morphological replacements of sub-word units based on rules derived by processing bilingual word lists.

All denotes the combination of both previous replacements by applying morphological ones after the lexicographic ones.

13.2 German and Bavarian

Table 13.1: Evaluation Metrics for Bavarian against the German reference

Data Quality	Feature Validity	Perturbation Type	Experiment	BLEU	chrF2	TER
naive	none	none	PERT	16.745	28.9858	97.9605
naive	none	none	NLLB	4.4749	14.3248	119.8208

Table 13.1 shows the differences between the Standard German and the Bavarian variant in terms of word-level (PERT) and in terms of translation model performance (or robustness to the dialect) by using the results of the German-to-English translations as reference (NLLB).

Table 13.2: Evaluation Metrics for Bavarian-German (standardized as part of preprocessing)

Data Quality	Feature Validity	Perturbation Type	Experiment	BLEU	chrF2	TER
naive	guess	lex	PERT	13.787	38.4879	75.925
naive	guess	mor	PERT	6.7303	19.5736	85.1412
naive	guess	all	PERT	6.6107	17.5	86.3544
naive	guess	lex	NLLB	4.322	19.6195	127.2325
naive	guess	mor	NLLB	0.4404	10.9578	181.0172
naive	guess	all	NLLB	0.424	12.2833	223.8087

Table 13.2 shows how standardized text (Bavarian text that has been perturbed to resemble Standard German text) performs compared to the unaltered Standard German text. Again (PERT) indicates experiments of modifying text data, while (NLLB) indicates the translation to English compared to the Standard German text translated into English.

Table 13.3 shows how translating English text into German and then, in turn, perturbing this text into a more Bavarian text, compares to the original Bavarian sentences that have already been aligned with the English input sentences.

Table 13.4 shows how dialectized text (Standard German text that has been perturbed to resemble Bavarian text) performs compared to the unaltered Standard

Table 13.3: Evaluation Metrics for English-Bavarian (dialectized as part of postprocessing)

Data Quality	Feature Validity	Perturbation Type	Experiment	BLEU	chrF2	TER
naive	guess	all	POST	5.092	15.8652	83.6179
naive	guess	mor	POST	5.1327	17.119	83.5147
naive	guess	lex	POST	29.1958	59.5913	54.0825

Table 13.4: Evaluation Metrics for German-Bavarian (dialectized)

Data Quality	Feature Validity	Perturbation Type	Experiment	BLEU	chrF2	TER
naive	guess	lex	PERT	51.5588	72.5826	24.3256
naive	guess	mor	PERT	10.2236	18.9192	70.7583
naive	guess	all	PERT	10.2056	17.3224	70.7903
naive	guess	lex	NLLB	38.2932	51.2928	65.3969
naive	guess	mor	NLLB	1.0201	9.3358	152.2364
naive	guess	all	NLLB	0.7376	8.2818	161.4027

German text. Again (PERT) indicates experiments of modifying text data, while (NLLB) indicates the translation to English compared to the Standard German text translated into English.

13.3 New Update

Table 13.5: Evaluation Metrics for Bavarian and German without any operations or modifications (How different are the Bavarian texts from their aligned German texts?)

Data Quality	Feature Validity	Perturbation Type	Experiment	BLEU	chrF2	TER
naive	none	none	PERT	16.745	28.9858	97.9605
clean	none	none	PERT	22.0817	46.0659	68.867

Table 13.6: Evaluation Metrics for Bavarian and their English translations (How different are the English translations of Bavarian text compared to the English translations of their aligned German text?)

Data Quality	Feature Validity	Perturbation Type	Experiment	BLEU	chrF2	TER
naive	none	none	NLLB	4.4749	14.3248	119.8208

Table 13.7: Evaluation Metrics for German-Bavarian and their perturbations (How much does the modified German text resemble the original aligned Bavarian text?)

Data Quality	Feature Validity	Perturbation Type	Experiment	BLEU	chrF2	TER
naive	guess	lex	PERT	65.1959	81.9966	16.2399
naive	guess	mor	PERT	0.0109	8.2787	100.0043
naive	reason	lex	PERT	65.1963	82.0015	16.2399
naive	reason	mor	PERT	24.928	62.5813	48.9352
clean	guess	lex	PERT	61.0166	81.4545	18.0258
clean	guess	mor	PERT	0.0141	8.1328	100.0018
clean	guess	all	PERT	61.0162	81.4362	18.0258
clean	reason	lex	PERT	61.0164	81.4217	18.0258
clean	reason	mor	PERT	18.2865	61.422	54.113
clean	reason	all	PERT	61.0164	81.4	18.0258

Table 13.8: Evaluation Metrics for Bavarian-German and their perturbations (How much did Bavarian text change?)

Data Quality	Feature Validity	Perturbation Type	Experiment	BLEU	chrF2	TER
naive	guess	lex	PERT	11.7641	27.2298	101.3146
naive	guess	mor	PERT	0.0063	11.9015	123.348
naive	guess	all	PERT	11.7586	27.224	101.3231
naive	reason	lex	PERT	11.7597	27.2123	101.325
naive	reason	mor	PERT	3.4873	24.3305	109.4471
naive	reason	all	PERT	11.763	27.2314	101.3171
clean	guess	lex	PERT	14.228	42.0906	75.2158
clean	guess	mor	PERT	0.0089	14.0903	106.5991
clean	guess	all	PERT	14.2266	42.0916	75.2094
clean	reason	lex	PERT	14.2286	42.1197	75.1975
clean	reason	mor	PERT	3.7324	38.1686	88.1964
clean	reason	all	PERT	14.2245	42.101	75.1912

Table 13.9: Evaluation Metrics for Bavarian-German and their perturbations (How different are the English translations of perturbed Bavarian text (looking more like German text), compared to the English translations of their aligned German text?)

Data Quality	Feature Validity	Perturbation Type	Experiment	BLEU	chrF2	TER
naive	guess	lex	NLLB	4.322	19.6195	127.2325
naive	guess	mor	NLLB	0.4404	10.9578	181.0172
naive	guess	all	NLLB	0.424	12.2833	223.8087

Chapter 14

Discussion

Chapter 15

Limitations

15.1 Expected Challenges

The following is a list of challenges, what they entail and how I plan to combat them:

1: Sparse Data

- Sparse data is one of the main problems that I deal with in my work. Depending on the applied methods, a certain level of data density will have to be reached to create useful synthetic data.

→ This challenge is intertwined with the scope of this work. If I find enough (and good enough) data for a number of Low-Resource languages and their dialects, then it will be less of a problem. If push comes to shove, I will have to focus on fewer, maybe even a single dialect in this particular work.

2: Lack of Benchmarks

- Evaluation might be exceptionally difficult due to missing benchmarks.

→ I plan to apply established evaluation metrics and techniques but based on the varying availability of each targeted language, the evaluations will have to be considered in detail and geared toward the circumstances of the different languages.

2: Writing Systems

- Language processing in (many) different scripts.

→ I will have to depend on available text processing tools, which I will have to integrate into my workflow and my eventual experiment pipeline. In some cases, I might have to add special rules for certain characters or encodings.

3: Neural Machine Translation

- Applying state-of-the-art methods of Neural Machine Translation.

→ To be able to apply myself to state-of-the-art applications in NLP is not only crucial for this work, but also of importance for my future in this field of research. This is why I am already working my way through a lot of literature in preparation.

4: Foreign Languages

- Language processing of languages that I do not speak myself.

→ This is just another reason to meticulously document all data, decisions and processes to guarantee the validity of my work. Contact with and help from native speakers will enable me to validate at least samples of my data.

5: Time is Running

- The scope encompassed identifying linguistic rules, acquiring sufficient enough and useful data, in the context of Low-Resource dialectal diversity, exploring synthetic data creation, building Neural Machine Translations and evaluating the experiment's outcomes despite established benchmarks. The resources scant; a dude and his supervisors. The time frame is just short of six months.

→ Having learned valuable experience while writing my bachelor thesis and to prevent running out of time, this work will follow a time plan that has been laid out in detail and put together in cooperation with my supervisors.

Chapter 16

Conclusion

Chapter 17

Future Work

References

Ahmadi, Sina, Milind Agarwal, and Antonios Anastasopoulos (May 2023). "PALI: A Language Identification Benchmark for Perso-Arabic Scripts." In: *Tenth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2023)*. Dubrovnik, Croatia: Association for Computational Linguistics, pp. 78–90. (Visited on 06/27/2023).

Ahmadi, Sina and Antonios Anastasopoulos (May 2023). *Script Normalization for Unconventional Writing of Under-Resourced Languages in Bilingual Communities*. DOI: 10.48550/arXiv.2305.16407. arXiv: 2305.16407 [cs]. (Visited on 06/14/2023).

Alam, Md Mahfuz Ibn, Sina Ahmadi, and Antonios Anastasopoulos (May 2023). *CODET: A Benchmark for Contrastive Dialectal Evaluation of Machine Translation*. DOI: 10.48550/arXiv.2305.17267. arXiv: 2305.17267 [cs]. (Visited on 06/14/2023).

Ansell, Alan et al. (Feb. 2023a). *Composable Sparse Fine-Tuning for Cross-Lingual Transfer*. DOI: 10.48550/arXiv.2110.07560. arXiv: 2110.07560 [cs]. (Visited on 02/19/2023).

— (June 2023b). *Distilling Efficient Language-Specific Models for Cross-Lingual Transfer*. arXiv: 2306.01709 [cs]. (Visited on 06/29/2023).

Artemova, Ekaterina and Barbara Plank (Apr. 19, 2023). *Low-Resource Bilingual Dialect Lexicon Induction with Large Language Models*. arXiv: 2304.09957 [cs]. URL: http://arxiv.org/abs/2304.09957 (visited on 11/15/2023). preprint.

Artetxe, Mikel et al. (2018). "UNSUPERVISED NEURAL MACHINE TRANSLATION." In.

Badawi, Soran (Jan. 2023). "A Transformer-Based Neural Network Machine Translation Model for the Kurdish Sorani Dialect." In: *UHD Journal of Science and Technology* 7.1, pp. 15–21. ISSN: 2521-4217. DOI: 10.21928/uhdjst.v7n1y2023.pp15-21. (Visited on 05/04/2023).

Bafna, Niyati et al. (May 23, 2023). *A Simple Method for Unsupervised Bilingual Lexicon Induction for Data-Imbalanced, Closely Related Language Pairs*. arXiv: 2305.14012 [cs]. URL: http://arxiv.org/abs/2305.14012 (visited on 12/11/2023). preprint.

Bali, Kalika et al. (Dec. 2019). "ELLORA: Enabling Low Resource Languages with Technology." In: *UNESCO International Conference on Language Technologies for All (LT4All)*. (Visited on 07/25/2023).

Bandyopadhyay, Saptarashmi (May 2023). "Factored Neural Machine Translation on Low Resource Languages in the COVID-19 Crisis." In: (visited on 06/20/2023).

Bapna, Ankur and Orhan Firat (Nov. 2019). "Simple, Scalable Adaptation for Neural Machine Translation." In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China: Association for Computational Linguistics, pp. 1538–1548. DOI: 10.18653/v1/D19-1165. (Visited on 06/27/2023).

Baroni, Marco (Dec. 2019). "Linguistic Generalization and Compositionality in Modern Artificial Neural Networks." In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 375.1791, p. 20190307. DOI: 10.1098/rstb.2019.0307. (Visited on 07/04/2023).

Batsuren, Khuyagbaatar, Gábor Bella, and Fausto Giunchiglia (Mar. 2022). "A Large and Evolving Cognate Database." In: *Language Resources and Evaluation* 56.1, pp. 165–189. ISSN: 1574-0218. DOI: 10.1007/s10579-021-09544-6. (Visited on 02/14/2023).

Bird, Steven (Dec. 2020). "Decolonising Speech and Language Technology." In: *Proceedings of the 28th International Conference on Computational Linguistics*. Barcelona, Spain (Online): International Committee on Computational Linguistics, pp. 3504–3519. DOI: 10.18653/v1/2020.coling-main.313. (Visited on 08/12/2023).

Bizzoni, Yuri et al. (2020). "How Human Is Machine Translationese? Comparing Human and Machine Translations of Text and Speech." In: *Proceedings of the 17th International Conference on Spoken Language Translation*. Online: Association for Computational Linguistics, pp. 280–290. DOI: 10.18653/v1/2020.iwslt-1.34. (Visited on 04/25/2023).

Bogoychev, Nikolay and Rico Sennrich (Oct. 2020). *Domain, Translationese and Noise in Synthetic Data for Neural Machine Translation*. DOI: 10.48550/arXiv.1911.03362. arXiv: 1911.03362 [cs, stat]. (Visited on 06/27/2023).

Bugliarello, Emanuele et al. (2022a). "IGLUE: A Benchmark for Transfer Learning across Modalities, Tasks, and Languages." In.

Bugliarello, Emanuele et al. (July 2022b). *IGLUE: A Benchmark for Transfer Learning across Modalities, Tasks, and Languages*. arXiv: 2201.11732 [cs]. (Visited on 04/25/2023).

Casas, Noe et al. (July 2021). "Linguistic Knowledge-Based Vocabularies for Neural Machine Translation." In: *Natural Language Engineering* 27.4, pp. 485–506. ISSN: 1351-3249, 1469-8110. DOI: 10.1017/S1351324920000364. (Visited on 07/09/2023).

Cooper Stickland, Asa, Xian Li, and Marjan Ghazvininejad (Apr. 2021). "Recipes for Adapting Pre-trained Monolingual and Multilingual Models to Machine Translation." In: *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. Online: Association for Computational Linguistics, pp. 3440–3453. DOI: 10.18653/v1/2021.eacl-main.301. (Visited on 06/27/2023).

Crystal, David (2000). *Language Death*. Cambridge: Cambridge University Press. DOI: 10.1017/CBO9781139106856. (Visited on 08/12/2023).

Czarnowska, Paula et al. (Oct. 22, 2019). *Don't Forget the Long Tail! A Comprehensive Analysis of Morphological Generalization in Bilingual Lexicon Induction*. DOI: 10.48550/arXiv.1909.02855. arXiv: 1909.02855 [cs]. preprint.

de Vries, Wietse et al. (2021). "Adapting Monolingual Models: Data Can Be Scarce When Language Similarity Is High." In: *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pp. 4901–4907. DOI: 10.18653/v1/2021.findings-acl.433. arXiv: 2105.02855 [cs]. (Visited on 07/15/2023).

Dekker, Kelly and Rob van der Goot (May 2020). "Synthetic Data for English Lexical Normalization: How Close Can We Get to Manually Annotated Data?" In: *Proceedings of the Twelfth Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, pp. 6300–6309. ISBN: 979-10-95546-34-4. (Visited on 06/27/2023).

Demszky, Dorottya et al. (2021). "Learning to Recognize Dialect Features." In: *Proceedings of the 2021 Conference of the North American Chapter of the Association*

for Computational Linguistics: Human Language Technologies. Online: Association for Computational Linguistics, pp. 2315–2338. DOI: 10.18653/v1/2021.naacl-main.184. (Visited on 07/07/2023).

Dou, Zi-Yi, Antonios Anastasopoulos, and Graham Neubig (Nov. 2020). "Dynamic Data Selection and Weighting for Iterative Back-Translation." In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, pp. 5894–5904. DOI: 10.18653/v1/2020.emnlp-main.475. (Visited on 06/26/2023).

Doval, Yerai, Jesús Vilares, and Carlos Gómez-Rodríguez (Sept. 2020). *Towards Robust Word Embeddings for Noisy Texts*. DOI: 10.48550/arXiv.1911.10876. arXiv: 1911.10876 [cs]. (Visited on 06/27/2023).

Duan, Sufeng et al. (Apr. 2020). *Syntax-Aware Data Augmentation for Neural Machine Translation*. DOI: 10.48550/arXiv.2004.14200. arXiv: 2004.14200 [cs]. (Visited on 06/27/2023).

Dupont, Quinn (Jan. 2017). "The Cryptological Origins of Machine Translation, from al-Kindi to Weaver." In: *amodern*.

Edunov, Sergey et al. (Oct. 2018). "Understanding Back-Translation at Scale." In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics, pp. 489–500. DOI: 10.18653/v1/D18-1045. (Visited on 06/27/2023).

Foster, Jennifer and Øistein E. Andersen (2009). "GenERRate: Generating Errors for Use in Grammatical Error Detection." In: *Proceedings of the Fourth Workshop on Innovative Use of NLP for Building Educational Applications - EdAppsNLP '09*. Boulder, Colorado: Association for Computational Linguistics, pp. 82–90. ISBN: 978-1-932432-37-4. DOI: 10.3115/1609843.1609855. (Visited on 06/27/2023).

Gao, Fei et al. (July 2019). "Soft Contextual Data Augmentation for Neural Machine Translation." In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, pp. 5539–5544. DOI: 10.18653/v1/P19-1555. (Visited on 06/27/2023).

Garcia, Xavier et al. (June 2021). "Harnessing Multilinguality in Unsupervised Machine Translation for Rare Languages." In: *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Online: Association for Computational Linguistics, pp. 1126–1137. DOI: 10.18653/v1/2021.naacl-main.89. (Visited on 06/27/2023).

Gouws, Stephan, Yoshua Bengio, and Greg Corrado (Feb. 4, 2016). *BilBOWA: Fast Bilingual Distributed Representations without Word Alignments*. DOI: 10.48550/arXiv.1410.2455. arXiv: 1410.2455 [cs, stat]. preprint.

Green, Lisa J. (2002). *African American English: A Linguistic Introduction*. Cambridge University Press. DOI: 10.1017/CBO9780511800306.

Ha, Thanh-Le, Jan Niehues, and Alexander Waibel (Nov. 2016). *Toward Multilingual Neural Machine Translation with Universal Encoder and Decoder*. DOI: 10.48550/arXiv.1611.04798. arXiv: 1611.04798 [cs]. (Visited on 06/27/2023).

Hasan, Tahmid et al. (2020). "Not Low-Resource Anymore: Aligner Ensembling, Batch Filtering, and New Datasets for Bengali-English Machine Translation." In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). Online: Association for Computational Linguistics, pp. 2612–2623. DOI: 10.18653/v1/2020.emnlp-main.207.

Held, Will, Caleb Ziems, and Diyi Yang (May 2023). *TADA: Task-Agnostic Dialect Adapters for English*. DOI: 10.48550/arXiv.2305.16651. arXiv: 2305.16651 [cs]. (Visited on 07/06/2023).

Hu, Junjie et al. (Sept. 2020). *XTREME: A Massively Multilingual Multi-task Benchmark for Evaluating Cross-lingual Generalization*. DOI: 10.48550/arXiv.2003.11080. arXiv: 2003.11080 [cs]. (Visited on 04/29/2023).

Karakanta, Alina, Jon Dehdari, and Josef Van Genabith (June 2018). "Neural Machine Translation for Low-Resource Languages without Parallel Corpora." In: *Machine Translation* 32.1-2, pp. 167–189. ISSN: 0922-6567, 1573-0573. DOI: 10.1007/s10590-017-9203-5. (Visited on 04/25/2023).

Kiela, Douwe et al. (June 2021). "Dynabench: Rethinking Benchmarking in NLP." In: *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Online: Association for Computational Linguistics, pp. 4110–4124. DOI: 10.18653/v1/2021.naacl-main.324. (Visited on 07/07/2023).

Koehn, Philipp (2009). *Statistical Machine Translation*. Cambridge: Cambridge University Press. ISBN: 978-0-521-87415-1. DOI: 10.1017/CBO9780511815829. (Visited on 07/25/2023).

— (2020). *Neural Machine Translation*. Cambridge: Cambridge University Press. ISBN: 978-1-108-49732-9. DOI: 10.1017/9781108608480. (Visited on 07/07/2023).

Kornai, András (Oct. 2013). "Digital Language Death." In: *PLOS ONE* 8.10, e77056. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0077056. (Visited on 06/16/2023).

Lameli, Alfred (Aug. 2008). *Deutsche Sprachlandschaften*. (Visited on 07/05/2023).

Lauscher, Anne et al. (May 2020). *From Zero to Hero: On the Limitations of Zero-Shot Cross-Lingual Transfer with Multilingual Transformers*. arXiv: 2005.00633 [cs]. (Visited on 04/29/2023).

Lavie, Alon and Abhaya Agarwal (June 2007). "METEOR: An Automatic Metric for MT Evaluation with High Levels of Correlation with Human Judgments." In: *Proceedings of the Second Workshop on Statistical Machine Translation*. Prague, Czech Republic: Association for Computational Linguistics, pp. 228–231. (Visited on 07/08/2023).

Li, Yaoyiran, Anna Korhonen, and Ivan Vulić (Oct. 21, 2023). *On Bilingual Lexicon Induction with Large Language Models*. arXiv: 2310.13995 [cs]. URL: http://arxiv.org/abs/2310.13995 (visited on 11/15/2023). preprint.

Liu, Yanchen, William Held, and Diyi Yang (May 2023). *DADA: Dialect Adaptation via Dynamic Aggregation of Linguistic Rules*. DOI: 10.48550/arXiv.2305.13406. arXiv: 2305.13406 [cs]. (Visited on 07/04/2023).

Lusito, Stefano, Edoardo Ferrante, and Jean Maillard (June 2022). *Text Normalization for Endangered Languages: The Case of Ligurian*. DOI: 10.48550/arXiv.2206.07861. arXiv: 2206.07861 [cs]. (Visited on 06/27/2023).

Malykh, Valentin, Varvara Logacheva, and Taras Khakhulin (Nov. 2018). "Robust Word Vectors: Context-Informed Embeddings for Noisy Texts." In: *Proceedings of the 2018 EMNLP Workshop W-NUT: The 4th Workshop on Noisy User-generated Text*. Brussels, Belgium: Association for Computational Linguistics, pp. 54–63. DOI: 10.18653/v1/W18-6108. (Visited on 06/27/2023).

Martin, Stefan and Walt Wolfram (1998). "The sentence in African-American Vernacular English." In: *African-American English: Structure, history and use*. Ed. by Salikoko S. Mufwene et al. London: Routledge, pp. 11–36. ISBN: 978-0-415-11732-6 978-0-415-11733-3.

Matras, Yaron (June 2017). "Preliminary Findings from the Manchester Database." In.

Matras, Yaron (Jan. 2019). "Revisiting Kurdish Dialect Geography: Findings from the Manchester Database Introduction: Database Method and Scope." In: *Revisiting Kurdish dialect geography: Findings from the Manchester Database. In: Haig, Geoffrey, Öpengin, Ergin & Gundoğlu, Songül, eds. Current Issues in Kurdish Linguistics. Bamberg: Bamberg University Press. 225-241.* (Visited on 06/07/2023).

Moseley, Christopher and Alexandre Nicolas (2010). *Atlas of the World's Languages in Danger*. 3rd ed., entirely rev., enl., upd. Paris : UNESCO, 2010. ISBN: 978-92-3-104096-2 (corr.) (Visited on 07/25/2023).

Ngo, Thi-Vinh et al. (Dec. 2022). "An Efficient Method for Generating Synthetic Data for Low-Resource Machine Translation." In: *Applied Artificial Intelligence* 36.1, p. 2101755. ISSN: 0883-9514. DOI: 10.1080/08839514.2022.2101755. (Visited on 06/24/2023).

Nigmatulina, Iuliia, Tannon Kew, and Tanja Samardzic (Dec. 2020). "ASR for Non-standardised Languages with Dialectal Variation: The Case of Swiss German." In: *Proceedings of the 7th Workshop on NLP for Similar Languages, Varieties and Dialects*. VarDial 2020. Ed. by Marcos Zampieri et al. Barcelona, Spain (Online): International Committee on Computational Linguistics (ICCL), pp. 15–24. URL: https://aclanthology.org/2020.vardial-1.2 (visited on 11/25/2023).

Papineni, Kishore et al. (July 2002). "BLEU: A Method for Automatic Evaluation of Machine Translation." In: *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*. ACL '02. USA: Association for Computational Linguistics, pp. 311–318. DOI: 10.3115/1073083.1073135. (Visited on 04/29/2023).

Parović, Marinela et al. (July 2022). "BAD-X: Bilingual Adapters Improve Zero-Shot Cross-Lingual Transfer." In: *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Seattle, United States: Association for Computational Linguistics, pp. 1791–1799. DOI: 10.18653/v1/2022.naacl-main.130. (Visited on 04/25/2023).

Pfeiffer, Jonas et al. (Nov. 2020). "MAD-X: An Adapter-Based Framework for Multi-Task Cross-Lingual Transfer." In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, pp. 7654–7673. DOI: 10.18653/v1/2020.emnlp-main.617. (Visited on 07/04/2023).

Popović, Maja (Sept. 2015). "chrF: Character n-Gram F-score for Automatic MT Evaluation." In: *Proceedings of the Tenth Workshop on Statistical Machine Translation*. Lisbon, Portugal: Association for Computational Linguistics, pp. 392–395. DOI: 10.18653/v1/W15-3049. (Visited on 06/27/2023).

Post, Matt (Oct. 2018). "A Call for Clarity in Reporting BLEU Scores." In: *Proceedings of the Third Conference on Machine Translation: Research Papers*. Brussels, Belgium: Association for Computational Linguistics, pp. 186–191. DOI: 10.18653/v1/W18-6319. (Visited on 06/27/2023).

Rei, Ricardo et al. (Nov. 2020). "COMET: A Neural Framework for MT Evaluation." In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, pp. 2685–2702. DOI: 10.18653/v1/2020.emnlp-main.213. (Visited on 06/27/2023).

Reimers, Nils and Iryna Gurevych (Nov. 2020). "Making Monolingual Sentence Embeddings Multilingual Using Knowledge Distillation." In: *Proceedings of the 2020*

Conference on Empirical Methods in Natural Language Processing (EMNLP). Online: Association for Computational Linguistics, pp. 4512–4525. DOI: 10.18653/v1/2020.emnlp-main.365. (Visited on 07/10/2023).

Riley, Parker et al. (June 2023). "FRMT: A Benchmark for Few-Shot Region-Aware Machine Translation." In: *Transactions of the Association for Computational Linguistics* 11, pp. 671–685. ISSN: 2307-387X. DOI: 10.1162/tacl_a_00568. (Visited on 07/11/2023).

Ruder, Sebastian et al. (Nov. 2021). "XTREME-R: Towards More Challenging and Nuanced Multilingual Evaluation." In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, pp. 10215–10245. DOI: 10.18653/v1/2021.emnlp-main.802. (Visited on 07/10/2023).

Salam Khalid, Hewa (Feb. 2015). "KURDISH DIALECT CONTINUUM, AS A STANDARDIZATION SOLUTION." In: *IJOKS - International Journal of Kurdish Studies* 1, pp. 27–39. DOI: 10.21600/ijks.95271.

— (Apr. 2020). "KURDISH LANGUAGE, ITS FAMILY AND DIALECTS *." In.

Sánchez-Cartagena, Víctor M. et al. (Nov. 2021). "Rethinking Data Augmentation for Low-Resource Neural Machine Translation: A Multi-Task Learning Approach." In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, pp. 8502–8516. DOI: 10.18653/v1/2021.emnlp-main.669. (Visited on 06/27/2023).

Sellam, Thibault, Dipanjan Das, and Ankur Parikh (July 2020). "BLEURT: Learning Robust Metrics for Text Generation." In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, pp. 7881–7892. DOI: 10.18653/v1/2020.acl-main.704. (Visited on 07/22/2023).

Sennrich, Rico, Barry Haddow, and Alexandra Birch (Aug. 2016). "Improving Neural Machine Translation Models with Monolingual Data." In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Berlin, Germany: Association for Computational Linguistics, pp. 86–96. DOI: 10.18653/v1/P16-1009. (Visited on 06/27/2023).

Snover, Matthew et al. (Aug. 2006). "A Study of Translation Edit Rate with Targeted Human Annotation." In: *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*. Cambridge, Massachusetts, USA: Association for Machine Translation in the Americas, pp. 223–231. (Visited on 07/08/2023).

Tavadze, Givi (2019). "Spreading of the Kurdish Language Dialects and Writing Systems Used in the Middle East." In: 13.1.

Team, NLLB et al. (Aug. 2022). *No Language Left Behind: Scaling Human-Centered Machine Translation*. DOI: 10.48550/arXiv.2207.04672. arXiv: 2207.04672 [cs]. (Visited on 06/19/2023).

Tripathi, Sneha (2010). "Approaches to Machine Translation." In: (visited on 07/25/2023).

Tsujii, Junichi (Dec. 2021). "Natural Language Processing and Computational Linguistics." In: *Computational Linguistics* 47.4, pp. 707–727. ISSN: 0891-2017. DOI: 10.1162/coli_a_00420. (Visited on 08/12/2023).

Üstün, Ahmet et al. (Oct. 2021). *Multilingual Unsupervised Neural Machine Translation with Denoising Adapters*. arXiv: 2110.10472 [cs]. (Visited on 06/04/2023).

Vaswani, Ashish et al. (Dec. 2017). *Attention Is All You Need*. DOI: 10.48550/arXiv.1706.03762. arXiv: 1706.03762 [cs]. (Visited on 04/29/2023).

Vulić, Ivan and Marie-Francine Moens (June 2013). "Cross-Lingual Semantic Similarity of Words as the Similarity of Their Semantic Word Responses." In: *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. NAACL-HLT 2013. Ed. by Lucy Vanderwende, Hal Daumé III, and Katrin Kirchhoff. Atlanta, Georgia: Association for Computational Linguistics, pp. 106–116. URL: https://aclanthology.org/N13-1011 (visited on 12/13/2023).

— (July 2015). "Bilingual Word Embeddings from Non-Parallel Document-Aligned Data Applied to Bilingual Lexicon Induction." In: *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*. ACL-IJCNLP 2015. Ed. by Chengqing Zong and Michael Strube. Beijing, China: Association for Computational Linguistics, pp. 719–725. DOI: 10.3115/v1/P15-2118.

Waldendorf, Jonas et al. (2022). "Improving Translation of Out Of Vocabulary Words Using Bilingual Lexicon Induction in Low-Resource Machine Translation." In: Conference of the Association for Machine Translation in the Americas. URL: https://www.semanticscholar.org/paper/Improving-Translation-of-Out-Of-Vocabulary-Words-in-Waldendorf-Birch/1694ecf55300c66d6b67c4520f9de5081679b69b (visited on 10/17/2023).

Wang, Alex et al. (2018). "GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding." In: *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*. Brussels, Belgium: Association for Computational Linguistics, pp. 353–355. DOI: 10.18653/v1/W18-5446. (Visited on 07/08/2023).

Wang, Alex et al. (2019). "SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems." In.

Xia, Mengzhou et al. (July 2019). "Generalized Data Augmentation for Low-Resource Translation." In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, pp. 5786–5796. DOI: 10.18653/v1/P19-1579. (Visited on 06/27/2023).

Xie, Ziang et al. (Mar. 2017). *Data Noising as Smoothing in Neural Network Language Models*. DOI: 10.48550/arXiv.1703.02573. arXiv: 1703.02573 [cs]. (Visited on 06/27/2023).

Ye, Jiacheng et al. (Dec. 2022). "ZeroGen: Efficient Zero-shot Learning via Dataset Generation." In: *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, pp. 11653–11669. (Visited on 07/10/2023).

Zhang*, Tianyi et al. (Sept. 2019). "BERTScore: Evaluating Text Generation with BERT." In: *International Conference on Learning Representations*. (Visited on 06/27/2023).

Zhang, Zhirui et al. (Mar. 2018). *Joint Training for Neural Machine Translation Models with Monolingual Data*. DOI: 10.48550/arXiv.1803.00353. arXiv: 1803.00353 [cs]. (Visited on 06/27/2023).

Ziems, Caleb et al. (May 2022). "VALUE: Understanding Dialect Disparity in NLU." In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Dublin, Ireland: Association for Computational Linguistics, pp. 3701–3720. DOI: 10.18653/v1/2022.acl-long.258. (Visited on 07/04/2023).

Ziems, Caleb et al. (May 2023). *Multi-VALUE: A Framework for Cross-Dialectal English NLP*. DOI: 10.48550/arXiv.2212.08011. arXiv: 2212.08011 [cs]. (Visited on 07/04/2023).

Acronyms

NLP natural language processing. 1

Glossary

Alemannic is a major dialect group of German. 35, 80

Bavarian is a major dialect group of German. 35, 79

Caucasus is a region in 14

Central Bavarian is a sub-dialect of Bavarian. 35

Central Kurdish is a major dialect group of Kurdish, mainly spoken in Iraq and Iran.. 79

Erbil names an administrative district and also a city in Iran. 42

Franconian is a major dialect group of German. 35

German is a west germanic language, written in latin script and mainly spoken in Germany. 35, 79, 80

Germany is a country in Europe. 79

Iran is a country 79, 80

Iraq is a country 14, 79

Khorasan is a region in 14

Kobani also written Kobanî, is a sub-dialect of Northern Kurdish. 14, 35

Kurdish is an Indo-Iranian language. 14

Kurmanji also written Kurmanjî, is a name often used for Northern Kurdish. 14

Mahabad also written Mehabad, names an administrative district and also a city in Iraq. 40, 42

Mukriyani (Mahabad, Mukriyani, Mukri) is a sub-dialect of Central Kurdish. 35

Northern Bavarian is a sub-dialect of Bavarian. 35

Northern Kurdish is a major dialect group of Kurdish. 14, 35, 79

Sanandaji (Sanandajî, Sanayi, Sanayî, Senayi, Senayî, Sine, Sine'i, Sine'î, Ardalani, Ardalanî) is a sub-dialect of Central Kurdish. 35

Saxon is a major dialect group of German. 35

Southern Bavarian is a sub-dialect of Bavarian. 35

Sulaymaniyah names an administrative district and also a city in Iran. 42

Swiss German is any of the Alemannic dialects spoken in the German-speaking part of Switzerland. 35

Switzerland is a country in Europe. 80

Syria is a country 14