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, 2024

"*Wege entstehen dadurch, dass man sie geht.*"
(Paths are made by walking them.)

Franz Kafka

Acknowledgements

I am deeply grateful to my thesis supervisors Dr. Seid Muhie Yimam and Dr. Sina Ahmadi. Their meticulous review of drafts and invaluable feedback significantly helped me to improve the style and coherence of this thesis. Their goal-oriented guidance was instrumental in keeping me focused on the core objectives, helping me navigate the ever-expanding labyrinth of intriguing side topics. It was a particular honor to work with Sina on these topics; I'm especially thankful for his willingness to supervise me externally and across significant time zones, first from Virginia and later from Zurich. I would like to extend my appreciation to Prof. Dr. Chris Biemann, my thesis examiner, for his time and expertise in evaluating this work and providing the opportunity for me to write this thesis in the Language Technology Group at University of Hamburg.

I have to thank Raman, who, even if by mere happenstance (during a conversation in our student-dormitory's shared kitchen), aimed me like an arrow towards the topic of my thesis. I would also like to express my heartfelt gratitude to Anran and Domi, whose companionship and support made this journey not just possible, but enjoyable. Their assistance extended beyond emotional support to practical help, exemplified by Domi driving me and all my belongings across the country on the weekend of the thesis submission - a gesture that truly encapsulates the true spirit of love.

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 aims to explore language-agnostic methods for improving neural machine translation of low-resource language variants. Extracting linguistic features enables the creation of synthetic text data oriented by a standard or more dominant language variant. In the main focus are the German dialect groups Bavarian and Alemannic.

Keywords: Natural language processing, dialectal varieties, linguistic information, low-resourced language, neural machine translation

Contents

List of Figures

List of Tables

List of Abbreviations and Acronyms

AI artificial intelligence. 15

BLI bilingual lexicon induction. vi, 21, 22, 26

CA conversational agent. 1, 14

CL computational linguistics. vi, 1, 2, 9, 15, 65

IR information retrieval. 14

LLM large language model. 2, 19, 21, 22

MT machine translation. vi, vii, 1, 3–5, 10–15, 17–22, 25, 30, 34–37, 43, 45, 65, 67

NER named entity recognition. 14

NLP natural language processing. vi, 1–3, 5, 9–11, 14, 15, 21–24, 39, 40, 62

NMT neural machine translation. 3, 4, 15, 19–21, 26, 27, 44

QA questioning & answering. 14

SA sentiment analysis. 1, 14

They're just words we use 'cos we don't know anything better.

— Pip Williams

1 Introduction

This chapter presents the linguistically informed translation in low-resource scenarios as the object of this thesis. The goals in form of research questions are then followed by an outline of the remaining thesis.

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. Additionally, Machine translation (MT) is a crucial component of this field, enabling the translation of text from one language to another using computational methods. 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 (SA), Conversational agents (CAs), and MT, with the latter being the focus of this work.

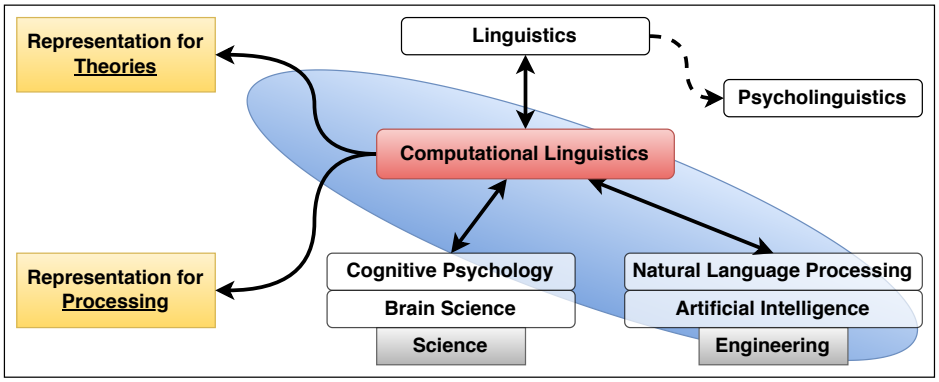


Figure 1.1: A schematic view of the research disciplines surrounding CL. The bottom disciplines are concerned with the processing of language, while the top discipline (linguistics), is concerned with language rules (Tsujii, 2021).

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 have a tradition long before the first computer was even built and insights gained in this field serve as the backbone of language technology (Scherrer, 2012). Despite the rapid progress mainly due to deep learning, such as Large language models (LLMs) like GPT, linguistics still remains an essential component in analyzing the performance of models and studying the languages of the world.

Both, 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 (see Figure 1.1). This schema places the more theoretical computational linguistics as a sub-field of linguistics and the more engineering-oriented NLP 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. It cannot be understated how crucial linguistic foundations are for progress in NLP and what great potential NLP can provide for linguistic investigations.

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 LLM, 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).

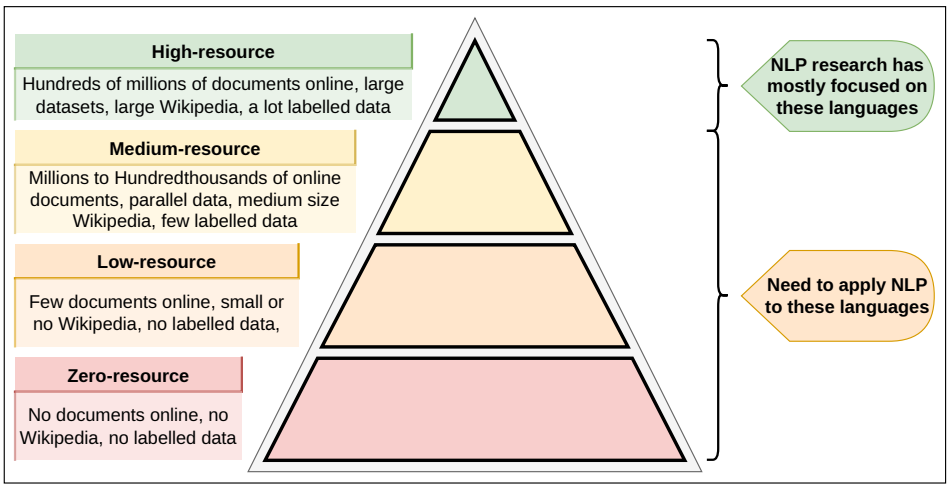


Figure 1.2: A conceptual view of the NLP resource hierarchy, with only very few languages at the top and most of the world's languages at the bottom. Figure adapted from[2].

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

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

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 large amounts of data to be applicable. Starting with rule-based and dictionary-based MTs (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.

By advancing MT 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 MT 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 linguistically motivated synthetic data acquisition methods.

1.4 Objectives and Research Questions

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

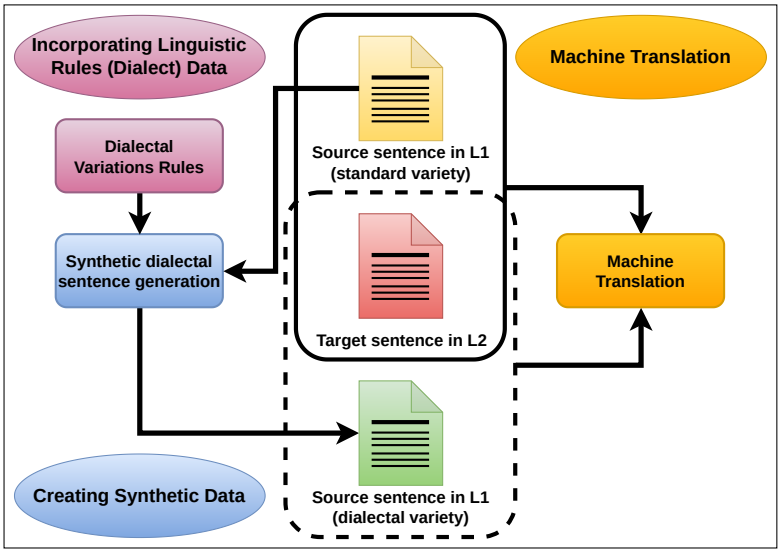


Figure 1.3: Main conceptualization of using standard variant data combined with dialectal variation rules to improve MT.

This work attempts to advance NMT of dialect variations 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-consuming and expensive.

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 MT 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 MT. The main concept of this work is shown in Figure 1.3.

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 languages can be processed more efficiently?

RQ3: What are requirements for deriving tools and processes that can be applied to vastly different languages from various language families?

1.5 Outline

Chapters 2 and 3 lay the foundation by exploring the motivation behind this research, providing essential theoretical background, and reviewing relevant related work. This sets the stage for Chapter 4, which starts by identifying the specific research gap this thesis aims to address, before presenting the methodological approach, detailing the techniques and processes employed in this study. It also provides a comprehensive overview of both the available data and the specific datasets utilized in the research. Chapter 5 follows closely, presenting the results of the experiments and offering an evaluation of the findings. Chapter 6 begins with a critical examination of the study's limitations, framing the subsequent discussion of the research outcomes. This chapter also includes recommendations for future work, highlighting potential avenues for further investigation and concludes this thesis with a synthesis of the key findings and their implications.

I'm sure there are plenty of wonderful words flying around that have never been written on a slip of paper. I want to record them.

— Pip Williams

2 Motivation

This chapter describes the situation and problem context surrounding this thesis. At first, it reveals the intricacies of language variation to then explore the notion of low-resource in NLP and beyond. It concludes with some first impressions of the problems that MT have to contend with by facing dialectal variation.

2.1 Languages and Dialects

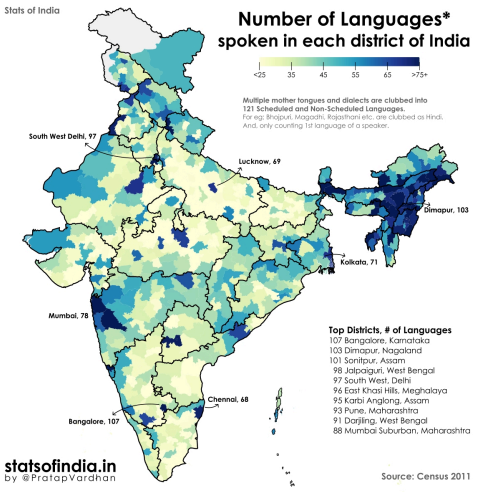
Languages come in different flavors, called dialects, that can be considered to be leaves on a language branch. These dialects are often unknown outside their local sphere and find little recognition in the wider, global, population and research alike. That Hindi is the main language of India and that 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 shown in Figure 2.1.

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. Some maps indicate very complex borders, as in Figure 2.2, while others, are surprisingly simplified and generalize languages to adhere to a certain purpose or agenda by the author of the map. Identifying true native speakers (e.g. for sending out field workers to acquire language data) can be challenging.

There are many languages that come in multiple regional varieties, such as English (British English, American English, Australian English, Indian English, etc.), French (Canadian French, Belgian French, etc.), Spanish (Mexican Spanish, Chilean Spanish, Argentinean Spanish, etc.), which can all display lexical, grammatical, and orthographical distinctions (Aepli et al., 2023).

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

(A) Diversity of local languages in India

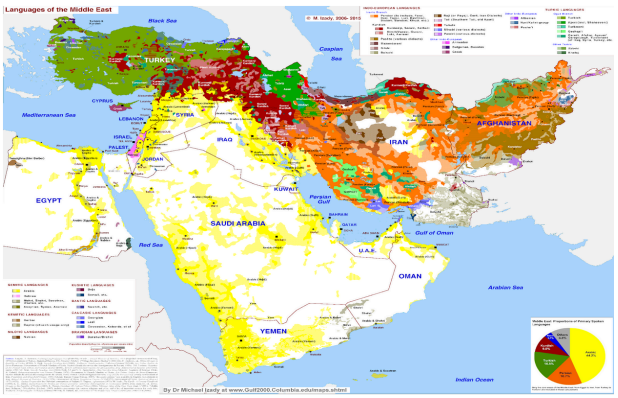


(B) Vast numbers of languages sharing a geolocation

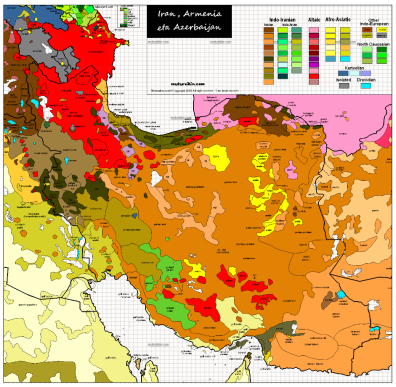
FIGURE 2.1: Linguistic diversity in India. Credit: statsofindia.in based on Indian census 2011

This becomes a serious issue, once these varieties do not display any standardized spelling. Languages for which this is the case can be found in high to medium-resource settings, such as Arabic (Darwish et al., 2021) and Italian (Ramponi, 2024), but also in many low-resource scenarios (Bird, 2022), with affected languages located all around the globe: Africa (Adebara and Abdul-Mageed, 2022), Asia (Roark et al., 2020; Darwish et al., 2021), Oceania (Solano, Nicholas, and Wray, 2018), and America (Littell et al., 2018; Mager et al., 2018).

Ultimately, what is considered a language and what a dialect is often an arbitrary notion that has to be suited to the scope of their investigations (Scherrer, 2012). For example, data samples collected from locations in Bavaria without any knowledge about local varieties may simply be labeled as Bavarian language data by one researcher, while the next team working on the data might filter for linguistic features and come to introduce a finer level of (now dialect) labels. The related controversy of distinguishing



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



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

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

between language and dialect has even been pointed out to be rather futile, considering that there are no simple solutions for doing so linguistically (Derince, Opengin, and Haig, 2008). Scherrer (2012) describes a dialect as a language variety that is defined by the geographical origin of its members. On the other hand, it is said that "a language is a dialect with an army and a navy"[3] and Chambers and Trudgill (1998) who provide the idea of regarding dialects as dialects of a language, elevating the standard variety of a language to be autonomous and above the subordinated dialects (see Figure 2.3) in their mainly oral use. Going forward, the terms language and dialect refer to the concept of language variety as described by (Chambers and Trudgill, 1998) as a way of referring to a way of speaking of a group of humans that we consider as a single entity.

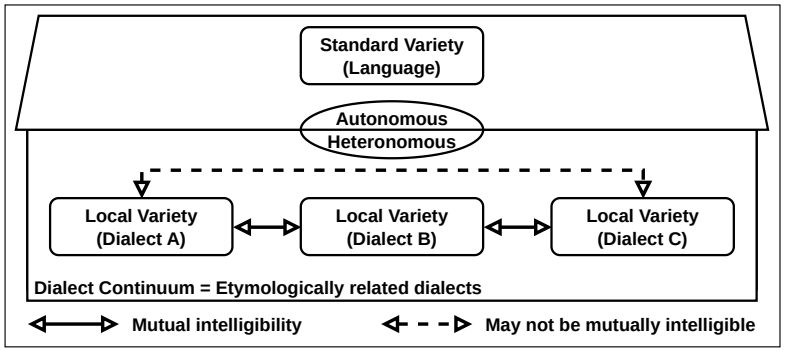
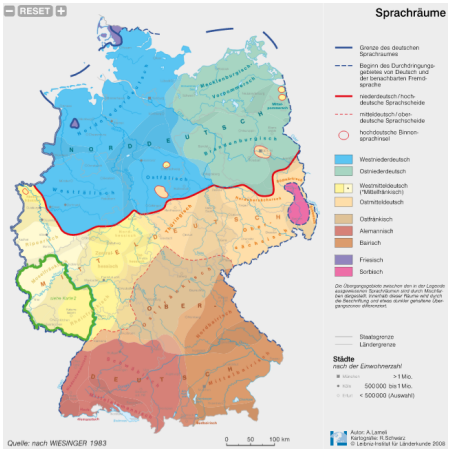


FIGURE 2.3: The house of dialects shows one possible conceptualization of the relation between a language and its dialects. This figure was created based on the ideas explored in (Scherrer, 2012).

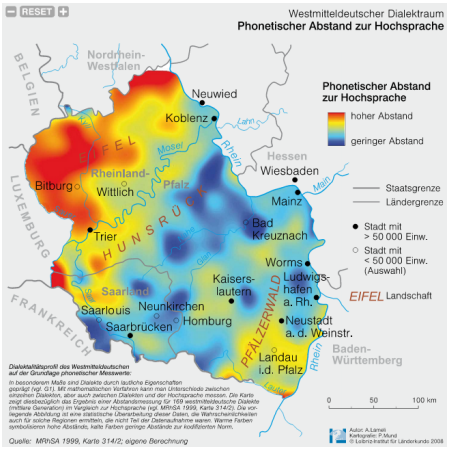
2.1.1 Dialect Continua

In addition to the previously-discussed issues, comes the smoothness of many language and dialect transitions. Chambers and Trudgill (1998) speak of a dialect continuum once a set of dialects are etymologically related. Languages and their speakers do not exist in perfect isolation from each other; they interact and intermingle with each other. This is not limited to creol languages such as Mauritian Creole which combines many aspects from locally dominating languages (in this case French and English) at the time of their development, but also already established languages slowly over time. 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, 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 (Khalid, 2020). This and similar processes have resulted in many dialect continua (Khalid, 2015). Figure 2.4 exemplifies this via a set of German words which gradually change while moving through dialect regions.

[3]This quip, sometimes called Weinreich witticism (refer to https://en.wikipedia.org/wiki/A_language_is_a_dialect_with_an_army_and_navy for its origin), points to the arbitrariness of distinguishing between language and dialect, while at the same time suggesting that political action is necessary for a language to establish itself long-term.



(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 observed in a set of German words. English translations from left to right: I, doing, village, that, pound, apple.

FIGURE 2.4: Distribution and transitions of German dialects. Source: (Lameli, 2008)

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 (Khalid, 2020). This leads to problems for processing data from such language varieties, illustrated on an example sentence in Table 2.1. The three Bavarian variations differ sometimes drastically from each other, but it can be expected that they would simply all be labeled as Bavarian in most current datasets.

Language	**Translated example sentence**
English	*Although I like her I won't marry her.*
Northern Bavarian	*Trotzdean das'e's moch, hairon tou'e's niat.*
Central Bavarian	*Obwoi i's mog, heirodn dua e's ned.*
Southern Bavarian	*Trotz dass i's mog, hairatn tua i's net.*
Standard German	*Obwohl ich sie mag, heiraten tu ich sie nicht.*

TABLE 2.1: Dialectal and orthographic variation in Bavarian Source: (Blaschke et al., 2024)

2.2 The Notion of Low-Resource

There are many reasons that can lead to a language or a dialect being considered low-resource. The most obvious one is a low number of native speakers, as for the Saterland Frisian which has approximately 2,000 speakers[4], but also political, 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 text, image or speech.

Solving, or at least circumventing, this issue is the main motivation of the current work. It can be argued, that 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).

Even though, in the field of NLP and CL, the term **low-resource** is frequently used to describe languages or language varieties with limited available data or technological support. However, this term can be ambiguous and multifaceted, as it intersects with various concepts from linguistics, sociolinguistics, and language preservation efforts. This section aims to explore different models and considerations that contribute to our understanding of what constitutes a **low-resource** language.

Kornai (2013) proposes a classification system of digitally alive and dead languages that categorizes languages based on their vitality in the digital age. This classification highlights the importance of considering a language's digital vitality when assessing its resource status in NLP contexts.

Still: Languages with very minimal or no digital presence

Heritage: Used for cultural purposes, nut not in everyday communication

Borderline: Limited digital presence, and at risk of digital extinction

Vital: With significant digital resources but less extensive than thriving ones

Thriving: With a strong online presence and active digital communities

The UNESCO Atlas of the World's Languages in Danger (Moseley and Nicolas, 2010) classification provides a more traditional view of language endangerment, focusing on intergenerational transmission and speaker numbers. This classification is valuable for understanding the overall health of a language but may not directly correlate with its resource status in NLP.

Extinct (EX): No living speakers

Critically Endangered (CR): Youngest speakers are grandparents and older

Severely Endangered (SE): Language is spoken by grandparents and older

Definitely Endangered (DE): No longer a mother tongue learned by children

Vulnerable (VU): Most children speak it, but restricted to certain domains

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

Safe (NE): Language is spoken by all generations and not endangered

The Graded Intergenerational Disruption Scale (GIDS) (Fishman, 2001) provides a more nuanced view of language vitality with 8 categories, focusing on intergenerational transmission and societal use of the language. This scale ranges from Stage 8 (most endangered) to Stage 1 (least endangered), considering factors such as literacy, educational use, and community support.

With the Expanded Graded Intergenerational Disruption Scale (EGIDS), Lewis and Simons (2010) expand on Fishman's work, providing a 13-point scale that offers a more detailed assessment of language vitality. This scale ranges from Level 10 (Extinct) to Level 0 (International), incorporating additional factors such as official status, standardization, and institutional support.

Bird (2022) introduces a multipolar model in order to respect local language ecologies, their orality and multilingualism.

Standardized Languages: Major international languages that are fully translatable

Local Languages: Languages primarily used in specific geographic or cultural contexts

Contact Languages: Trade, or vehicular languages as connections between standardized and local languages

Third Spaces: Linguistic environments where multiple languages or varieties coexist and interact

Bird calls for a more local and speaker-centered approach to solving MT problems, in which it shall not be the aim to achieve full digital language equality by the introduction of disruptive language technology (Joshi 2019) in so called under-resourced languages. Instead Bird (2022, p. 9) "suggests a new opportunity for language technology, not how to improve translation for 'under-resourced' languages, but how to support people to work together in third space, and to navigate a metaphysical divide."

The variety of classification systems and concepts presented above demonstrates the complexity of defining what exactly low-resource in language processing is supposed to mean. While NLP often focuses on the availability of digital resources and annotated datasets, these classifications remind us that language vitality is multifaceted. The classifications can involve factors such as number of speakers, intergenerational transmission, digital presence and online communities, institutional support and standardization, sociolinguistic status and domains of use.

On the other hand, classifying a language in the context of NLP and MT will focus on factors such as limited available corpora or annotated datasets, lack of digital tools and resources (e.g., part-of-speech taggers, parsers), absence of standardized orthography or multiple competing orthographies, complex morphology or syntax not well-handled by existing NLP techniques, limited computational resources dedicated to the language. It is crucial to recognize that a language's resource status in NLP may not directly correlate with its vitality or number of speakers. For instance, a language with millions of speakers might still be considered low-resource in NLP if it lacks digital resources or standardized writing systems.

Furthermore, the dynamic nature of language use and technology development means that a language's resource status can change over time. Efforts in language documentation, resource creation, and community engagement can transform a previously low-resource language into one with more substantial NLP support.

Ultimately, when discussing low-resource languages in the context of NLP and MT, it is essential to clearly define what aspects of **low-resource** are being considered. Researchers should be aware of the broader context of language vitality and the potential discrepancies between a language's overall health and its resource status in NLP. This nuanced understanding can inform more effective and culturally sensitive approaches to developing language technologies for diverse linguistic communities in accordance to the actual and real needs and desires of the affected language communities.

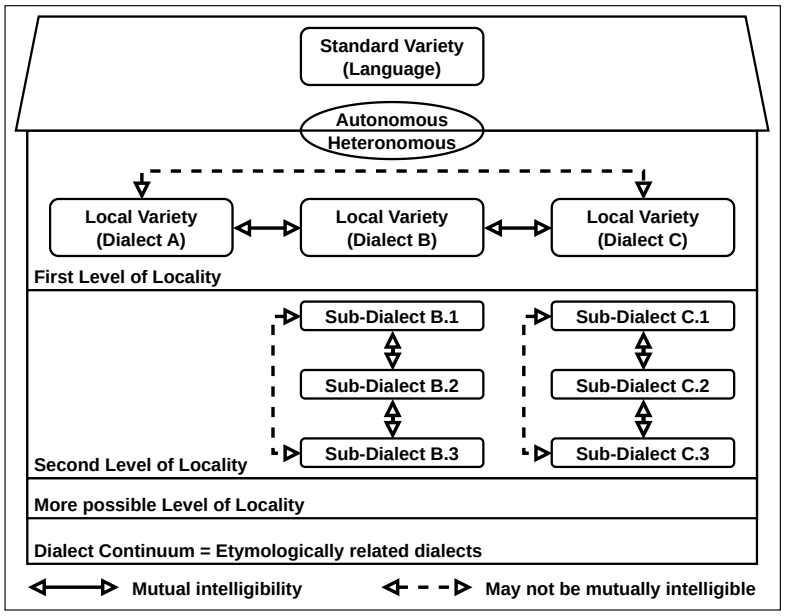


FIGURE 2.5: House of dialects with additional floors for sub-dialects. This figure was created based on the ideas explored in (Bird, 2022; Scherrer, 2012).

2.3 Language Classification Issues

As seemingly custom by now, the exact notation and names of many language varieties are shrouded in mystery and can probably only be found spoken of in the ancient legends of old. Some languages have an ISO-code or are well enough established to have agreed-upon denominations. Others can be found in many different writings, dependent on the authors knowledge and convictions. Then there are many local and to a lesser extend explored languages, that are often named after the location of data collection. All of this leads to an often very inconsistent use of language labels. Modern language identification tools support more and more languages, but are still limited to the existing classifications and data used for training them. Investigations of previously unexplored language varieties get stopped dead in their tracks if simply recognizing the object of interest already requires access to considerable resources and help of a group of native speakers.

Human language is of such a diverse nature that it seems virtually impossible to design a system which perfectly handles each and every instance that can occur be it in formal and often constrained or colloquial and more relaxed settings. Nevertheless the difficulties, coming up with methods that are more language-agnostic promises to benefit many more languages more swiftly than handling them one-by-one. And be it just to have them get their foot into the door of a digitalized house of dialects to then be further explored and refined by future research.

2.4 Machine Translation and Dialects

Preliminary experiments are conducted to assess the performance of modern MT systems on dialect text data. These experiments utilize Google Translate and NLLB (Costa-jussà et al., 2024), various datasets (primarily CODET (Alam, Ahmadi, and Anastasopoulos, 2023)), and the Fairseq implementation of BLEU. Translating the available text data results in new text documents for which the respective word counts can be found in Table 2.3 and further evaluation follow below.

Language	**Translated text**
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.
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.
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.

Table 2.2: 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[5]

Table 2.2 gives an impression of how badly even very popular and large translation systems fail to properly process text from less represented dialects. 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 Saxon and Danube Bavarian (named after the regions from which the data origins), the translations are close to useless and almost consist of simply copying the input text.

The CODET project provides data from five dialects in Bangladesh, namely Jessore, Khulna, Kushtia, Barisal, and Dhaka. As expected, Figure 2.6 indicates that translation systems perform better on dialects closer to Standard Bengali (Jessore, Kushtia) compared to those further away (Khulna, Barisal, Dhaka). Notably, the word count of dialect translations differs by about 13% from the original text (see Table 2.3).

Contrary to expectations based on previous research (Alam, Ahmadi, and Anastasopoulos, 2023), the Mahabad dialect outperforme the Sulaimanya dialect, which is widely used in media. The Erbil dialect shows surprisingly low BLEU scores (see Figure 2.7), despite being considered an epicenter of Central Kurdish varieties. This discrepancy may indicate a significant imbalance in resource availability between the Central Kurdish dialects of the north and those of the south.

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

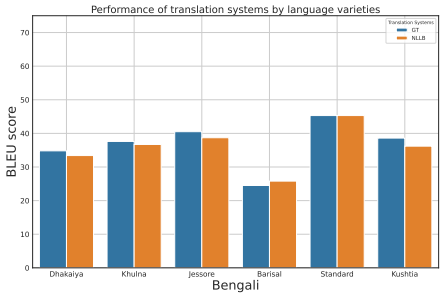


FIGURE 2.6: Language varieties of Bengali (from left to right: Dhakaiya, Khulna, Jessore, Barisal, Standard Bengali, Kushtia).

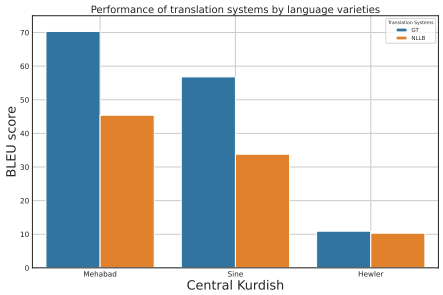


FIGURE 2.7: Language varieties of Central Kurdish (from left to right: Mehabad, Sulaimanya, Erbil).

Language	**Variety**	**Words (Source)**	**Words (GT)**	**Words (NLLB)**
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	Sulaimanya	1880	2075	2492
Central Kurdish	Erbil	1824	2166	2362

TABLE 2.3: Word counts of used data per language and their varieties respectively. Also shown are the translations generated by GoogleTranslate (GT) and (NLLB).

These exploratory experiments demonstrate significant performance disparities in MT systems when handling various dialects, particularly for low-resource scenarios. The results underscore the potential for research in improving dialect-specific MT, which this thesis attempts through leveraging morphological and lexical features in synthetic data generation.

Words are like stories, don't you think? They change as they are passed from mouth to mouth; their meanings stretch or truncate to fit what needs to be said. The Dictionary can't possibly capture every variation, especially since so many have never been written down-

— Pip Williams

3 Related Work

This chapter provides a comprehensive overview of the research field in question and related scientific work. First, a top-down foundation is provided, from Natural language processing and linguistics to Machine translation and its evaluation, before zooming in on specific foci. This chapter explores why it is so difficult to evaluate in low-resource scenarios, how research can still be done, even if data is scarce, and finally, how data can synthetically be enriched based on dialectal features.

3.1 NLP building on Linguistics

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 (Scherrer, 2012).

The term 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 (SA), Questioning & answering (QA), Conversational agent (CA), Machine translation (MT) and many more. The remainder of this work focuses on MT.

3.2 Machine Translation

The investigation of using software to translate text or speech from a source language into a target language falls under MT, which can be considered to be a sub-field of CL. 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 MT, 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 AI make it challenging to predict the future of MT.

Neural Machine Translation By advancing MT via utilizing recurrent neural networks, the area of 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 MT 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. Noteworthy are attempts to push for alternatives such as Chinese-centric (Li et al., 2024) NMT or Afrocentric (Adebara and Abdul-Mageed, 2022) NLP. Sufficient text data of adequate quality is seen as 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.

Machine Translation Evaluation Evaluation of MT systems is a complex and multifaceted task, with numerous metrics and benchmarks having emerged over the years. The diversity of these evaluation methods reflects the challenges inherent in assessing translation quality.

The field of MT evaluation has evolved significantly, starting with what can now be considered **classic** metrics. These include the Bilingual Evaluation Understudy (BLEU) (Papineni et al., 2002; Post, 2018), which remains widely used despite known limitations. Other established metrics include METEOR[1] (Lavie and Agarwal, 2007), which incorporates semantic information, chrF[2] (Popović, 2015), which operates on character n-grams, and Translation Edit Rate (TER)[3] (Snover et al., 2006), which measures the number of edits required to match a reference translation. The choice of evaluation metric can significantly impact the perceived performance of MT systems. For instance, BLEU (see Figure 3.1), while widely used, has known limitations, particularly for low-resource languages. It relies heavily on exact matches and doesn't account well for legitimate variations in translation. TER (see Figure 3.2) provides a more intuitive measure of post-editing effort but may not capture semantic equivalence. chrF (see Figure 3.3), operating on character-level, can be more suitable for morphologically rich languages often found in low-resource scenarios.

[1] http://www.cs.cmu.edu/~alavie/METEOR/
[2] https://github.com/m-popovic/chrF
[3] https://github.com/jhclark/tercom

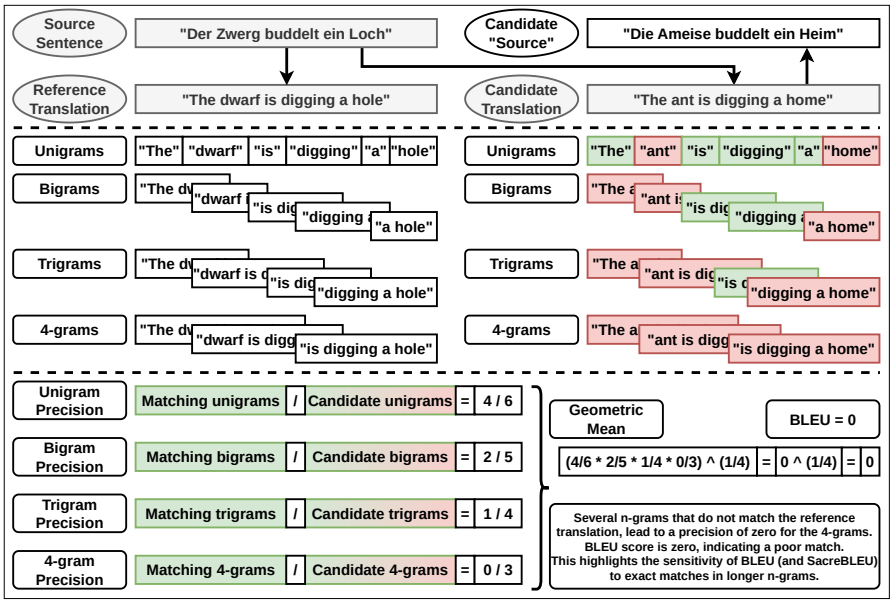


FIGURE 3.1: Evaluation metric BLEU and how it works shown on an example translation.

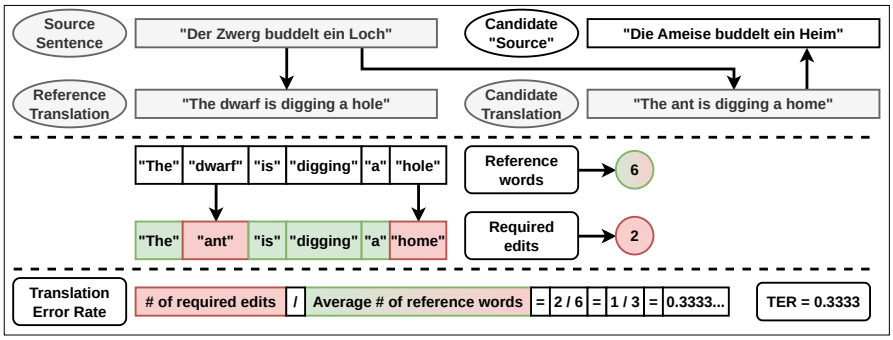


FIGURE 3.2: Evaluation metric TER and how it works shown on an example translation.

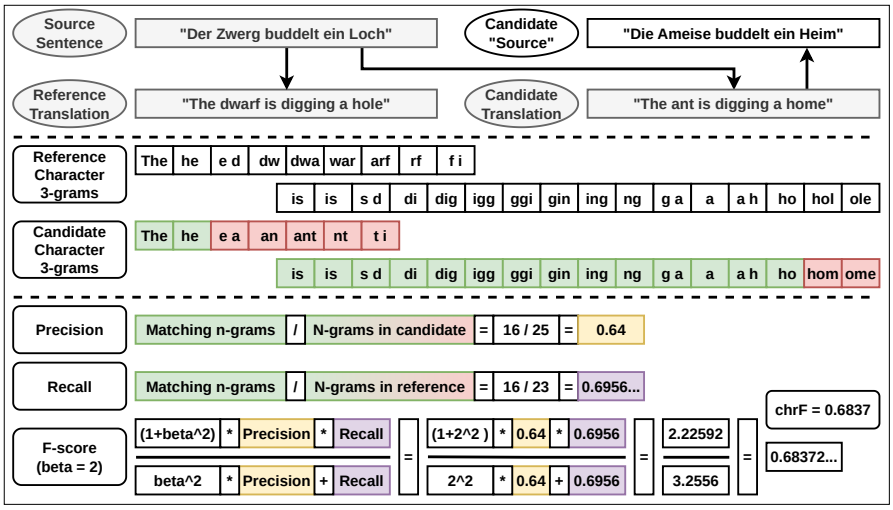


FIGURE 3.3: Evaluation metric chrF and how it works shown on an example translation.

Recent years have seen the development of more sophisticated evaluation methods. BLEURT (Sellam, Das, and Parikh, 2020) leverages pre-trained language models for evaluation. The GLUE[4] benchmark (Wang et al., 2018) and its successor SuperGLUE (Wang et al., 2019) have become standard for evaluating general language understanding, while COMET[5] (Rei et al., 2020) focuses specifically on MT evaluation.

Cross-lingual and multilingual evaluation has gained significant attention, particularly relevant for low-resource scenarios. Benchmarks such as XTREME[6][7] (Hu et al., 2020) and its revised version XTREME-R (Ruder et al., 2021) aim to evaluate cross-lingual transfer capabilities of multilingual models. IGLUE[8] (Bugliarello et al., 2022) extends this concept to vision-and-language tasks. These benchmarks are crucial for assessing how well models perform across diverse languages.

It is important to note that despite the proliferation of automated metrics, human evaluation by native speakers remains the gold standard for MT evaluation, especially for low-resource languages. However, as noted, this approach is often constrained by time and resource limitations. This is also true for most of the advanced metrics and evaluation techniques which require at least some amount of gold-quality reference data in order to be applicable to low-resource or new languages.

3.3 Low-Resource Machine Translation Evaluation

The challenge of evaluating MT for low-resource languages becomes particularly apparent when considering dialect-specific translation. Recent approaches like VALUE (Ziems et al., 2022), Multi-VALUE (Ziems et al., 2023), FRMT (Riley et al., 2023), and CODET (Alam, Ahmadi, and Anastasopoulos, 2023) have been developed to address this specific need. These methods aim to capture the nuances of dialectal variations, which is crucial for many low-resource language scenarios where standard evaluation metrics may fall short.

Evaluating MT in low-resource scenarios presents unique challenges that extend beyond those encountered in general MT evaluation. Recent research has highlighted the limitations of existing evaluation metrics, which are often exacerbated in low-resource contexts. Kumar et al. (2021) and Bapna et al. (2022) raise concerns about the reliability of current evaluation metrics, particularly when applied to low-resource languages. These concerns are further amplified when dealing with dialectal variations, as demonstrated by Alam, Ahmadi, and Anastasopoulos (2023). They present a benchmark specifically designed to evaluate machine translation with dialectal variation on the source side, addressing a critical gap in existing evaluation frameworks. The approach proposed by Riley et al. (2023) focus on language varieties that were included in the model's pre-training data, highlighting the challenges of evaluating MT for truly low-resource dialects that lack standardized representations in large language models. This limitation is further emphasized by Sun et al. (2023) and Aepli et al. (2023), who demonstrate the inadequacies of existing metrics in reliably evaluating text generation outputs for dialects without a standard orthography.

To address these challenges, Aepli et al. (2023) introduce an innovative approach that incorporates character-level noise during metric training. This technique builds upon previous work showing the benefits of such noise in cross-lingual transfer to language varieties lacking standardized orthography (Aepli and Sennrich, 2022; Srivastava

[4] https://gluebenchmark.com/
[5] https://github.com/Unbabel/COMET
[6] https://sites.research.google/xtreme
[7] https://github.com/google-research/xtreme
[8] https://github.com/e-bug/iglue

and Chiang, 2023; Blaschke, Schütze, and Plank, 2023). This approach represents a significant step towards more robust evaluation methods for low-resource dialects.

Despite these advancements, Aepli et al. (2023) emphasize that realistic evaluation of MT still requires obtaining human-translated reference texts and human judgments for translation hypotheses. This underscores the ongoing importance of human evaluation, particularly in low-resource scenarios where automated metrics are shown to fall short.

The issue of dialect preference bias, as documented by Riley et al. (2023) and Abu Farha and Magdy (2022), presents another challenge in low-resource MT evaluation. To mitigate this, Aepli et al. (2023) recommend that recruited annotators and translators should be native speakers of the dialect in question, especially when translating from a standard variant into a dialect. This approach helps ensure that evaluations capture the nuances and preferences specific to the target dialect.

To assess a system's inter-dialect robustness, Sun et al. (2023) propose the use of challenge sets. These sets are designed to compare metric scores across multiple language varieties and between a variety and versions with introduced meaning changes. This method provides a more comprehensive evaluation of MT systems' performance across dialectal variations that can significantly impact translation quality.

While recent research has made significant strides in addressing the limitations of existing metrics and proposing new evaluation frameworks, there is still a defined need for continued research in this area. The combination of innovative automated metrics, carefully designed challenge sets, and judicious use of human evaluation might be the most promising path forward for future research on accurately assessing MT quality in low-resource contexts.

3.4 Approaches to Low-Resource Machine Translation

MT for low-resource languages, both as source and target language, has been a significant focus of research in recent years. This area presents unique challenges due to the scarcity of parallel corpora, limited computational resources, and the linguistic diversity of low-resource languages. Especially the task of translating into low-resource languages and their varieties has attracted considerable attention due to its potential impact on language preservation and accessibility of information. Numerous approaches have been proposed, each with its own benefits and limitations. While some of these methods can be combined for enhanced results, others provide valuable insights that inform the direction of research in this field.

Rule-based Scherrer (2011) pioneering work on translating into Swiss German dialects, demonstrating the feasibility of rule-based approaches for closely related language varieties. This work lays the groundwork for subsequent research on dialect translation. Haddow et al. (2013) explore the use of pivot languages for translating into low-resource languages, a technique that has proven effective when direct parallel corpora are scarce. Their work demonstrates how high-resource languages could be leveraged to improve translation quality for low-resource target languages.

Pivot-based Fancellu, Way, and O'Brien (2014) investigate the challenges of translating into minority languages, focusing on the specific case of Scottish Gaelic. Their work highlights the importance of considering morphological complexity and domain-specific vocabulary in low-resource MT.

Unsupervised Myint Oo, Kyaw Thu, and Mar Soe (2019) focus on MT for Myanmar ethnic languages, highlighting the challenges and opportunities in developing MT systems for extremely low-resource languages within a multilingual country. Wan et al. (2020) propose innovative techniques for unsupervised NMT for low-resource languages, demonstrating how monolingual data could be leveraged to improve translation quality in the absence of parallel corpora.

Large language model Garcia and Firat (2022) explore the use of multilingual models and few-shot learning for translating into low-resource languages. Their work showcases the potential of large pre-trained models in addressing the data scarcity problem in low-resource MT.

Several strategies have been proposed to enhance translation quality in low-resource scenarios. These approaches can be broadly categorized into data augmentation techniques, model adaptation methods, and innovative architectures for cross-lingual transfer.

Data augmentation It has been a key focus area, with back-translation emerging as a prominent technique, to augment data. Sennrich, Haddow, and Birch (2016) introduce the use of monolingual data to improve neural machine translation models through back-translation. This concept is further explored by Edunov et al. (2018), who investigate back-translation at scale. Dou, Anastasopoulos, and Neubig (2020) propose dynamic data selection and weighting for iterative back-translation, employing TF-IDF to select relevant sentences. In a related vein, Zhang et al. (2018) present a method for joint training of neural machine translation models with monolingual data.

Adaptation Model adaptation techniques have also shown promise in low-resource settings. Bapna and Firat (2019) introduced a simple and scalable adaptation approach for neural machine translation. Pfeiffer et al. (2020) propose MAD-X[9], an adapter-based framework for multi-task cross-lingual transfer. Cooper Stickland, Li, and Ghazvininejad (2021) offers recipes for adapting pre-trained monolingual and multilingual models to machine translation tasks. Üstün et al. (2021) explore multilingual unsupervised neural machine translation using denoising adapters.

Transfer learning Several studies focus on innovative architectures and approaches for cross-lingual transfer. Ansell et al. (2023a) introduce composable sparse fine-tuning for cross-lingual transfer[10], incorporating a variant of the Lottery Ticket Hypothesis. The same authors also investigate distilling efficient language-specific models for cross-lingual transfer (Ansell et al., 2023b). Garcia et al. (2021) explore harnessing multilinguality in unsupervised machine translation for rare languages. Costa-jussà, Zampieri, and Pal (2018) explore the use of NMT for similar languages, proposing techniques to leverage the similarities between closely related languages to improve translation quality. This approach is particularly relevant for low-resource languages with more resourced relatives. In a similar vain, Lakew, Erofeeva, and Federico (2018) investigate the use of transfer learning and multilingual models for low-resource NMT. Their work demonstrates how knowledge from high-resource language pairs could be transferred to improve translation for low-resource languages.

[9] https://adapterhub.ml/

[10] https://github.com/cambridgeltl/composable-sft

The limitations of zero-shot language transfer with multilingual transformers are examined by Lauscher et al. (2020), while Parović et al. (2022) propose BAD-X[11], a method using bilingual adapters to improve zero-shot cross-lingual transfer.

Representation learning Another promising strategy is proposed by Reimers and Gurevych (2020), introducing 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 the outlook of their model being extendable to multiple languages in the same training process.

Linguistically driven Linguistic knowledge incorporation is also explored, with Casas et al. (2021) proposing linguistic knowledge-based vocabularies for neural machine translation. The seminal work of Vaswani et al. (2017) on the Transformer architecture continues to influence the field, with adaptations for low-resource scenarios such as the factored neural machine translation approach proposed by Bandyopadhyay (2020).

For extremely low-resource situations, Karakanta, Dehdari, and Van Genabith (2018) investigate neural machine translation without parallel corpora. Vries et al. (2021) explore adapting monolingual models in scenarios where data is scarce but language similarity is high. A large-scale effort to address translation for low-resource languages is undertaken by Costa-jussà et al. (2024) in their **No Language Left Behind** project, resulting in a publicly released framework called fairseq[12] and language models on Huggingface[13].

These diverse approaches demonstrate the ongoing research efforts to improve translation quality in low-resource settings and while significant progress has been made, many challenges remain. Future research directions may include further exploration of cross-lingual transfer learning, improved techniques for leveraging monolingual data, and the development of more efficient models that can perform well with limited computational resources. As the field continues to evolve, it is likely that a combination of these approaches, along with novel techniques leveraging recent advancements in large language models and few-shot learning, will lead to further improvements in MT for low-resource languages and their varieties.

3.5 Synthetic Text Data Generation

The generation of synthetic text data has emerged as a crucial strategy for improving MT, particularly in low-resource environments. This approach has been recognized and developed in numerous studies over the past decade, offering significant benefits while also acknowledging potential risks and limitations.

Early work by Foster and Andersen (2009) lays the groundwork for utilizing synthetic data in linguistic research. Followed by Ha, Niehues, and Waibel (2016) proposing a universal encoder and decoder approach for multilingual NMT, leveraging shared linguistic information across multiple languages. Since then, the field has seen a proliferation of sophisticated data augmentation techniques. Xie et al. (2017) and Gao et al. (2019) explore various methods of text augmentation to improve model robustness. Xia et al. (2019) and Duan et al. (2020) further advance these techniques, applying them specifically to

[11] https://github.com/parovicm/BADX
[12] https://github.com/facebookresearch/fairseq/tree/nllb
[13] https://huggingface.co/docs/transformers/model_doc/nllb

low-resource scenarios. Sánchez-Cartagena et al. (2021) demonstrate the effectiveness of synthetic data in improving NMT for low-resource languages.

The application of synthetic data extends beyond direct text augmentation. Malykh, Logacheva, and Khakhulin (2018) utilize synthetic data to create robust word vectors, while Doval, Vilares, and Gómez-Rodríguez (2020) focus on developing robust embeddings. These approaches have significantly enhanced the quality of word representations, particularly beneficial for low-resource languages.

In the realm of NMT, synthetic data has proven to be a game-changer. Artetxe, Labaka, and Agirre (2018) demonstrate the potential of unsupervised MT using only monolingual corpora, effectively leveraging synthetic parallel data. Ngo et al. (2022) further explore this concept, focusing on improving NMT for low-resource languages. Bogoychev and Sennrich (2020) investigate the use of synthetic data in multilingual NMT systems, showing improvements in translation quality across multiple language pairs.

The versatility of synthetic data is evident in its application to various NLP tasks. Dekker and Goot (2020) employ synthetic data for lexical normalization, addressing the challenge of non-standard language varieties. Ahmadi and Anastasopoulos (2023) utilize synthetic data for script normalization, a crucial task for many low-resource languages with multiple writing systems. Lusito, Ferrante, and Maillard (2022) apply similar techniques to text normalization for Ligurian, demonstrating the potential of this approach for endangered languages.

A notable advancement in the field is the ZEROGEN approach being proposed by Ye et al. (2022). This method focuses on efficient zero-shot learning via dataset generation, potentially revolutionizing how we approach low-resource language tasks.

In the context of dialectal variations, Scherrer (2012) explore an innovative approach to generating Swiss German sentences from Standard German, employing a multi-dialectal strategy. This work highlights the potential of synthetic data in bridging the gap between standard languages and their dialectal variants.

3.6 Bilingual Lexicon Induction

Bilingual lexicon induction (BLI) has emerged as a crucial method in NLP to bridge the gap between languages and enable high-quality machine translation. Although earlier works may exist in this field, the foundations of BLI can be traced back to the works of Yamamoto, Matsumoto, and Kitamura (2001), followed by (Fung and Chen, 2004; Sahlgren and Karlgren, 2005; Caseli, Nunes, and Forcada, 2006) and henceforth continuously being further developed and refined (Lardilleux, Gosme, and Lepage, 2010; Scherrer and Cartoni, 2012; Scherrer and Sagot, 2013; Irvine and Callison-Burch, 2013). Czarnowska et al. (2019) describe BLI as a well-established choice for evaluating cross-lingual word embedding models.

Recent advancements in BLI focus on addressing challenges in low-resource and closely related language pairs. Bafna et al. (2023) present methods for unsupervised BLI for data-imbalanced closely related language pairs, while Waldendorf et al. (2022) explore the use of BLI to improve the translation of out-of-vocabulary words in low-resource machine translation.

The intersection of BLI with LLMs has opened new avenues for research. Li, Korhonen, and Vulić (2023) leverage LLMs for developing bilingual lexicons, while Artemova and Plank (2023) extend this approach to low-resource language varieties, such as German dialects.

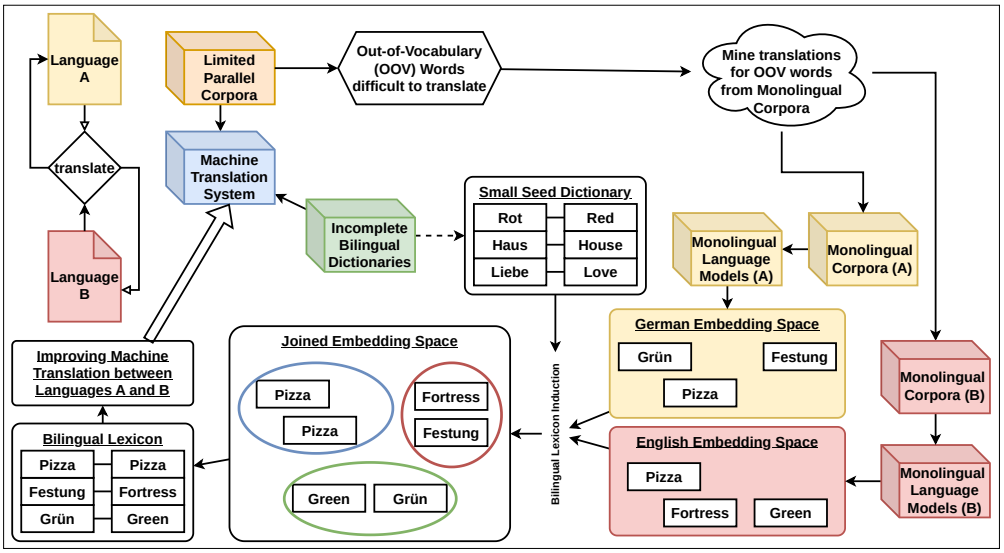


FIGURE 3.4: Conceptualization of bilingual lexicon induction utilizing mainly monolingual data with the aim to improve machine translation. Inspiration for this figure is drawn from (Irvine and Callison-Burch, 2017; Wang, Fan, and Frantzen, 2021).

Cross-linguistic research has led to the development of massively multilingual datasets, such as CogNet (Batsuren, Bella, and Giunchiglia, 2022). These resources provide direct links between languages, avoiding the need to use English as a pivot language. This is particularly advantageous as English, being morphologically poor, is less suitable for analyzing morphological generalization (Czarnowska et al., 2019).

The development of cross-lingual word embeddings has been crucial to the advancement of BLI. Vulić and Moens (2013) explored cross-lingual semantic similarity of words as the similarity of their semantic word responses. They further develop this concept in their Vulić and Moens (2015) Vulić and Moens (2015) work on bilingual word embeddings from non-parallel document-aligned data. Gouws, Bengio, and Corrado (2016) introduce BilBOWA, a method for fast bilingual distributed representations without word alignments.

Recent work also focus on improving BLI for low-resource scenarios. Waldendorf et al. (2022) investigate improving the translation of out-of-vocabulary words using BLI in low-resource MT. Bafna et al. (2023) proposeda simple method for unsupervised BLI for data-imbalanced, closely related language pairs. Czarnowska et al. (2019) provide a comprehensive analysis of morphological generalization in BLI, emphasizing the importance of considering the **long tail** of less frequent words.

In conclusion, BLI has evolved from its early foundations to become a crucial tool in cross-lingual NLP, with recent advancements focusing on low-resource languages, the integration of LLMs, and improved techniques for closely related language pairs. Figure 3.4 provides an intuitive display of the interplay between BLI and MT.

3.7 Linguistic Dialectal Features for Translation

The intersection of linguistic features, dialectal variations, and their application in translation has been a growing area of research in recent years. Baroni (2019) provides a comprehensive overview of the history of modern artificial neural networks and their role in linguistic generalization, including insights into their compositionality. The importance of addressing dialectal variations has been increasingly recognized in the field.

Dialect feature detection Demszky et al. (2021) emphasize this need, proposing an approach to learn dialect feature detection for English variations based on a small set of minimal pairs. This method circumvents the requirement for large-scale annotated corpora, which are often unavailable for many dialects.

Benchmarking and evaluation A series of influential publications from Stanford University has significantly contributed to this area. Ziems et al. (2022) introduce VALUE, a challenging variant of GLUE[14] (Wang et al., 2018), designed to understand disparities in current models and facilitate more dialect-competent NLP systems. VALUE expands on established benchmarks that previously contained only Standard American English, such as GLUE and SuperGLUE (Wang et al., 2019). The initial release of VALUE focused on 11 linguistic features of African American Vernacular English, validated by fluent speakers through linguistic acceptability judgments in a participatory design approach. Building upon this work, Ziems et al. (2022) develop Multi-VALUE (Ziems et al., 2023), expanding the scope to 50 different dialects of English and 189 unique linguistic features. This comprehensive benchmark addresses the previously noted lack of systematic studies on cross-dialectal model performance, though, yet again very English-centric and not trivially transferred to languages with fewer resources.

Model adaptation Liu, Held, and Yang (2023) further advance this line of research with DADA, which builds on Multi-VALUE. DADA proposes adapting trained language models via feature adapters, each corresponding to a specific linguistic feature. This approach demonstrates the potential for reusing feature adapters across various language variants, aligning with the concept of flexible boundaries between dialects.

Cross-task transfer The rationale for these approaches is supported by Held, Ziems, and Yang (2023), who argue that current methods for improving dialect robustness, particularly in English, have been limited by their focus on single tasks. They propose a more scalable approach enabling task-agnostic zero-shot transfer, using perturbations from Multi-VALUE and dialect variants of GLUE to empirically demonstrate its effectiveness.

Data synthesis Recent work by Alam, Ahmadi, and Anastasopoulos (2024a) proposes an innovative approach to address data scarcity in low-resource scenarios. Their method synthesizes parallel data by combining morpho-syntactic information and bilingual lexicons, utilizing a small amount of seed parallel data. This approach has shown improvements even with as few as five seed sentences combined with a bilingual lexicon, as demonstrated on the English-Kurmanjî language pair. However, a significant limitation of this method is its reliance on the Stanza model, which only supports a very limited range of languages.

Linguistic resources These developments highlight a current trend in research in which high-resource languages are leveraged to benefit their low-resource counterparts, especially dialectal variations. While much of this work focuses on English variants, it provides a framework that can guide similar research on other languages in comparable settings. Despite how exciting and promising they are, most of these approaches can not simply be applied to new target languages, especially in the limited scope of a thesis. For example, the World Atlas of Language Structures Online (WALS) (Dryer and Haspelmath, 2013), a well-established resource for linguistic features provides information for

[14] https://gluebenchmark.com/

159 features for English, and 157 for German. However, looking at dialect-level there are 17 varieties of German listed, each with only exactly a single linguistic feature. Hence, many of the discussed approached entail an initial dialectal feature extraction, which, in turn, necessitates available datasets that have to be specifically filtered or initially labeled on dialect-level. This is currently required for each new, to be explored, target language, making the development of a general, more language-agnostic method even more desirable.

To conclude this little sidestep into the realm of the related work of this thesis, the field of linguistic features and dialectal variations in translation is rapidly evolving, with recent research focusing on developing more comprehensive benchmarks, adaptive models, and innovative data synthesis techniques. These advancements are particularly crucial for guiding development of NLP for low-resource languages and dialectal variations, paving the way, even if just stone-by-stone, for more inclusive and effective NLP systems.

The Dictionary is a history book. If it has taught me anything, it is that the way we conceive of things now will most certainly change.

— Pip Williams

4 Methodology and Data

Now, that the related work has been unfolded, this chapter follows by outlining the goal of this thesis. A methodology is proposed and first challenges are faced, before the applied methods and used data for pursuing the outlined goals are described in detail.

Low-resource languages and their dialectal varieties continue to pose substantial challenges for Machine translation. This thesis aims to address this gap by developing a novel, data-driven approach to improve MT for language varieties in low-resource scenarios, with the ambitious objective to minimize the need for extensive linguistic expertise and native speaker involvement, which are costly and time consuming. To achieve this, we propose a language-agnostic method that can potentially be applied to a wide range of language pairs, thereby benefiting numerous language communities.

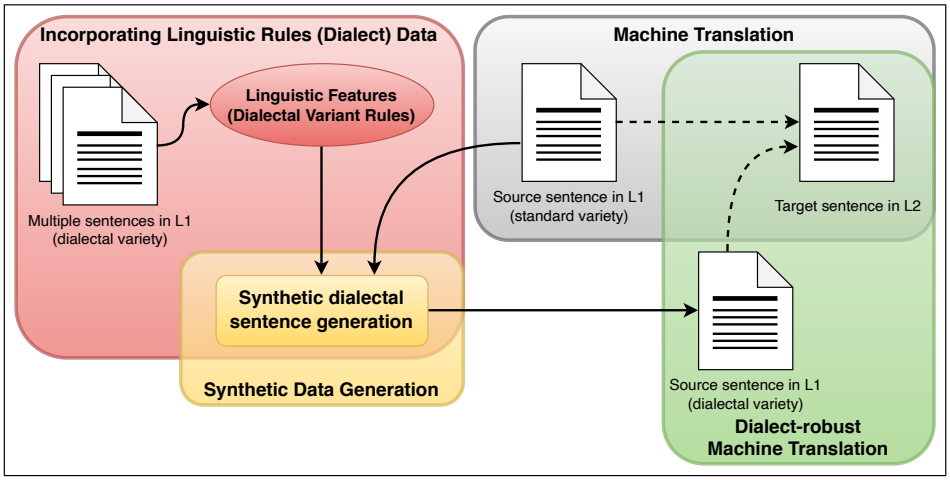


FIGURE 4.1: Concept of this work for arriving at dialect-robust machine translation for low-resource language variants via synthetic data generation.

The results of this thesis hope to help pave the way to developing dialect-robust MT systems (see Figure 4.1). We exemplify this method using data for German varieties, in particular Bavarian dialects.

4.1 Proposed Methodology

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.

This research presents a novel, data-driven approach for cross-lingual transformation and NMT enhancement, with a particular focus on low-resource language scenarios. The method is designed to be *language-agnostic* and minimally dependent on linguistic expertise or native speaker input, making it readily applicable to a wide range of language pairs. This methodology aims to provide a *reproducible* framework that researchers can apply to diverse language pairs, particularly benefiting low-resource scenarios and languages with limited linguistic documentation. The proposed methodology has developed into five main phases, presented below.

4.1.1 Data Acquisition and Alignment

For the following steps, two types of data are required. First, monolingual data for each target language, and second, a seed dictionary of aligned words across the target languages. We derive additional aligned text data for the target language varieties on word-level via utilizing BLI. This process can potentially be expanded to the sentence-level in future work, further discussed in Section 6.2.4.

The initial data for this work is drawn from multiple corpora, acquired via OpusTools. OpusTools provides cross-lingually aligned text data for a large number of languages. For Bavarian a number of aligned entries (many sentences, sometimes just single words) could be collected for German and English. As of the time of writing, the Alemannic language itself is not included, on the other hand, one of its subdialects, Swabian can be found. The scope of this thesis though, prevents the consideration of each angle and level of locality at this point, further discussed in Section 6.2.2.

4.1.2 Linguistic Feature Extraction

The identification of dialect features requires considerable linguistic expertise and definitions only exist for a very limited number of languages (Sun et al., 2023). Hence, this thesis approaches the acquisition of linguistic features from a *data-oriented* angle. Deriving cross-lingual transformation rules happens in two stages (see Figure 4.2). First, sub-word alignment and divergence detection through string matching. Second, context-aware rule filtration and validation.

The idea is such that a set of practical criteria are used to narrow down and improve the quality of the rules, *ergo*: heuristic-based rule refinement. More specifically, using contextual expansion and frequency thresholding in order to derive a data-driven functionality. This phase results in a set of dialectal rules (or linguistic features) specific to the language pair.

4.1.3 Synthetic Data Generation

Applying the derived linguistic features perturbs existing data, thereby synthetically creating novel data for each language variety. Starting from a more dominant language variety (often the standard variety) and changing the text to resemble the low-resource

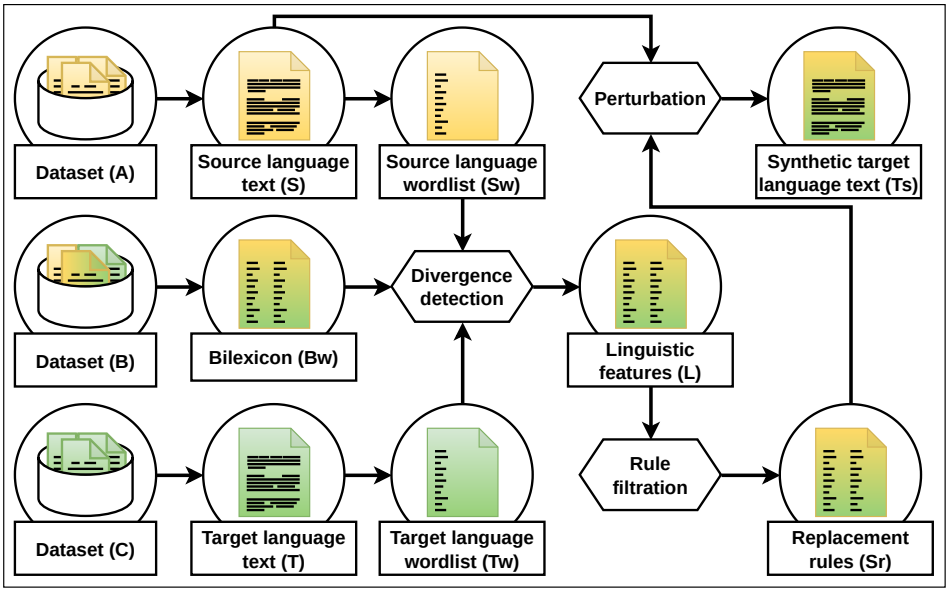


FIGURE 4.2: Methodology overview to derive replacement rules to then perturb text, transforming them between language varieties.

variant (usually called dialect) is a process known as **dialectalization**. In cases where the available text that was produced by native speakers of a dialect is transformed to rather resemble the standard variant, the process is called **standardization**. Once the linguistic ravine between language varieties is bridged, the ability to transform freely between the different varieties promises to result in considerably higher value gain for the low-resourced language varieties. Figure 4.3 shows a high-level overview of the relations between language varieties and pivot language involved in the creation and evaluation of synthetic data.

This rule-based system works by morphosyntactically analyzing the words aligned between the standard and the dialectal variety and sequentially applying a set of dialect-specific rewriting rules to generate perturbed output. We apply string-matching on aligned words to derive linguistic features describing differences between language varieties. From these, we select a suitable set of replacement rules for text perturbation. Lexical and sub-word replacements are applied to perturb Standard German text into Bavarian variants and correspondingly perturb Bavarian text to Standard German.

Due to the lack of specialized test suites for dialectal variants (discussed in Section 3.3), we translate different versions of the same text data into English and evaluate performance using reference translations. Our evaluation metrics include BLEU, TER, and chrF scores. These scores are calculated using the provided implementation in Sockeye (Hieber et al., 2018; Hieber et al., 2020; Hieber et al., 2022), using the recommended default parameters[1].

To compare Bavarian NMT, we examine results of current state of the art represented by Her and Kruschwitz (2024). Their approach, which ignores sub-dialectal differences, still achieves decent results, applying a training with n-fold cross-validation in combination with back-translation or transfer learning to assess their effect on translation directions. Her and Kruschwitz (2024) report a higher improvement and overall better performance of translating into Standard German, while translating into the low-resource variant stayed behind in each experimental setup and suspect multiple sub-dialects as a possible reason for this.

[1] https://github.com/awslabs/sockeye

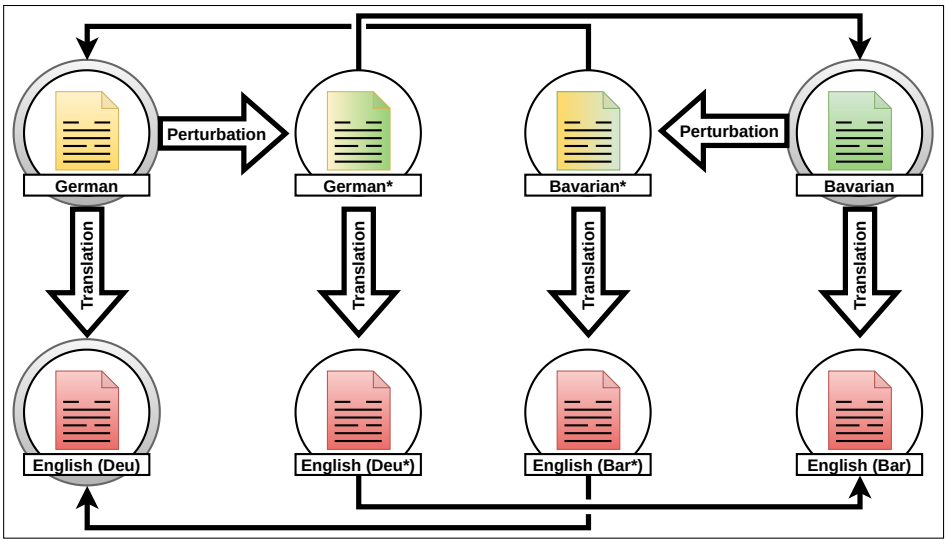


FIGURE 4.3: Simple overview of the experimental setup that aims to transform (perturb) text from one language variety into another. Comparing translations of the perturbed data can help evaluate the method.

We attempt to clean Wikidump data and filter for sub-dialects, similar to Lambrecht, Schneider, and Waibel (2022)'s work that lead to great improvements for Alemannic data from Wikipedia. Due to the scope of this thesis, instead of training a dialect classifier for each available sub-dialect, we filter the data based on word frequencies observed in already tagged article contents (see Appendix B). However, this process, though a promising next step, proves time-consuming and beyond the scope of this work.

Feature Validity The lowest level (dubbed internally as **Guess**) is a very simple approach as it is based on automatic functions and data-driven 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 are high in number (thousands) but also include single character replacements and removals without concern for the context inside the text. Mainly intended as a tech demo, the use of these rules leads to results of such low performance, that they can serve as something of a sanity check being the lower-bound of that, which this part of the method can achieve.

More sophisticated stands the level called **Reason** as 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 comparable approach to include the context for the lexicographic replacements still needs testing and more data to produce enough applicable rules. Currently this window has a length of 1 in each direction. Therefore, now the aligned word-pair *fochgebiet* → *fachgebiet* would not result in a rule replacing $o \rightarrow a$, but *foc* → *fac*.

The last level introduced for this thesis is internally called **Relaxed** and applies almost the same rules as the aforementioned **Reason**, but now with loosened exclusion criteria. In this setting, the derived features can have a length of 6 characters and only have to appear once in the data instead of 5 times in order to be included. The idea behind this level is to capture as much information from the limited data as possible. It

can be expected, that this will also introduce additional noise, which will degrade the evaluation, but comparing the different applied metrics might reveal new insights into the performance of our method.

Perturbation Types **Lex** denotes lexicographic replacements of entire words based on bilingual word lists (see Table 4.1 for examples in either direction). This type of perturbation is limited by the scope of the available bidictionary and can only affect known words. Though, less impact full, does it bring the advantage of disturbing less words compared to the replacement of sub-word units. The selected example shown in Table 4.1 showcases, that this is not a guarantee, as missing context can still lead to faulty replacements: Neither of the two Bavarian *de*, even if such a rule exists, should be replaced with the German *drei* (eng.: three) in this sentence.

German	**Perturbed (dialectized)**
Eine ***Gemeindeverbindungsstraße führt*** *zur Kreisstraße AN 52 nach* ***Aichau*** *bzw. Böckau.*	*Eine* ***Gmoavabindungsstroßn fiat*** *zur Kreisstraße AN 52 nach* ***Thürnhofen*** *bzw. Böckau.*
Bavarian	**Perturbed (standardized)**
Duachn Ort valafft ***de*** *Kreisstroß* ***AN*** *52,* ***de*** *noch Unterahorn oda noch Böckau* ***fiaht****.*	*Duachn Dorf valafft* ***drei*** *Kreisstroß* ***vielen*** *52,* ***drei*** *noch Unterahorn oda noch Böckau* ***verläuft****.*

Table 4.1: Examples of the perturbed German-Bavarian texts based on lexicographic rules. English: A community connection road leads to the AN 52 district road to Aichau or Böckau.

Mor denotes morphological replacements of sub-word units based on rules derived by processing bilingual word lists (see Table 4.2 for details).

German	**Perturbed (dialectized)**
Eine ***Gemeindeverbindungsstraße führt zur Kreisstraße AN*** *52* ***nach*** *Aichau bzw. Böckau.*	***Einen Gemeindevabindungsstroß führts zua Kreisstroß on*** *52* ***noch*** *Aichau bzw. Böckau.*
Bavarian	**Perturbed (standardized)**
Duachn Ort valafft de Kreisstroß AN *52,* ***de noch*** *Unterahorn* ***oda noch*** *Böckau* ***fiaht****.*	***Duchen Orts valaffts den Kreisstraße Ans*** *52,* ***den nachs*** *Unterahoren* ***oder nachs*** *Bösckaus* ***fiahts****.*

Table 4.2: Examples of the perturbed German-Bavarian texts based on rules derived from subword-units. English: A community connection road leads to the AN 52 district road to Aichau or Böckau.

All denotes the combination of both previous replacements by applying morphological ones after the lexicographic ones. During this process, the lexicographic replacement function marks each replaced word (by adding special characters **@@@** to the front of it), such that they are ignored during the morphological replacement to not distort properly replaced words as an aftereffect.

The limited scope of this thesis necessitated restrictions on investigated items. Complex linguistic phenomena, such as syntactic changes between language varieties, were excluded as they would require more elaborate parsing techniques and substantially more data to be applicable without linguistic expertise.

Language Variety Granularity Level The acquired results indicate that the proposed approach in combination with currently available data is very limited, likely due to sub-dialectal noise in the data (further discussed in Section 6). For example, the prefix *ein* (eng.: one) from the German side is observed to be aligned on the Bavarian side with *ei*, *eih*, and *oa*, which corresponds to the descriptions of different Bavarian sub-dialects. This highlights the need for more refined data cleaning and filtering processes. The degree of data scarcity vastly differs between language varieties (Alam, Ahmadi, and Anastasopoulos, 2024b) 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 and this approach in it's entirety.

The current granularity level uses the data as it was provided from various sources and labeled on language level (such as German, Kurdish, English) and sometimes dialect level (such as Alemannic, Bavarian, Kurmanjî, Soranî). 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.

4.1.4 Evaluation Framework Development

To evaluate the outcomes and ensure the validity of experiments and their methods is always of utmost importance in scientific research. Due to the lack of meaningful evaluation data or actual benchmarks, and even established evaluation metrics for MT having difficulties in assessing dialectal or in general low-resource language data, certain considerations have to be made in this line of research. In cases where no data exists to enable a direct comparison of the results and no native speakers are accessible for a study involving human participants, indirect approaches can help to provide a partial remedy. One such approach for meaningfully evaluating results in the absence of standard benchmarks or high-quality reference data is to include a new language over which to pivot.

For example, to evaluate a newly trained model for translating between Kobanî and Mauritian Creole, two languages for which no parallel data yet exists, the produced translations might be compared with outputs generated by two other models, which translate between Kobanî-English and Mauritian Creole-English respectively. Naturally, this introduces new sources of possible errors and inconsistencies, but still, it enables at least some kind of evaluation. Similar to this is the idea of translating original and perturbed text into a third language to then compare how close to each other these translations are (see Figure 4.3).

4.1.5 Limits of Available Data

About identified sources:

- OpusTools (Tiedemann, 2009)
  - This source is well established in the field of MT, providing aligned text data for a large number of language pairings, such as Bavarian-German[2] and Bavarian-English[3].
  - The coverage of dialectal variants is currently still quite limited.

[2] https://opus.nlpl.eu/results/bar&de/corpus-result-table
[3] https://opus.nlpl.eu/results/bar&en/corpus-result-table

- DialectBLI (Artemova and Plank, 2023)
  - Bitexts and bidictionaries for Alemannic-German and Bavarian-German that are automatically derived from Wikipedia articles.
  - Due to the origin of the texts, this data is severely restricted in domain and style.
- Wikidump[4]
  - Wikipedia articles and their meta data, which can serve the name of local language varieties, if the original author provided this information.
  - This data, again, is very restricted in domain and style.
- World Atlas of Language Structures Online (WALS) (Dryer and Haspelmath, 2013)
  - This collection of linguistic features would be exactly what is needed for this line of work and will at some point enable future research in a similar direction.
  - As it stands, dialectal variations (at least for German) are close to non-existent.

4.1.6 Data Quality and Cleaning

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 are incorrectly-aligned, sentences that are of low quality, and even text from different languages. Automatic corpus creation methods used to quickly output large amounts of data, can lead to corpora in which the aligned sentence is the same in both languages or the assigned language labels getting mixed up. As a consequence, sometimes sentences entirely consisting of Chinese or Bengali characters are labeled to be German or Bavarian text.

Clean is the data that has gone through 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. Given our focus on dialects, we assume that a sentence more than 3 times the length of another sentence, can hardly be considered to be well-aligned.

Additionally, manual review of the collected data confirms that the processing of the original corpus creators does not suffice and that further automated cleaning steps are required. These include as a first step the removal of entries that contain more than 20% non-German characters, which indicates some kind of mix up during corpus creation. Next, all entries, where one sentence is found to be more than 3 times as long as its aligned counterpart is excluded from the data (some examples shown in Figure 4.4), as well as extremely long entries (some entries exceed many thousands of characters, such as exhaustive lists of locations and links, sometimes found on Wikipedia pages). Additionally, parenthesized text, or text found in brackets is removed, as well as programming codes, HTML and website-related content, and special characters. Considering the origin of the data, these exclusions affected parts of the texts, but also entire entries in the aligned text data (see Figure 4.5).

This chapter details the experimental journey undertaken during this thesis. Starting with an exploration of various sources of data and tools for bilingual lexicon induction.

[4] https://dumps.wikimedia.org/

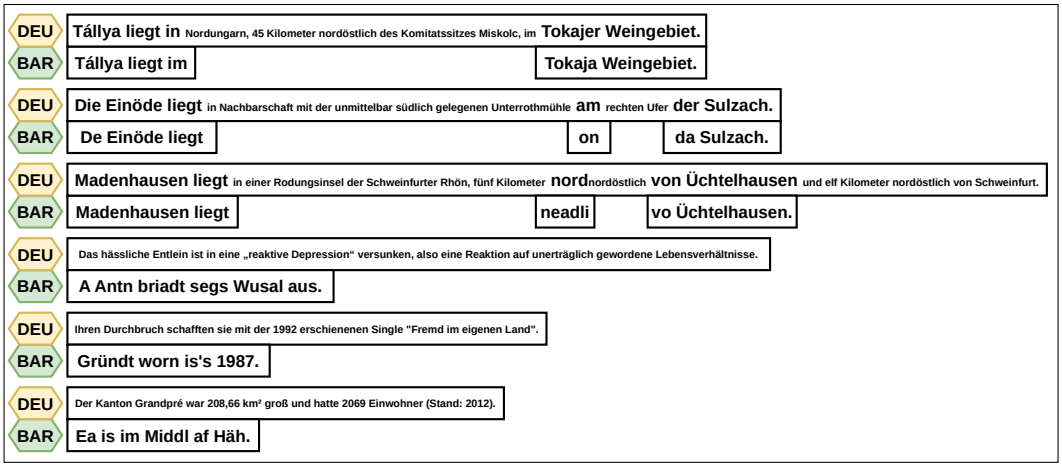


FIGURE 4.4: Examples of aligned Bavarian and Standard German sentences that are excluded from the dataset due to noise.

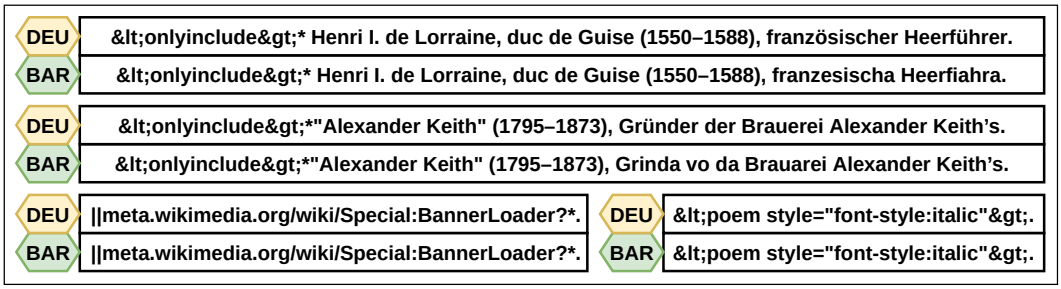


FIGURE 4.5: Examples of noise in aligned data containing some type of code and URLs, motivating the development of a script-solution for handling many different fringe cases of noise.

Then follows the extraction of linguistic features based on string matching and the selection of suitable replacement rules for creating perturbed text. Identifying challenges in developing a more language-agnostic method for cross-lingual transformation and dialectal machine translation enhancement.

4.2 An Approach with Multiple Angles

Conceptualization of an experimental setup which covers multiple angles of exploration (see Figure 4.6). Considering the direction of a translation system and the effect this can have on the quality, it stands to reason, to investigate the performance from three different angles. First, the translation from the standard variant into English, here dubbed **baseline**. Second, the transformation from the dialectal variant into the standard variant (**standardization**), prior to translating into English, here called **preprocess**. Third, the translation from English into the standard variant prior to transforming into the dialectal variant (**dialectization**), here called **postprocess**.

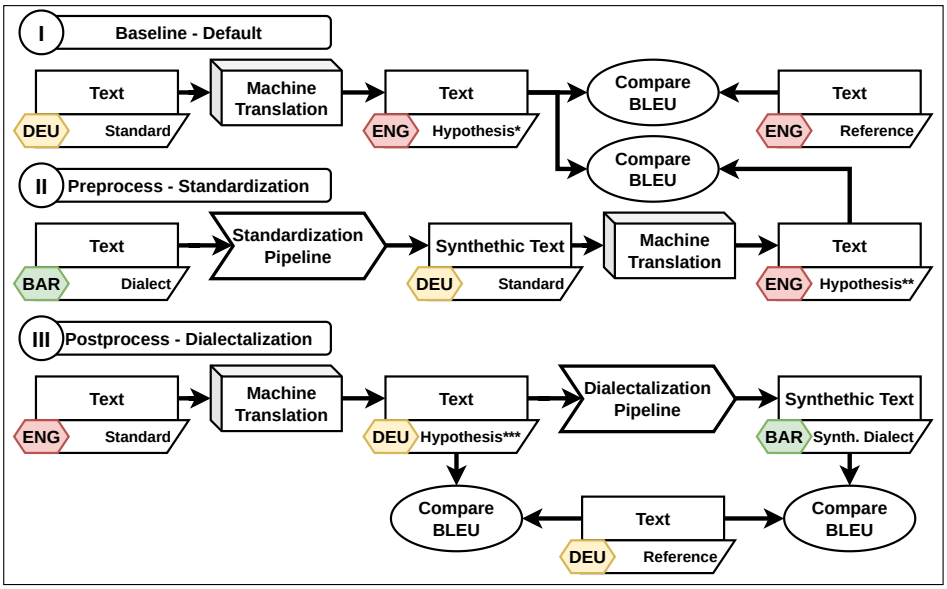


Figure 4.6: Experimental setup combining three approaches.

For a foundation of applicable data we first aim to extract aligned words for linguistic feature extraction. We explore the following methods:

- The FastAlign tool can be used to extract aligned words from aligned texts. It can be applied to any language, but is severely limited the amount of available data. Experiments involving sorting the tools output by frequency of occurrence reveal a lot of noise and ill-aligned words unsuitable for extraction of high quality replacement rules.
- We attempt to make use of this large language model (ChatGPT) to translate German words into Bavarian and to collect lists of prefixes and suffixes. While initially promising, this approach proves cumbersome. The model produces considerable amounts of noisy and faulty outputs requiring significant manual effort and linguistic expertise to sort out, contradicting our goal of a language-agnostic method.

4.3 Utilizing Data from DialectBLI

We shift to using German-aligned words from DialectBLI (Artemova and Plank, 2023) for Alemannic-German and Bavarian-German pairs. The provided data is first filtered by the datasets own quality labels, partially based on human annotations, partially on language model estimations. A manual review of randomly selected data samples prompted additional cleaning (following the descriptions in Section 4.1.6) aimed to remove as much noise as possible, and simultaneously accepting further reduction of the available data quantities as a trade-off. Even though, this data still requires substantial cleaning effort due to noise, it enables us to investigate the previously proposed method despite a working BLI-method. The path to acquire synthetic data developed during the work step-by-step (see Figure 4.7).

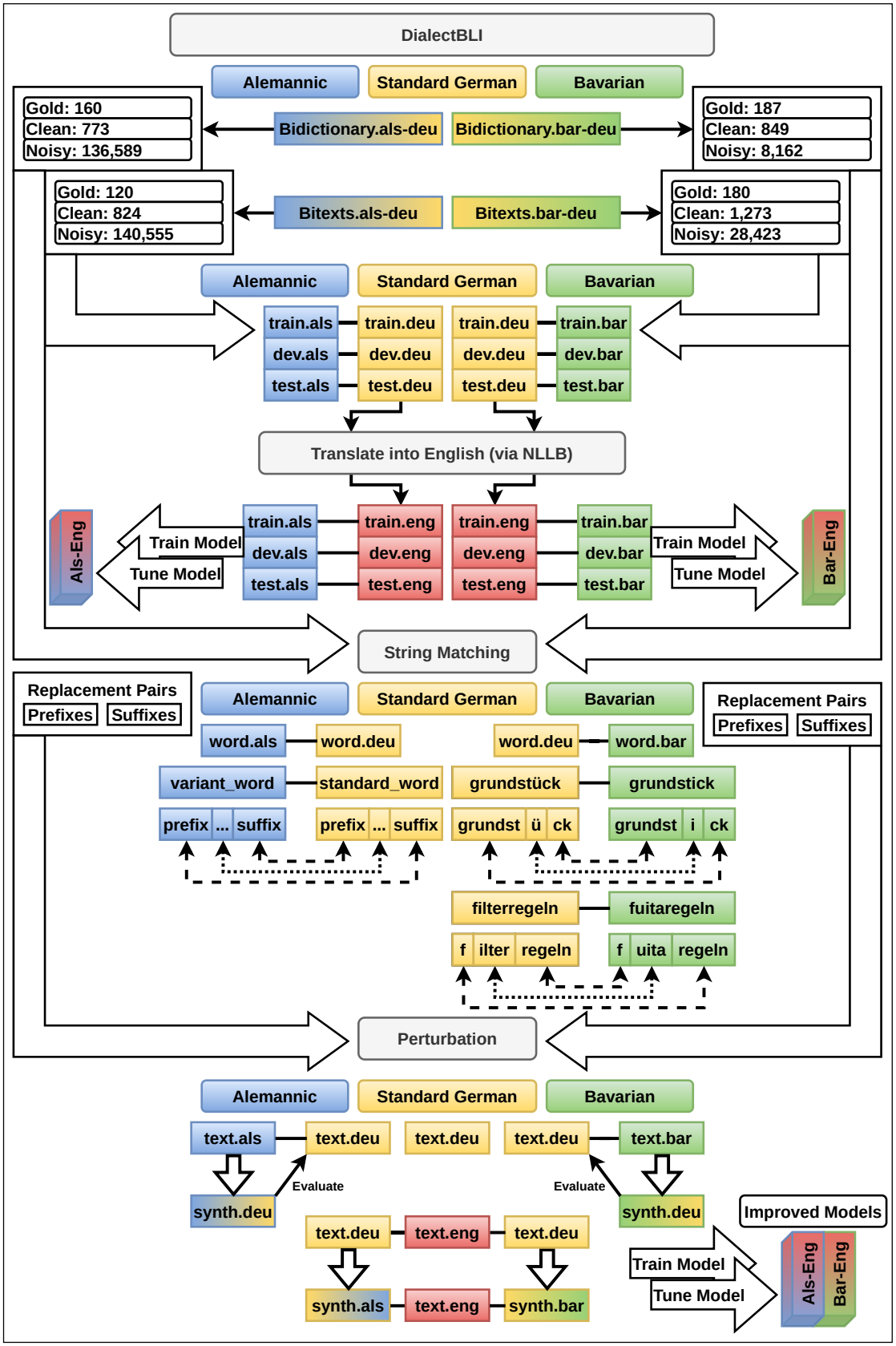


Figure 4.7: The path to improving MT via generating synthetic data based on aligned words. All measures denote the number of text lines.

There was no end to the words. No end to what they meant, or the ways they had been used. Some words' histories stretched so far back that our modern understanding of them was nothing more than an echo of the original, a distortion.

— Pip Williams

5 Results and Evaluation

This chapter presents the obtained results of the previously described experiments and their evaluation. Starting with insights into the produced perturbations, followed by an exploration of MT tools and ending with a metric-based evaluation of the reached translation performance.

5.1 Created Perturbations

For an impression of the created synthetic data, a selection of perturbations for the same pair of aligned sentences are shown with the standard variety German as starting point (see Table 5.1) and with the dialect variety Bavarian as starting point (see Table 5.2). Each row contains the result of text perturbations that are done via linguistic rules that are derived in different ways.

Rules	**Text lines**
None	*Die Sprache ist vom Aussterben bedroht, da sie nur von über 60-Jährigen gesprochen wird.*
Lex	*Die Lateinischn ist vom Aussterm bedroht, da sie nur von über 60-Jährigen gsprochn wird.*
All	*Die Lateinischn ist vom Aussterm bedroht, da sie nur von über 60-Jährigen gsprochn wird.*
Mor	*Dien Sproch ists vom Aussterbn bedrohts, da sien nua von üba 60-Jähring gsprochn wird.*
All∗	*Die Lateinischn ist vom Aussterm bedroht, da sie nur von über 60-Jährigen gsprochn wird.*
Mor∗	*Dien Sprouch ischt vom Aassderbn popbedrohd, da'n sien nua vonisch ušpopba 60-Ja'ähring gsprouchn wirdo.*

Table 5.1: Selected results of various (Dialectalization) perturbation methods of German (first row) text aiming to produce Bavarian looking text. English translation provided in Table 5.5

During dialectalization (see Table 5.1) of the sentence, the lexicographically-informed replacement rules replace the word *Sprache* (eng.: language) with *lateinischn* (eng.:

Latin), a repeatedly observed phenomenon, indicating that lexical replacement rules require a similar context window as the rules build on morphological information. On the other hand, rules purely informed on sub-word information, turn *Sprache* into *Sproch*, which correspond to the Bavarian word for language (sometimes *Sprooch*).

Rules	**Text lines**
None	*De Sproch is vom Aasterm bedroht wei se nua vo Leid de iwa 60 Joar oid san gredd wiad.*
Lex	*Drei Sprach fiel vom Aasterm bedroht wei se nua seit Leid drei iwa 60 Jahr oid bilden gredd wiad.*
All	*Drei Sprach bedeutet vom Aasterm bedroht wei se nua seit Leid drei iwa 60 Jahres oid saint gredd wiad.*
Mor	*Den Sprache ises voms Aasterms bedrohts weich sen nur vos Leit den iwas 60 Joars oit sans gredds wiert.*
All∗	*Drei Sprach kommt vom Aasterm bedroht wei se nua seit Leid drei iwa 60 Jahre oid saint gredd wiad.*
Mor∗	*Den Sprachs ies voms Austerms gelbedarohts weuch wassersen nur vos Leuten den iwas 60 Jahr oit wassersern geredet wiert.*

Table 5.2: Selected results of various (Standardization) perturbation methods of Bavarian (first row) text aiming to produce German looking text. English translation provided in Table 5.5

The perturbations marked with ∗ stem from a more generous selection of replacement rules in an attempt to capture more features than before. One such setting results in the replacement of *ist* (eng.: is) with *ischt*, but simultaneously impair multiple other words such as *Die* (eng.: the) → *Dien* and *über* (eng.: over) → *üspopba*.

In the reversed direction attempting standardization of Bavarian text into German text, the dubbed **Mor**-rules (based on sub-word unit information) show again great potential. Only the **Mor**-perturbation manages to transform the Bavarian word *Sproch* (eng.: language) into the corresponding German word *Sprache*. And it is the Mor∗-perturbation that creates the German *Leuten* (eng.: people) and *geredet* (eng.: spoken) from the Bavarian *Leid* and *gredd*. Each of these accomplishments is accompanied by sometimes severe degradation of other words, such as *bedroht* → *gelbedarohts*.

This reveals some of the difficulties of fine-tuning this method. Is the heuristic selection too strict, then the text does barely change, is it too lax, can the perturbations quickly degrade the entire text and turn it useless for any following MT task.

5.2 Machine Translation Baseline for German-English

To gauge the general performance across MT systems for translating between English and German in both directions (see Tables 5.4 and 5.3) the initial performance comparison includes the tools Argos[1], NLLB[2] and Googgle Translate[3].

The evaluation of the results in this thesis focuses exclusively on NLLB, which performed better than Argos in either direction and provides the ability to run differently sized models locally as well as to fine tune them, which is of crucial interest for potential future work building on this thesis.

[1] Argos Translate: https://github.com/argosopentech/argos-translate
[2] No Language Left Behind: https://github.com/facebookresearch/fairseq/tree/nllb
[3] Translate Shell (Google Translate): https://github.com/soimort/translate-shell

Model	Source	Target	BLEU	chrF2	TER
Argos	Deu	Eng	36.30	63.56	48.30
NLLB	Deu	Eng	42.20	66.72	43.14
Google	Deu	Eng	**48.18**	**71.21**	**37.70**

Table 5.3: Machine translation from German to English using the NLLB Seed Devtest.

Model	Source	Target	BLEU	chrF2	TER
Argos	Eng	Deu	31.75	60.49	53.76
NLLB	Eng	Deu	34.24	61.58	52.00
Google	Eng	Deu	**43.39**	**68.87**	**43.70**

Table 5.4: Machine translation from English to German using data from the NLLB Seed Devtest.

The scores shown in Table 5.4 are surprisingly low. Manually review of the translated texts indicate perfectly fine translations that simply do not use the exact same words that were found in the references of the test data. This, yet again, underlines how failable these evaluation metrics still are, even for a language pair that is considered to be as high-resource as English-German.

5.3 Perturbation-Based Machine Translation

Table 5.5 shows the resulting translations based on previously discussed perturbed text data from Tables 5.1 and 5.2, whereby † marks the perturbed texts.

Data	Text lines
Deu	*Die Sprache ist vom Aussterben bedroht, da sie nur von über 60-Jährigen gesprochen wird.*
Deu → Eng	The language is endangered, as it is only spoken by people over 60.
Deu†	*Dien Sprouch ischt vom Aassderbn popbedrohd, da'n sien nua vonisch ušpopba 60-Ja'ühring gsprouchn wirdo.*
Deu† → Eng	He 's threatened by the Aassderbn Pop because he 's a new member of the 60-year-old pop group.
Bar	*Den sprachs ies voms austerms gelbedarohts weuch wassersen nur vos Leuten den iwas 60 jahr oit wassersern geredet wiert.*
Bar → Eng	The only people who could speak the language of the world were those who had spoken it sixty years ago.
Bar†	*De Sproch is vom Aasterm bedroht wei se nua vo Leid de iwa 60 Joar oid san gredd wiad.*
Bar† → Eng	De Sproch is threatened by Aasterm we se nua vo Leid de iwa 60 years oid san gredd wiad.

Table 5.5: Selected English translations for some of the perturbations of the German-Bavarian aligned text from Tables 5.1 and 5.2.

Table 5.6 shows the resulting evaluation metrics of comparing originally German text that is dialectized to approximate the Bavarian version of the same text. High values for BLEU or chrF2 and low values for TER indicate that the compared texts are

very similar. The applied cleaning operations lead to better scores, probably due to the exclusion of sentence pairs if their lengths differ too much from each other which can be caused by sub-sentences missing on one side of the alignment or words translating into multi-word-expressions. Lexicographic replacements slightly reduce these scores, while replaced sub-word units result in severe degradation of the BLEU scores. The observed impact on the more precise chrF2 score is less extreme though.

Quality	Feature	Perturbation	BLEU	chrF2	TER
naive	none	none	16.745	28.9858	97.9605
clean	none	none	**22.0817**	**46.0659**	**68.867**
clean	guess	mor	0.0308	9.2788	119.8652
clean	reason	lex	17.0723	45.3778	81.284
clean	reason	mor	10.19	41.1213	87.47
clean	relaxed	mor	4.7539	32.6961	97.3855

Table 5.6: Evaluation for German perturbed to Bavarian (dialectization).

Table 5.7 shows the resulting evaluation metrics of comparing originally Bavarian text that is standardized to approximate the German version of the same text. This setting shows similar tendencies as before: Some degradation of the scores by replacing entire words, which is more pronounced for sub-word replacements.

Quality	Feature	Perturbation	BLEU	chrF2	TER
clean	none	none	**22.0817**	**46.0659**	**68.867**
clean	guess	mor	0.0089	14.0903	118.0105
clean	reason	lex	14.1974	42.1197	83.7689
clean	reason	all	14.2245	42.101	75.1912
clean	reason	mor	3.7269	38.1686	98.0128
clean	relaxed	mor	3.0415	31.1524	99.9885

Table 5.7: Evaluation for Bavarian perturbed to German (standardization).

Table 5.8 shows the resulting evaluation metrics of the English translation of originally Bavarian text that is standardized prior to translation compared with the English translation of the originally German text that in turn is dialectized. Here, the cleaning of the data has again a positive effect on the metrics, even more so than in the setting from above, which is only concerned with Bavarian-German.

Quality	Feature	Perturbation	BLEU	chrF2	TER
naive	none	none	4.4749	14.3248	119.8208
clean	none	none	**26.071**	**44.3625**	**74.0701**
clean	guess	mor	0.3164	17.0362	254.8441
clean	reason	lex	19.0943	41.2929	83.2599
clean	reason	mor	14.299	38.3677	87.4861
clean	relaxed	mor	6.0176	28.5109	109.5909

Table 5.8: Evaluation for Bavarian text, first perturbed to German and then translated into English compared with the English translations of the German text.

Our thinking was limited by convention (the most subtle but oppressive dictator).
Please forgive our lack of imagination.

— Pip Williams

6 Discussion and Conclusion

This chapter discusses the results in light of the limitations of this thesis. This is accompanied by explanations of why this line of research still holds a lot of potential before discussing implications and providing an outlook for future work.

6.1 Interpretation of the Results

6.1.1 Quantitative Analysis

Simply cleaning the data with automatic functions (admittedly geared towards processing German text), considerably improves the performance on all evaluation metrics. This effect is even more pronounced in the context of translating into English. This suggests that current practices in NLP, especially during data cleaning and corpus creation, are not strict enough, which can be attributed to the fear of excluding too much of the little data that is often available.

The inclusion of replacement rules based on **guess** is shown to absolutely demolish the performance in any of the experiment settings. The manual review of the perturbed texts confirms that the replacement rules without consideration of the context (applying a context window of length 0) leads to nonsensical creations in which some words have each and every of their letters replaced by sometimes entire morphemes.

In each setting, the application of lexicographic replacements (**lex**) results in a decrease of all scores. This can most likely be attributed to missing information about the context in which the to-be-replaced word is used. The data that serves as foundation for the replacement rules often includes many different options for aligned words such as the singular (*Vulkan* (eng.: volcano)) and the plural form (*Vulkane* (eng.: volcanos)) of a word or its noun (*Vulkan* (eng.: volcano)) and adjective (*vulkanig* (eng.: volcanic)) version. Section 5.1 already displayed how some of these replacements appear to be reasonable, while others tangle up the meaning of a sentence.

Interestingly, the perturbations on sub-word level (**mor**) show worse consequences regarding the BLEU score than the chrF2 score. Considering the same level of feature validity (**reason**) the perturbation types **lex** to **mor** are losing 7 BLEU and 4 chrF2 in Table 5.6, 10 BLEU and 4 chrF2 in Table 5.7, and 4 BLEU and 3 chrF2 in Table 5.8. On the one hand, this indicates that even though words get distorted, some character sequences still are strong and valid, arguing for the potential of fine-grained approaches.

On the other hand, is this degradation not desirable and possible causes need to be examined.

Finally, the approach of expanding the range of included features, and therefore rules as part of (**relaxed**), culminates in even worse results in all metrics. This, in combination with the observations from Table 5.2 (in which **Mor**$*$ denotes the **relaxed** setup of **mor**), where no other setup is able to replace the Bavarian *gredd* with the German *geredet* (eng.: spoken), paint a new picture. These contradictory observations serve as indication for the assumed explanation of noise due to mixed-in sub-dialectal data as cause for these unexpected inconsistencies.

6.1.2 Qualitative Analysis

Another reason is the ambiguous distinction between various dialects. Even authors aware of the mixed dialect situation decide to release datasets with very generalized or less-local dialect-tags. The few available and utilizable datasets restrict applied methods by the simplicity of used language label schemes. Her and Kruschwitz (2024) note: "Another challenge lies in multiple sub-dialects. This phenomenon can be observed in our corpus, which is mined from the Bavarian Wikipedia, where articles are written in different regional dialects. We argue that these sub-dialects in the parallel corpus lead to translation confusion, resulting in translation outputs which consist of mixed accents."

Literature (Burghardt, Granvogl, and Wolff, 2016; Artemova, Blaschke, and Plank, 2024) indicates that Bavarian can reasonably be divided into five dialect families, even though, there is no consistent naming in place yet (i.e. Northern Middle Bavarian vs. Central Northern Bavarian). Emphasize lies on the word families, as these grouped dialects can still drastically differ from each other (see Table 6.1). To make things even more difficult, there is the Bavarian-Alemannic transition zone (Lanwermeyer et al., 2016) to be considered, which expands the question of text language classification beyond the borders of Bavarian into neighboring dialects. To just name one example: the German word *bietet* (eng.: offers) results in four different, lexicographic replacements in Bavarian, namely *biat*, *buit*, *biatd*, and *bieat*.

6.2 Ongoing Challenges and Future Work

Having analyzed the quantitative and qualitative aspects of our results, we now turn our attention to the ongoing challenges in this field and the potential avenues for future research. While the current approach shows promise, several challenges remain to be addressed in future work. These challenges primarily revolve around improving data quality and refining the dialect identification process.

6.2.1 Dialectal Data Availability

The world wide web has been the primary source for data in NLP, being easily accessible and more affordable than acquiring help from experts. Regrettably, it displays a very uneven data distribution for the worlds languages and depending on the task domain, anything besides English, which makes up approximately 50% of the available data[2], can easily fall under the low-resourced category. This makes it challenging to explore new approaches for less-represented languages, as this work aims to do.

Initially, we utilize German-/English-aligned sentences from OpusTools (Aulamo et al., 2020) for Bavarian-German and Bavarian-English pairs. However, this data is notably noisy, and no equivalent data is available for Alemannic.

[2] https://w3techs.com/technologies/overview/content_language

Central Bavarian Western Variant	**Central Bavarian Eastern Variant**	**Standard German**	**English**
Isogloss: ui vs. üü			
v**ui**	v**üü**	viel	much
Schb**ui**, schb**ui**n	Schb**üü**, schb**üü**n	Spiel, spielen	Game, playing
i w**ui**	i w**üü**	ich will	I want
mia w**oi**n	mia w**öö**n/woin	wir wollen	we want
Isogloss: å vs. oa			
i f**å**(r), mia f**å**ma	i f**oa**, mia f**oa**n	ich fahre, wir fahren	I drive, we drive
h**å**(r)t, heata	h**oa**t, heata	hart, härter	hard, harder
Gf**å**, gf**â**li	Gf**oa**, gf**ea**li	Gefahr, gefährlich	danger, dangerous
Isogloss: oa vs. â			
oans, zw**oa**, gl**oa**	**â**ns, zw**â**, gl**â**	eins, zwei, klein	one, two, small
h**oa**ß, h**oa**zn	h**â**ß, h**â**zn	heiß, heizen	hot, heating
dah**oa**m, St**oa**	dah**â**m, St**â**	daheim, Stein	at home, stone
Isogloss: o vs. à			
i k**à**f	i k**ò**f	ich kaufe	I buy
mia k**à**ffa(n)	mia k**ò**ffa(n)	wir kaufen	we buy
Isogloss: no rule			
i k**i**mm	i k**u**mm	ich komme	I come
mia k**e**mma(n)	mia k**u**mma(n)	wir kommen	we come

Table 6.1: Examples of the differences found in Western and Eastern Variants of Central Bavarian[1].

6.2.2 Dialectal Data Processing

While data availability presents a significant hurdle, dialectal data processing introduces its own set of complexities that warrant careful consideration. In relation to the first stated research question *What is the performance of the current state-of-the-art models in translating dialects?* the findings of this thesis indicate a situation as abysmal as expected.

As no suitable framework for this type of experimental setup is currently available, most of the code and scripts for this thesis are written by the author and customized to the currently available data. Computer science methods are usually limited by the available data that can be processed. Ideally, each and any source of information would neatly funnel into the pipeline and contribute towards improving the results.

Not merely collected bitexts and seed dictionaries, but all kind of data: Starting with simple word lists, often found on Wikipedia pages of subdialects or in standardized form such as Swadesh lists (add reference). Linguistic rules attributed to certain dialect(s) as derived by experts and described in linguistic literature. Data from all layers of locality, be it the generalized term for an entire group of dialects i.e. Alemannic, the name of one of its dialects i.e. Low Alemannic, or even one of its many subdialects i.e. Upper Rhine Alemannic or Lake Constance Alemannic. This would have to happen most dynamically and at the same time static enough to still be able to run near-fully automatic.

6.2.3 Sub-Dialect Identification

One of the key challenges is the accurate identification and classification of sub-dialects within the broader dialect categories. Our initial attempts to automatically tag articles with sub-dialect labels faces significant hurdles:

We manually collect word lists for various sub-dialects from sources including Wikipedia, scientific literature, and other websites. However, these lists often contain words infrequently used in Wikipedia articles (e.g., personal pronouns like "I", "You", "She", "He"), resulting in very few articles being successfully labeled.

To address this, we sort articles from the Bavarian Wikipedia by their officially assigned dialect tags and create frequency dictionaries for each sub-dialect. Our proposed next steps include:

- Filtering these dictionaries (e.g., removing words with fewer than 5 occurrences)
- Using the filtered dictionaries to assign potential dialect-tag labels to each article
- Developing a heuristic to determine the most appropriate dialect-tag for each article

Developing such an effective heuristic for dialect identification presents several complexities. First, the question of frequency vs. uniqueness: A simplistic approach based solely on the number of words matching each dialect's dictionary may be misleading. For instance, an article with 15 words matching dialect A, 10 matching dialect B, and 5 matching dialect C might seem to belong to dialect A. However, if the 5 words from C are highly unique to that dialect and never used in others, while the 15 words from A are also common across dialects, the article might actually be more representative of dialect C. The same can be argued, it the 5 words from dialect C appear many times in this article, while the 15 words from A each only appear once, leading into the second complexity. Second, considerations regarding the word rarity: An ideal heuristic would consider not just the presence of words, but their rarity both within the article being tagged and within the dialect dictionaries. Third, the data imbalance: Some dialects may have many more articles and thus larger dictionaries, potentially skewing the identification process.

To address these challenges, we propose developing a more sophisticated dialect identification system that considers:

1. The number of words matching each dialect's dictionary
2. The uniqueness or exclusivity of matched words to specific dialects
3. The frequency of matched words within their respective dialects
4. The relative sizes of different dialect corpora
5. The distribution of dialect-specific words across all tagged articles

This system would aim to weight the importance of each word in the dialect identification process based on these factors, potentially providing a more accurate sub-dialect classification and hopefully better the utilization of even few data. Implementing and refining this system represents a significant avenue for future work, with the potential to greatly enhance the granularity and accuracy of our dialect transformation and machine translation processes.

6.2.4 Improving Replacement Rules

Having addressed the challenges of sub-dialect identification, we now turn our attention to refining the replacement rules, a crucial component in improving the overall performance of our approach. The second research question *Can we incorporate linguistic information in MT to synthetically generate sentences in language variants so that dialects of various languages can be processed more efficiently?* leads to a complex exploration of increasingly fine-grained dialectal variations.

To enhance the validity of linguistic features from which replacement rules are derived, an alternative to the current method of extraction from aligned words could be the direct use of rules described in scientific literature by expert linguists. While this approach may deviate from the language-agnostic ideal and limit the set of viable languages, it could serve as a valuable reference point for research on other language varieties. Researchers could potentially use this as a template to develop rules for new language varieties.

As discussed in Section 4.1.1, the data acquisition process could be enhanced by introducing automatic sentence-level alignment methods early in the processing. Although beyond the scope of this thesis, such an approach could build upon previous work on improving sentence alignment methods for low-resource language pairs (Tien et al., 2021). However, it is crucial to consider the known limitations of these methods, as outlined by (Forgac et al., 2023). These limitations include dependence on external tools, potential computational expense for larger datasets, and the need for specific resources such as dictionaries or pre-calculated vector embeddings.

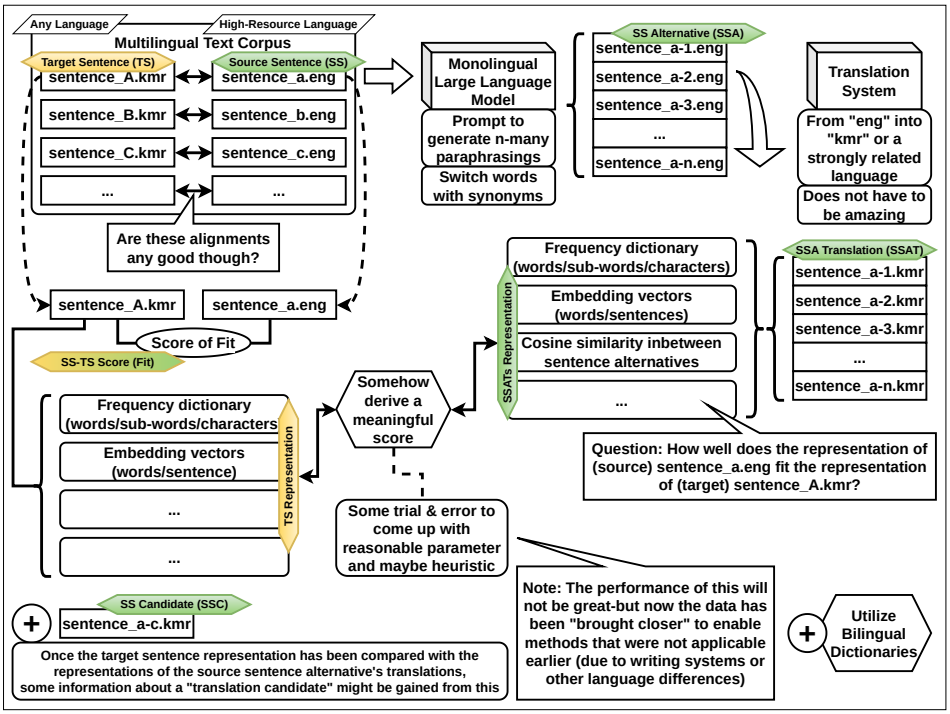


FIGURE 6.1: Idea for a method to improve text alignments for a language found in available but very noisy text corpora.

Given the data quality issues encountered during this work, developing new methods to improve the cleaning process of aligned data would be highly beneficial. Working with low-resource language varieties presents numerous limitations, with a major concern being the quality of data alignments in the few available datasets. These alignments are often created automatically and contain significant noise, which can negatively impact

the performance of downstream tasks and experiments. Figure 6.1 illustrates an experimental approach to removing noisy data (alignment pairs), which could prove valuable for future work in this area.

6.2.5 Enabling Evaluation on Dialect-Level

Finally, the third research question *What are requirements for deriving tools and processes that can be applied to vastly different languages from various language families?* demands more than an echoed call for more data. Simply adding larger amounts of data is an approach only suited for already well-established languages. Low-resource language varieties, which describes most dialects, require sophisticated strategies to be able to catch up. These are difficult to come by, if new attempts first require to set up an entirely new benchmark or evaluation schema, just because the targeted language varieties have not yet been covered by previous studies.

Benchmarks and test suites of very high quality, such as Macketanz et al., 2021 present for German, are direly needed to enable a comprehensive exploration of approaches for dialectal NMT. Attempts to expand the range of language varieties included in modern benchmarks (Alam, Ahmadi, and Anastasopoulos, 2024b; Aepli et al., 2023) aim into a recommendable direction, but are often still very limited in the number of varieties captured and fail to encompass many of the truly low-resourced languages to date.

6.2.6 Neural Machine Translation Enhancement

While establishing robust evaluation methods is critical, enhancing the neural machine translation models themselves presents another important avenue for improvement. One

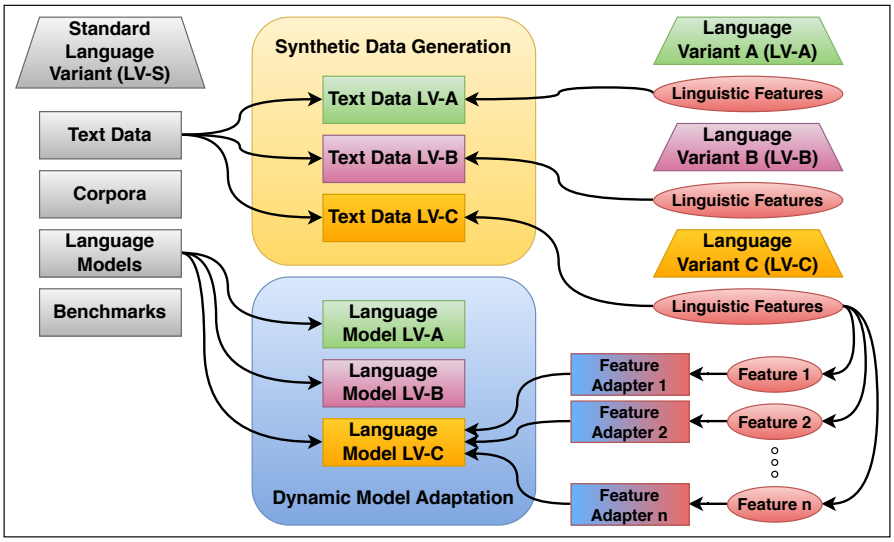


Figure 6.2: Utilization of the resources that a language standard variation brings with it in order to benefit the low-resource variants or dialects by generating synthetic text data (yellow) or deriving pretrained language models (blue) based on the variant's characteristic linguistic features (red).

approach to improve NMT models can be found in dynamic model adaptation based on the 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 6.2 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)). Regrettably, this step of the proposed methodology goes beyond the scope of this thesis and is left for future research.

6.3 Conclusion

As we have explored the various challenges and potential directions for future work, it is clear that this field of study offers significant opportunities for advancement. In light of these considerations, we can now draw some overarching conclusions from our research.

This experimental process underscores the significant challenges in conducting such research without access to expert linguists and native speakers. It highlights the need for more robust, language-agnostic methods and improved data quality for dialectal and low-resource language varieties. This need does not start at the moment of inserting aligned data into the training process for a MT model, but long before that and on a more foundational level. Producing and releasing properly labeled data, multi-varietal (capture a broader spectrum), and at the same time specific and standardized (reducing the ambiguousness), is indeed challenging, but can massively benefit future research over time.

Language proficiency is only one of the many puzzle pieces required for solving such problems. Ideally this line of work should be done by a team of computer scientists, expert linguists and decent pool of available native speakers, all working on eye-level. This might be a great way to merge the requirements and potential of computational systems with linguistic expertise meaningfully framed by the real needs of language communities. Despite all these hurdles and challenges, we still see great potential in advancing MT for dialects.

References

Abu Farha, Ibrahim and Walid Magdy (Dec. 2022). "The Effect of Arabic Dialect Familiarity on Data Annotation." In: *Proceedings of the Seventh Arabic Natural Language Processing Workshop (WANLP)*. WANLP 2022. Ed. by Houda Bouamor, Hend Al-Khalifa, Kareem Darwish, Owen Rambow, Fethi Bougares, Ahmed Abdelali, Nadi Tomeh, Salam Khalifa, and Wajdi Zaghouani. Abu Dhabi, United Arab Emirates (Hybrid): Association for Computational Linguistics, pp. 399–408. DOI: 10.18653/v1/2022.wanlp-1.39. URL: https://aclanthology.org/2022.wanlp-1.39 (visited on 07/11/2024).

Adebara, Ife and Muhammad Abdul-Mageed (May 2022). "Towards Afrocentric NLP for African Languages: Where We Are and Where We Can Go." In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. ACL 2022. Ed. by Smaranda Muresan, Preslav Nakov, and Aline Villavicencio. Dublin, Ireland: Association for Computational Linguistics, pp. 3814–3841. DOI: 10.18653/v1/2022.acl-long.265. URL: https://aclanthology.org/2022.acl-long.265 (visited on 07/08/2024).

Aepli, Noëmi, Chantal Amrhein, Florian Schottmann, and Rico Sennrich (Nov. 28, 2023). *A Benchmark for Evaluating Machine Translation Metrics on Dialects Without Standard Orthography*. DOI: 10.48550/arXiv.2311.16865. arXiv: 2311.16865 [cs]. URL: http://arxiv.org/abs/2311.16865 (visited on 07/08/2024). Pre-published.

Aepli, Noëmi and Rico Sennrich (May 2022). "Improving Zero-Shot Cross-lingual Transfer Between Closely Related Languages by Injecting Character-Level Noise." In: *Findings of the Association for Computational Linguistics: ACL 2022*. Findings 2022. Ed. by Smaranda Muresan, Preslav Nakov, and Aline Villavicencio. Dublin, Ireland: Association for Computational Linguistics, pp. 4074–4083. DOI: 10.18653/v1/2022.findings-acl.321. URL: https://aclanthology.org/2022.findings-acl.321 (visited on 07/08/2024).

Ahmadi, Sina and Antonios Anastasopoulos (May 25, 2023). *Script Normalization for Unconventional Writing of Under-Resourced Languages in Bilingual Communities*. DOI: 10.48550/arXiv.2305.16407. arXiv: 2305.16407 [cs]. URL: http://arxiv.org/abs/2305.16407 (visited on 06/14/2023). Pre-published.

Alam, Md Mahfuz Ibn, Sina Ahmadi, and Antonios Anastasopoulos (May 26, 2023). *CODET: A Benchmark for Contrastive Dialectal Evaluation of Machine Translation*. DOI: 10.48550/arXiv.2305.17267. arXiv: 2305.17267 [cs]. URL: http://arxiv.org/abs/2305.17267v1 (visited on 06/14/2023). Pre-published.

Alam, Md Mahfuz Ibn, Sina Ahmadi, and Antonios Anastasopoulos (Feb. 2, 2024a). *A Morphologically-Aware Dictionary-based Data Augmentation Technique for Machine Translation of Under-Represented Languages*. arXiv: 2402.01939 [cs]. URL: http://arxiv.org/abs/2402.01939 (visited on 05/08/2024). Pre-published.

Alam, Md Mahfuz Ibn, Sina Ahmadi, and Antonios Anastasopoulos (Feb. 2, 2024b). *CODET: A Benchmark for Contrastive Dialectal Evaluation of Machine Translation*. DOI: 10.48550/arXiv.2305.17267. arXiv: 2305.17267 [cs]. URL: http://arxiv.org/abs/2305.17267 (visited on 07/08/2024). Pre-published.

Ammon, Ulrich (July 11, 2011). *Die deutsche Sprache in Deutschland, Österreich und der Schweiz: Das Problem der nationalen Varietäten*. De Gruyter. ISBN: 978-3-11-087217-0. DOI: 10.1515/9783110872170. URL: https://www.degruyter.com/document/doi/10.1515/9783110872170/html (visited on 07/22/2024).

Ansell, Alan, Edoardo Maria Ponti, Anna Korhonen, and Ivan Vulić (Feb. 9, 2023a). *Composable Sparse Fine-Tuning for Cross-Lingual Transfer*. DOI: 10.48550/arXiv.2110.07560. arXiv: 2110.07560 [cs]. URL: http://arxiv.org/abs/2110.07560 (visited on 02/19/2023). Pre-published.

Ansell, Alan, Edoardo Maria Ponti, Anna Korhonen, and Ivan Vulić (June 2, 2023b). *Distilling Efficient Language-Specific Models for Cross-Lingual Transfer*. arXiv: 2306.01709 [cs]. URL: http://arxiv.org/abs/2306.01709 (visited on 06/29/2023). Pre-published.

Artemova, Ekaterina, Verena Blaschke, and Barbara Plank (Feb. 3, 2024). *Exploring the Robustness of Task-oriented Dialogue Systems for Colloquial German Varieties*. arXiv: 2402.02078 [cs]. URL: http://arxiv.org/abs/2402.02078 (visited on 05/08/2024). Pre-published.

Artemova, Ekaterina and Barbara Plank (May 2023). "Low-Resource Bilingual Dialect Lexicon Induction with Large Language Models." In: *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*. NoDaLiDa 2023. Ed. by Tanel Alumäe and Mark Fishel. Tórshavn, Faroe Islands: University of Tartu Library, pp. 371–385. arXiv: 2304.09957 [cs]. URL: https://aclanthology.org/2023.nodalida-1.39 (visited on 04/26/2024).

Artetxe, Mikel, Gorka Labaka, and Eneko Agirre (2018). "Unsupervised Statistical Machine Translation." In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Brussels, Belgium: Association for Computational Linguistics, pp. 3632–3642. DOI: 10.18653/v1/D18-1399. URL: http://aclweb.org/anthology/D18-1399 (visited on 12/25/2023).

Auer, Peter (June 24, 2011). "Europe's Sociolinguistic Unity, or: A Typology of European Dialect/Standard Constellations." In: *Perspectives on Variation*. De Gruyter Mouton, pp. 7–42. ISBN: 978-3-11-090957-9. DOI: 10.1515/9783110909579.7. URL: https://www.degruyter.com/document/doi/10.1515/9783110909579.7/html (visited on 07/22/2024).

Aulamo, Mikko, Umut Sulubacak, Sami Virpioja, and Jörg Tiedemann (May 2020). "OpusTools and Parallel Corpus Diagnostics." In: *Proceedings of the Twelfth Language Resources and Evaluation Conference*. LREC 2020. Ed. by Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis. Marseille, France: European Language Resources Association, pp. 3782–3789. ISBN: 979-10-95546-34-4. URL: https://aclanthology.org/2020.lrec-1.467 (visited on 05/08/2024).

Bafna, Niyati, Cristina España-Bonet, Josef van Genabith, Benoît Sagot, and Rachel Bawden (May 23, 2023). *A Simple Method for Unsupervised Bilingual Lexicon Induction for Data-Imbalanced, Closely Related Language Pairs*. DOI: 10.48550/arXiv.2305.14012. arXiv: 2305.14012 [cs]. URL: http://arxiv.org/abs/2305.14012 (visited on 03/13/2024). Pre-published.

Bali, Kalika, Monojit Choudhury, Sunayana Sitaram, and Vivek Seshadri (Dec. 1, 2019). "ELLORA: Enabling Low Resource Languages with Technology." In: UNESCO International Conference on Language Technologies for All (LT4All). URL: https://www.microsoft.com/en-us/research/publication/ellora-enabling-low-resource-languages-with-technology/ (visited on 07/25/2023).

Bandyopadhyay, Saptarashmi (2020). "Factored Neural Machine Translation on Low Resource Languages in the COVID-19 Crisis." In: URL: https://www.semanticscholar.org/paper/Factored-Neural-Machine-Translation-on-Low-Resource-Bandyopadhyay/8e3e4c9fc4ae3cdce1875800605c87d22a775962 (visited on 07/29/2024).

Bapna, Ankur, Isaac Caswell, Julia Kreutzer, Orhan Firat, Daan van Esch, Aditya Siddhant, Mengmeng Niu, Pallavi Baljekar, Xavier Garcia, Wolfgang Macherey, Theresa Breiner, Vera Axelrod, Jason Riesa, Yuan Cao, Mia Xu Chen, Klaus Macherey, Maxim Krikun, Pidong Wang, Alexander Gutkin, Apurva Shah, Yanping Huang, Zhifeng Chen, Yonghui Wu, and Macduff Hughes (July 6, 2022). *Building Machine Translation Systems for the Next Thousand Languages*. DOI: 10.48550/arXiv.2205.03983. arXiv: 2205.03983 [cs]. URL: http://arxiv.org/abs/2205.03983 (visited on 07/08/2024). Pre-published.

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)*. EMNLP-IJCNLP 2019. Hong Kong, China: Association for Computational Linguistics, pp. 1538–1548. DOI: 10.18653/v1/D19-1165. URL: https://aclanthology.org/D19-1165 (visited on 06/27/2023).

Baroni, Marco (Dec. 16, 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. URL: https://royalsocietypublishing.org/doi/full/10.1098/rstb.2019.0307 (visited on 07/04/2023).

Batsuren, Khuyagbaatar, Gábor Bella, and Fausto Giunchiglia (Mar. 1, 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. URL: https://doi.org/10.1007/s10579-021-09544-6 (visited on 02/14/2023).

Bergmann, Gunter (1990). "Upper Saxon." In: *The Dialects of Modern German*. Routledge. ISBN: 978-1-315-00177-7.

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

Bird, Steven (May 2022). "Local Languages, Third Spaces, and Other High-Resource Scenarios." In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. ACL 2022. Ed. by Smaranda Muresan, Preslav Nakov, and Aline Villavicencio. Dublin, Ireland: Association for Computational Linguistics, pp. 7817–7829. DOI: 10.18653/v1/2022.acl-long.539. URL: https://aclanthology.org/2022.acl-long.539 (visited on 07/08/2024).

Blaschke, Verena, Barbara Kovačić, Siyao Peng, Hinrich Schütze, and Barbara Plank (Mar. 15, 2024). *MaiBaam: A Multi-Dialectal Bavarian Universal Dependency Treebank*. DOI: 10.48550/arXiv.2403.10293. arXiv: 2403.10293 [cs]. URL: http://arxiv.org/abs/2403.10293 (visited on 04/05/2024). Pre-published.

Blaschke, Verena, Hinrich Schütze, and Barbara Plank (Apr. 20, 2023). *Does Manipulating Tokenization Aid Cross-Lingual Transfer? A Study on POS Tagging for Non-Standardized Languages*. arXiv: 2304.10158 [cs]. URL: http://arxiv.org/abs/2304.10158 (visited on 11/15/2023). Pre-published.

Bogoychev, Nikolay and Rico Sennrich (Oct. 3, 2020). *Domain, Translationese and Noise in Synthetic Data for Neural Machine Translation*. DOI: 10.48550/arXiv.1911.0

3362. arXiv: 1911.03362 [cs, stat]. URL: http://arxiv.org/abs/1911.03362 (visited on 06/27/2023). Pre-published.

Bugliarello, Emanuele, Fangyu Liu, Jonas Pfeiffer, Siva Reddy, Desmond Elliott, Edoardo Maria Ponti, and Ivan Vulić (July 17, 2022). *IGLUE: A Benchmark for Transfer Learning across Modalities, Tasks, and Languages*. DOI: 10.48550/arXiv.2201.11732. arXiv: 2201.11732 [cs]. URL: http://arxiv.org/abs/2201.11732 (visited on 07/08/2023). Pre-published.

Burghardt, Manuel, Daniel Granvogl, and Christian Wolff (May 2016). "Creating a Lexicon of Bavarian Dialect by Means of Facebook Language Data and Crowdsourcing." In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*. LREC 2016. Ed. by Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis. Portorož, Slovenia: European Language Resources Association (ELRA), pp. 2029–2033. URL: https://aclanthology.org/L16-1321 (visited on 04/05/2024).

Casas, Noe, Marta R. Costa-jussà, José A. R. Fonollosa, Juan A. Alonso, and Ramón Fanlo (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. URL: https://www.cambridge.org/core/journals/natural-language-engineering/article/abs/linguistic-knowledgebased-vocabularies-for-neural-machine-translation/C1FAB80C1D6ADCD252EB627BA3B4082B (visited on 07/09/2023).

Caseli, Helena M., Maria das Graças V. Nunes, and Mikel L. Forcada (Mar. 1, 2006). "Automatic Induction of Bilingual Resources from Aligned Parallel Corpora: Application to Shallow-Transfer Machine Translation." In: *Machine Translation* 20.4, pp. 227–245. ISSN: 1573-0573. DOI: 10.1007/s10590-007-9027-9. URL: https://doi.org/10.1007/s10590-007-9027-9 (visited on 07/19/2024).

Chambers, J. K. and Peter Trudgill (1998). *Dialectology*. 2nd ed. Cambridge Textbooks in Linguistics. Cambridge: Cambridge University Press. ISBN: 978-0-521-59646-6. DOI: 10.1017/CBO9780511805103. URL: https://www.cambridge.org/core/books/dialectology/3B5DB46311E1C43A8B15003717350F58 (visited on 07/19/2024).

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*. EACL 2021. Online: Association for Computational Linguistics, pp. 3440–3453. DOI: 10.18653/v1/2021.eacl-main.301. URL: https://aclanthology.org/2021.eacl-main.301 (visited on 06/27/2023).

Costa-jussà, Marta R., James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, Jeff Wang, and NLLB Team (June 2024). "Scaling Neural Machine Translation to 200 Languages." In: *Nature* 630.8018, pp. 841–846. ISSN: 1476-4687. DOI: 10.1038/s41586-024-07335-x. URL: https://www.nature.com/articles/s41586-024-07335-x (visited on 07/29/2024).

Costa-jussà, Marta R., Marcos Zampieri, and Santanu Pal (Aug. 2018). "A Neural Approach to Language Variety Translation." In: *Proceedings of the Fifth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2018)*. VarDial 2018.

Ed. by Marcos Zampieri, Preslav Nakov, Nikola Ljubešić, Jörg Tiedemann, Shervin Malmasi, and Ahmed Ali. Santa Fe, New Mexico, USA: Association for Computational Linguistics, pp. 275–282. URL: https://aclanthology.org/W18-3931 (visited on 07/08/2024).

Crystal, David (2000). *Language Death*. Cambridge: Cambridge University Press. DOI: 10.1017/CBO9781139106856. URL: https://www.cambridge.org/core/books/language-death/0661E7D2E2E77ACB08FDACB0E7BD4951 (visited on 08/12/2023).

Czarnowska, Paula, Sebastian Ruder, Edouard Grave, Ryan Cotterell, and Ann Copestake (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]. URL: http://arxiv.org/abs/1909.02855 (visited on 12/11/2023). Pre-published.

Darwish, Kareem, Nizar Habash, Mourad Abbas, Hend Al-Khalifa, Huseein T. Al-Natsheh, Houda Bouamor, Karim Bouzoubaa, Violetta Cavalli-Sforza, Samhaa R. El-Beltagy, Wassim El-Hajj, Mustafa Jarrar, and Hamdy Mubarak (Mar. 22, 2021). "A Panoramic Survey of Natural Language Processing in the Arab World." In: *Commun. ACM* 64.4, pp. 72–81. ISSN: 0001-0782. DOI: 10.1145/3447735. URL: https://doi.org/10.1145/3447735 (visited on 07/08/2024).

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*. LREC 2020. Marseille, France: European Language Resources Association, pp. 6300–6309. ISBN: 979-10-95546-34-4. URL: https://aclanthology.org/2020.lrec-1.773 (visited on 06/27/2023).

Demszky, Dorottya, Devyani Sharma, Jonathan Clark, Vinodkumar Prabhakaran, and Jacob Eisenstein (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*. 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. URL: https://aclanthology.org/2021.naacl-main.184 (visited on 07/07/2023).

Derince, Mehmet Şerif, Ergin Opengin, and Geoffrey Haig (2008). *Online Course in Kurmanji Kurdish, Uppsala University*.

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)*. EMNLP 2020. Online: Association for Computational Linguistics, pp. 5894–5904. DOI: 10.18653/v1/2020.emnlp-main.475. URL: https://aclanthology.org/2020.emnlp-main.475 (visited on 06/27/2023).

Doval, Yerai, Jesús Vilares, and Carlos Gómez-Rodríguez (Sept. 30, 2020). *Towards Robust Word Embeddings for Noisy Texts*. DOI: 10.48550/arXiv.1911.10876. arXiv: 1911.10876 [cs]. URL: http://arxiv.org/abs/1911.10876 (visited on 06/27/2023). Pre-published.

Dryer, Matthew S. and Martin Haspelmath, eds. (2013). *WALS Online (V2020.3)*. Zenodo. DOI: 10.5281/zenodo.7385533. URL: https://doi.org/10.5281/zenodo.7385533.

Duan, Sufeng, Hai Zhao, Dongdong Zhang, and Rui Wang (Apr. 29, 2020). *Syntax-Aware Data Augmentation for Neural Machine Translation*. DOI: 10.48550/arXiv.2004.14200. arXiv: 2004.14200 [cs]. URL: http://arxiv.org/abs/2004.14200 (visited on 06/27/2023). Pre-published.

Dupont, Quinn (Jan. 1, 2017). "The Cryptological Origins of Machine Translation, from al-Kindi to Weaver." In: *amodern*. URL: https://www.researchgate.net/publication/319529566_The_Cryptological_Origins_of_Machine_Translation_from_al-Kindi_to_Weaver.

Durrell, Martin (1990). "Westphalian and Eastphalian." In: *The Dialects of Modern German*. Routledge. ISBN: 978-1-315-00177-7.

Durrell, Martin and Winifred V. Davies (1990). "Hessian." In: *The Dialects of Modern German*. Routledge. ISBN: 978-1-315-00177-7.

Edunov, Sergey, Myle Ott, Michael Auli, and David Grangier (Oct. 2018). "Understanding Back-Translation at Scale." In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. EMNLP 2018. Brussels, Belgium: Association for Computational Linguistics, pp. 489–500. DOI: 10.18653/v1/D18-1045. URL: https://aclanthology.org/D18-1045 (visited on 06/27/2023).

Fancellu, Federico, Andy Way, and Morgan O'Brien (June 16, 2014). "Standard Language Variety Conversion for Content Localisation via SMT." In: *Proceedings of the 17th Annual Conference of the European Association for Machine Translation*. EAMT 2014. Ed. by Mauro Cettolo, Marcello Federico, Lucia Specia, and Andy Way. Dubrovnik, Croatia: European Association for Machine Translation, pp. 143–149. URL: https://aclanthology.org/2014.eamt-1.34 (visited on 07/08/2024).

Fishman, Joshua A. (Jan. 5, 2001). "Can Threatened Languages Be Saved?" In: *Can Threatened Languages Be Saved?* Multilingual Matters. ISBN: 978-1-85359-706-0. DOI: 10.21832/9781853597060. URL: https://www.degruyter.com/document/doi/10.21832/9781853597060/html?lang=en (visited on 07/29/2023).

Forgac, Frantisek, Dasa Munkova, Michal Munk, and Livia Kelebercova (Nov. 17, 2023). "Evaluating Automatic Sentence Alignment Approaches on English-Slovak Sentences." In: *Scientific Reports* 13.1, p. 20123. ISSN: 2045-2322. DOI: 10.1038/s41598-023-47479-w. URL: https://www.nature.com/articles/s41598-023-47479-w (visited on 07/22/2024).

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*. The Fourth Workshop. Boulder, Colorado: Association for Computational Linguistics, pp. 82–90. ISBN: 978-1-932432-37-4. DOI: 10.3115/1609843.1609855. URL: http://portal.acm.org/citation.cfm?doid=1609843.1609855 (visited on 06/27/2023).

Fung, Pascale and Benfeng Chen (2004). "BiFrameNet: Bilingual Frame Semantics Resource Construction by Cross-Lingual Induction." In: *Proceedings of the 20th International Conference on Computational Linguistics - COLING '04*. The 20th International Conference. Geneva, Switzerland: Association for Computational Linguistics, 931–es. DOI: 10.3115/1220355.1220489. URL: http://portal.acm.org/citation.cfm?doid=1220355.1220489 (visited on 07/19/2024).

Gao, Fei, Jinhua Zhu, Lijun Wu, Yingce Xia, Tao Qin, Xueqi Cheng, Wengang Zhou, and Tie-Yan Liu (July 2019). "Soft Contextual Data Augmentation for Neural Machine Translation." In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. ACL 2019. Florence, Italy: Association for Computational Linguistics, pp. 5539–5544. DOI: 10.18653/v1/P19-1555. URL: https://aclanthology.org/P19-1555 (visited on 06/27/2023).

Garcia, Xavier and Orhan Firat (Feb. 23, 2022). *Using Natural Language Prompts for Machine Translation*. DOI: 10.48550/arXiv.2202.11822. arXiv: 2202.11822 [cs]. URL: http://arxiv.org/abs/2202.11822 (visited on 07/08/2024). Pre-published.

Garcia, Xavier, Aditya Siddhant, Orhan Firat, and Ankur Parikh (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. URL: https://aclanthology.org/2021.naacl-main.89 (visited on 07/29/2023).

Goltz, Reinhard H. and Alastair G. H. Walker (1990). "North Saxon." In: *The Dialects of Modern German*. Routledge. ISBN: 978-1-315-00177-7.

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]. URL: http://arxiv.org/abs/1410.2455 (visited on 12/13/2023). Pre-published.

Green, Lisa J. (2002). *African American English: A Linguistic Introduction*. Cambridge: Cambridge University Press. ISBN: 978-0-521-89138-7. DOI: 10.1017/CBO9780511800306. URL: https://www.cambridge.org/core/books/african-american-english/1AE59657F9CF1BBC3A2BF2B9BB29D1D0 (visited on 07/19/2024).

Green, W. A. I. (1990). "The Dialects of the Palatinate (Das Pfälzische)." In: *The Dialects of Modern German*. Routledge. ISBN: 978-1-315-00177-7.

Ha, Thanh-Le, Jan Niehues, and Alexander Waibel (Nov. 15, 2016). *Toward Multilingual Neural Machine Translation with Universal Encoder and Decoder*. DOI: 10.48550/arXiv.1611.04798. arXiv: 1611.04798 [cs]. URL: http://arxiv.org/abs/1611.04798 (visited on 06/27/2023). Pre-published.

Haddow, Barry, Adolfo Hernández, Friedrich Neubarth, and Harald Trost (Sept. 2013). "Corpus Development for Machine Translation between Standard and Dialectal Varieties." In: *Proceedings of the Workshop on Adaptation of Language Resources and Tools for Closely Related Languages and Language Variants*. Ed. by Cristina Vertan, Milena Slavcheva, and Petya Osenova. Hissar, Bulgaria: INCOMA Ltd. Shoumen, BULGARIA, pp. 7–14. URL: https://aclanthology.org/W13-5303 (visited on 07/08/2024).

Held, Will, Caleb Ziems, and Diyi Yang (May 26, 2023). *TADA: Task-Agnostic Dialect Adapters for English*. DOI: 10.48550/arXiv.2305.16651. arXiv: 2305.16651 [cs]. URL: http://arxiv.org/abs/2305.16651 (visited on 07/06/2023). Pre-published.

Her, Wan-Hua and Udo Kruschwitz (Apr. 12, 2024). *Investigating Neural Machine Translation for Low-Resource Languages: Using Bavarian as a Case Study*. DOI: 10.48550/arXiv.2404.08259. arXiv: 2404.08259 [cs]. URL: http://arxiv.org/abs/2404.08259 (visited on 05/08/2024). Pre-published.

Hieber, Felix, Michael Denkowski, Tobias Domhan, Barbara Darques Barros, Celina Dong Ye, Xing Niu, Cuong Hoang, Ke Tran, Benjamin Hsu, Maria Nadejde, Surafel Lakew, Prashant Mathur, Anna Currey, and Marcello Federico (Aug. 2, 2022). *Sockeye 3: Fast Neural Machine Translation with PyTorch*. DOI: 10.48550/arXiv.2207.05851. arXiv: 2207.05851 [cs]. URL: http://arxiv.org/abs/2207.05851 (visited on 07/26/2024). Pre-published.

Hieber, Felix, Tobias Domhan, Michael Denkowski, and David Vilar (2020). "SOCKEYE 2: A Toolkit for Neural Machine Translation." In: *Eamt 2020*. URL: https://www.amazon.science/publications/sockeye-2-a-toolkit-for-neural-machine-translation.

Hieber, Felix, Tobias Domhan, Michael Denkowski, David Vilar, Artem Sokolov, Ann Clifton, and Matt Post (June 1, 2018). *Sockeye: A Toolkit for Neural Machine Translation*. arXiv: 1712.05690 [cs, stat]. URL: http://arxiv.org/abs/1712.05690 (visited on 10/26/2023). Pre-published.

Hu, Junjie, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson (Sept. 4, 2020). *XTREME: A Massively Multilingual Multi-task Benchmark for Evaluating Cross-lingual Generalization*. DOI: 10.48550/arXiv.2003.11080. arXiv: 2003.11080 [cs]. URL: http://arxiv.org/abs/2003.11080 (visited on 04/29/2023). Pre-published.

Irvine, Ann and Chris Callison-Burch (2013). "Supervised Bilingual Lexicon Induction with Multiple Monolingual Signals." In: *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 518–523. URL: https://aclanthology.org/N13-1056.pdf.

Irvine, Ann and Chris Callison-Burch (June 2017). "A Comprehensive Analysis of Bilingual Lexicon Induction." In: *Computational Linguistics* 43.2, pp. 273–310. DOI: 10.1162/COLI_a_00284. URL: https://aclanthology.org/J17-2001 (visited on 12/25/2023).

Jacobs, Neil G. (2005). *Jacobs N.G. Yiddish. A Linguistic Introduction*. Cambridge University Press. ISBN: 0-521-77215-X. URL: https://www.academia.edu/30182787/Jacobs_N_G_Yiddish_A_Linguistic_Introduction (visited on 07/22/2024).

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. URL: http://link.springer.com/10.1007/s10590-017-9203-5 (visited on 04/25/2023).

Khalid, Hewa Salam (2015). "Kurdish Dialect Continuum, as a Standardization Solution." In: *International Journal of Kurdish Studies* 1.1. ISSN: 2149-2751. URL: https://www.academia.edu/28277681/Kurdish_Dialect_Continuum_as_a_Standardization_Solution (visited on 09/02/2023).

Khalid, Hewa Salam (Apr. 29, 2020). "Kurdish Language, Its Family and Dialects." In: *International Journal of Kurdiname*. URL: https://www.academia.edu/45288807/KURDISH_LANGUAGE_ITS_FAMILY_AND_DIALECTS.

Koehn, Philipp (2009). *Statistical Machine Translation*. Cambridge: Cambridge University Press. ISBN: 978-0-521-87415-1. DOI: 10.1017/CBO9780511815829. URL: https://www.cambridge.org/core/books/statistical-machine-translation/94EADF9F680558E13BE759997553CDE5 (visited on 07/25/2023).

Koehn, Philipp (2020). *Neural Machine Translation*. Cambridge: Cambridge University Press. ISBN: 978-1-108-49732-9. DOI: 10.1017/9781108608480. URL: https://www.cambridge.org/core/books/neural-machine-translation/7AAA628F88ADD64124EA008C425C0197 (visited on 07/07/2023).

Kornai, András (Oct. 22, 2013). "Digital Language Death." In: *PLOS ONE* 8.10, e77056. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0077056. URL: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0077056 (visited on 06/16/2023).

Kumar, Adarsh, Rajalakshmi Krishnamurthi, Surbhi Bhatia, Keshav Kaushik, Neelu Jyothi Ahuja, Anand Nayyar, and Mehedi Masud (2021). "Blended Learning Tools and Practices: A Comprehensive Analysis." In: *IEEE Access* 9, pp. 85151–85197. ISSN: 2169-3536. DOI: 10.1109/ACCESS.2021.3085844. URL: https://ieeexplore.ieee.org/abstract/document/9446138 (visited on 01/22/2024).

Lakew, Surafel Melaku, Aliia Erofeeva, and Marcello Federico (Oct. 2018). "Neural Machine Translation into Language Varieties." In: *Proceedings of the Third Conference on Machine Translation: Research Papers*. WMT 2018. Ed. by Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Christof Monz, Matteo Negri,

Aurélie Névéol, Mariana Neves, Matt Post, Lucia Specia, Marco Turchi, and Karin Verspoor. Brussels, Belgium: Association for Computational Linguistics, pp. 156–164. DOI: 10.18653/v1/W18-6316. URL: https://aclanthology.org/W18-6316 (visited on 07/08/2024).

Lambrecht, Louisa, Felix Schneider, and Alexander Waibel (June 2022). "Machine Translation from Standard German to Alemannic Dialects." In: *Proceedings of the 1st Annual Meeting of the ELRA/ISCA Special Interest Group on Under-Resourced Languages*. SIGUL 2022. Ed. by Maite Melero, Sakriani Sakti, and Claudia Soria. Marseille, France: European Language Resources Association, pp. 129–136. URL: https://aclanthology.org/2022.sigul-1.17 (visited on 04/18/2024).

Lameli, Alfred (Aug. 29, 2008). *Deutsche Sprachlandschaften*. URL: https://aktuell.nationalatlas.de/dialektraeume-9_08-2008-0-html/ (visited on 07/05/2023).

Lanwermeyer, Manuela, Karen Henrich, Marie J. Rocholl, Hanni T. Schnell, Alexander Werth, Joachim Herrgen, and Jürgen E. Schmidt (May 27, 2016). "Dialect Variation Influences the Phonological and Lexical-Semantic Word Processing in Sentences. Electrophysiological Evidence from a Cross-Dialectal Comprehension Study." In: *Frontiers in Psychology* 7. ISSN: 1664-1078. DOI: 10.3389/fpsyg.2016.00739. URL: https://www.frontiersin.org/journals/psychology/articles/10.3389/fpsyg.2016.00739/full (visited on 05/10/2024).

Lardilleux, Adrien, Julien Gosme, and Yves Lepage (May 2010). "Bilingual Lexicon Induction: Effortless Evaluation of Word Alignment Tools and Production of Resources for Improbable Language Pairs." In: *The Seventh Conference on International Language Resources and Evaluation (LREC'10)*. Ed. by Nicoletta Calzolari (Conference Chair), Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odjik, Stelios Piperidis, Mike Rosner, and Daniel Tapias. Valletta, Malta: European Language Resources Association (ELRA), pp. 252–256. URL: https://hal.science/hal-00488768 (visited on 07/19/2024).

Lauscher, Anne, Vinit Ravishankar, Ivan Vulić, and Goran Glavaš (May 1, 2020). *From Zero to Hero: On the Limitations of Zero-Shot Cross-Lingual Transfer with Multilingual Transformers*. arXiv: 2005.00633 [cs]. URL: http://arxiv.org/abs/2005.00633 (visited on 04/29/2023). Pre-published.

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*. WMT 2007. Prague, Czech Republic: Association for Computational Linguistics, pp. 228–231. URL: https://aclanthology.org/W07-0734 (visited on 07/08/2023).

Lewis, Melvyn and Gary Simons (Apr. 1, 2010). "Assessing Endangerment: Expanding Fishman's GIDS." In: *Revue Roumaine de Linguistique* 55. DOI: 10.1017/CBO9780511783364.003. URL: https://www.researchgate.net/publication/228384852_Assessing_endangerment_Expanding_Fishman%27s_GIDS.

Li, Bin, Yixuan Weng, Fei Xia, and Hanjun Deng (Mar. 1, 2024). "Towards Better Chinese-centric Neural Machine Translation for Low-Resource Languages." In: *Computer Speech & Language* 84, p. 101566. ISSN: 0885-2308. DOI: 10.1016/j.csl.2023.101566. URL: https://www.sciencedirect.com/science/article/pii/S0885230823000852 (visited on 03/14/2024).

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). Pre-published.

Littell, Patrick, Anna Kazantseva, Roland Kuhn, Aidan Pine, Antti Arppe, Christopher Cox, and Marie-Odile Junker (Aug. 2018). "Indigenous Language Technologies in

Canada: Assessment, Challenges, and Successes." In: *Proceedings of the 27th International Conference on Computational Linguistics*. COLING 2018. Ed. by Emily M. Bender, Leon Derczynski, and Pierre Isabelle. Santa Fe, New Mexico, USA: Association for Computational Linguistics, pp. 2620–2632. URL: https://aclanthology.org/C18-1222 (visited on 07/08/2024).

Liu, Yanchen, William Held, and Diyi Yang (May 22, 2023). *DADA: Dialect Adaptation via Dynamic Aggregation of Linguistic Rules*. DOI: 10.48550/arXiv.2305.13406. arXiv: 2305.13406 [cs]. URL: http://arxiv.org/abs/2305.13406 (visited on 07/04/2023). Pre-published.

Lusito, Stefano, Edoardo Ferrante, and Jean Maillard (June 15, 2022). *Text Normalization for Endangered Languages: The Case of Ligurian*. DOI: 10.48550/arXiv.2206.07861. arXiv: 2206.07861 [cs]. URL: http://arxiv.org/abs/2206.07861 (visited on 06/27/2023). Pre-published.

Macketanz, Vivien, Eleftherios Avramidis, Shushen Manakhimova, and Sebastian Möller (Nov. 2021). "Linguistic Evaluation for the 2021 State-of-the-art Machine Translation Systems for German to English and English to German." In: *Proceedings of the Sixth Conference on Machine Translation*. WMT 2021. Ed. by Loic Barrault, Ondrej Bojar, Fethi Bougares, Rajen Chatterjee, Marta R. Costa-jussa, Christian Federmann, Mark Fishel, Alexander Fraser, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Paco Guzman, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Tom Kocmi, Andre Martins, Makoto Morishita, and Christof Monz. Online: Association for Computational Linguistics, pp. 1059–1073. URL: https://aclanthology.org/2021.wmt-1.115 (visited on 06/05/2024).

Mager, Manuel, Ximena Gutierrez-Vasques, Gerardo Sierra, and Ivan Meza-Ruiz (Aug. 2018). "Challenges of Language Technologies for the Indigenous Languages of the Americas." In: *Proceedings of the 27th International Conference on Computational Linguistics*. COLING 2018. Ed. by Emily M. Bender, Leon Derczynski, and Pierre Isabelle. Santa Fe, New Mexico, USA: Association for Computational Linguistics, pp. 55–69. URL: https://aclanthology.org/C18-1006 (visited on 07/08/2024).

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*. WNUT 2018. Brussels, Belgium: Association for Computational Linguistics, pp. 54–63. DOI: 10.18653/v1/W18-6108. URL: https://aclanthology.org/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, John R. Rickford, Guy Bailey, and John Baugh. London: Routledge, pp. 11–36. ISBN: 978-0-415-11732-6 978-0-415-11733-3.

Matras, Yaron (Jan. 1, 2019). "Revisiting Kurdish Dialect Geography: Findings from the Manchester Database." 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.* URL: https://www.academia.edu/35487253/Revisiting_Kurdish_dialect_geography_findings_from_the_Manchester_Database_Introduction_Database_method_and_scope (visited on 06/07/2023).

Mattheier, Klaus J. and Peter Wiesinger (1994). *Dialektologie des Deutschen: Forschungsstand und Entwicklungstendenzen*. De Gruyter. ISBN: 978-3-11-095848-5. DOI: 10.1515/9783110958485. URL: https://www.degruyter.com/document/doi/10.1515/9783110958485/html?lang=de (visited on 07/22/2024).

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.) URL: https://unesdoc.unesco.org/ark:/48223/pf0000187026 (visited on 07/25/2023).

Myint Oo, Thazin, Ye Kyaw Thu, and Khin Mar Soe (June 2019). "Neural Machine Translation between Myanmar (Burmese) and Rakhine (Arakanese)." In: *Proceedings of the Sixth Workshop on NLP for Similar Languages, Varieties and Dialects*. VarDial 2019. Ed. by Marcos Zampieri, Preslav Nakov, Shervin Malmasi, Nikola Ljubešić, Jörg Tiedemann, and Ahmed Ali. Ann Arbor, Michigan: Association for Computational Linguistics, pp. 80–88. DOI: 10.18653/v1/W19-1408. URL: https://aclanthology.org/W19-1408 (visited on 07/08/2024).

Newton, G. (1990). "Central Franconian." In: *The Dialects of Modern German*. Routledge. ISBN: 978-1-315-00177-7.

Ngo, Thi-Vinh, Phuong-Thai Nguyen, Van Vinh Nguyen, Thanh-Le Ha, and Le-Minh Nguyen (Dec. 31, 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. URL: https://doi.org/10.1080/08839514.2022.2101755 (visited on 06/24/2023).

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu (July 6, 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. URL: https://dl.acm.org/doi/10.3115/1073083.1073135 (visited on 04/29/2023).

Parović, Marinela, Goran Glavaš, Ivan Vulić, and Anna Korhonen (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*. 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. URL: https://aclanthology.org/2022.naacl-main.130 (visited on 02/19/2023).

Pfeiffer, Jonas, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder (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)*. EMNLP 2020. Online: Association for Computational Linguistics, pp. 7654–7673. DOI: 10.18653/v1/2020.emnlp-main.617. URL: https://aclanthology.org/2020.emnlp-main.617 (visited on 07/04/2023).

Philipp, Marthe and Arlette Bothorel-Witz (1990). "Low Alemannic." In: *The Dialects of Modern German*. Routledge. ISBN: 978-1-315-00177-7.

Polenz, Peter von (June 1, 2011). *Deutsche Sprachgeschichte vom Spätmittelalter bis zur Gegenwart*. De Gruyter. ISBN: 978-3-11-082488-9. DOI: 10.1515/9783110824889. URL: https://www.degruyter.com/document/doi/10.1515/9783110824889/html (visited on 07/22/2024).

Popović, Maja (Sept. 2015). "chrF: Character n-Gram F-score for Automatic MT Evaluation." In: *Proceedings of the Tenth Workshop on Statistical Machine Translation*. WMT 2015. Lisbon, Portugal: Association for Computational Linguistics, pp. 392–395. DOI: 10.18653/v1/W15-3049. URL: https://aclanthology.org/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*. WMT 2018.

Brussels, Belgium: Association for Computational Linguistics, pp. 186–191. DOI: 10.18653/v1/W18-6319. URL: https://aclanthology.org/W18-6319 (visited on 06/27/2023).

Ramponi, Alan (2024). "Language Varieties of Italy: Technology Challenges and Opportunities." In: *Transactions of the Association for Computational Linguistics* 12, pp. 19–38. ISSN: 2307387X. DOI: 10.1162/tacl_a_00631. URL: https://www.proquest.com/docview/2923016916/abstract/B9E1BFFB88C44878PQ/1 (visited on 05/10/2024).

Rei, Ricardo, Craig Stewart, Ana C Farinha, and Alon Lavie (Nov. 2020). "COMET: A Neural Framework for MT Evaluation." In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. EMNLP 2020. Online: Association for Computational Linguistics, pp. 2685–2702. DOI: 10.18653/v1/2020.emnlp-main.213. URL: https://aclanthology.org/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)*. EMNLP 2020. Online: Association for Computational Linguistics, pp. 4512–4525. DOI: 10.18653/v1/2020.emnlp-main.365. URL: https://aclanthology.org/2020.emnlp-main.365 (visited on 07/10/2023).

Riley, Parker, Timothy Dozat, Jan A. Botha, Xavier Garcia, Dan Garrette, Jason Riesa, Orhan Firat, and Noah Constant (June 29, 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. URL: https://doi.org/10.1162/tacl_a_00568 (visited on 07/11/2023).

Roark, Brian, Lawrence Wolf-Sonkin, Christo Kirov, Sabrina J. Mielke, Cibu Johny, Isin Demirsahin, and Keith Hall (May 2020). "Processing South Asian Languages Written in the Latin Script: The Dakshina Dataset." In: *Proceedings of the Twelfth Language Resources and Evaluation Conference*. LREC 2020. Ed. by Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis. Marseille, France: European Language Resources Association, pp. 2413–2423. ISBN: 979-10-95546-34-4. URL: https://aclanthology.org/2020.lrec-1.294 (visited on 07/08/2024).

Rowley, Anthony R. (1990a). "East Franconian." In: *The Dialects of Modern German*. Routledge. ISBN: 978-1-315-00177-7.

Rowley, Anthony R. (1990b). "North Bavarian." In: *The Dialects of Modern German*. Routledge. ISBN: 978-1-315-00177-7.

Ruder, Sebastian, Noah Constant, Jan Botha, Aditya Siddhant, Orhan Firat, Jinlan Fu, Pengfei Liu, Junjie Hu, Dan Garrette, Graham Neubig, and Melvin Johnson (Nov. 2021). "XTREME-R: Towards More Challenging and Nuanced Multilingual Evaluation." In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. EMNLP 2021. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, pp. 10215–10245. DOI: 10.18653/v1/2021.emnlp-main.802. URL: https://aclanthology.org/2021.emnlp-main.802 (visited on 07/10/2023).

Russ, Charles V. J. (1990a). "High Alemannic." In: *The Dialects of Modern German*. Routledge. ISBN: 978-1-315-00177-7.

Russ, Charles V. J. (1990b). "Swabian." In: *The Dialects of Modern German*. Routledge. ISBN: 978-1-315-00177-7.

Sahlgren, M. and J. Karlgren (Sept. 2005). "Automatic Bilingual Lexicon Acquisition Using Random Indexing of Parallel Corpora." In: *Natural Language Engineering* 11.3, pp. 327–341. ISSN: 1469-8110, 1351-3249. DOI: 10.1017/S1351324905003876. URL: https://www.cambridge.org/core/journals/natural-language-engineering/article/abs/automatic-bilingual-lexicon-acquisition-using-random-indexing-of-parallel-corpora/AB1D596379C225CC3CE11046934B81C7 (visited on 07/19/2024).

Salzmann, Martin and Gerhard Schaden (Sept. 6, 2019). "The Syntax and Semantics of Past Participle Agreement in Alemannic." In: *Glossa: a journal of general linguistics* 4.1 (1). ISSN: 2397-1835. DOI: 10.5334/gjgl.756. URL: https://www.glossa-journal.org/article/id/5212/ (visited on 05/10/2024).

Scherrer, Yves (July 2011). "Syntactic Transformations for Swiss German Dialects." In: *Proceedings of the First Workshop on Algorithms and Resources for Modelling of Dialects and Language Varieties*. Ed. by Jeremy Jancsary, Friedrich Neubarth, and Harald Trost. Edinburgh, Scotland: Association for Computational Linguistics, pp. 30–38. URL: https://aclanthology.org/W11-2604 (visited on 07/08/2024).

Scherrer, Yves (2012). "Generating Swiss German Sentences from Standard German: A Multi-Dialectal Approach." Université de Genève. DOI: 10.13097/ARCHIVE-OUVERTE/UNIGE:26361. URL: https://archive-ouverte.unige.ch/unige:26361 (visited on 07/11/2024).

Scherrer, Yves and Bruno Cartoni (2012). "The Trilingual ALLEGRA Corpus: Presentation and Possible Use for Lexicon Induction." In: *LREC*, pp. 2890–2896. URL: http://www.lrec-conf.org/proceedings/lrec2012/pdf/685_Paper.pdf.

Scherrer, Yves and Benoît Sagot (Sept. 13, 2013). "Lexicon Induction and Part-of-Speech Tagging of Non-Resourced Languages without Any Bilingual Resources." In: RANLP Workshop on Adaptation of Language Resources and Tools for Closely Related Languages and Language Variants. URL: https://inria.hal.science/hal-00862693 (visited on 07/19/2024).

Schmidt, Jürgen E. (Aug. 22, 2008). "Die deutsche Standardsprache: eine Varietät – drei Oralisierungsnormen." In: *Standardvariation*. De Gruyter, pp. 278–305. ISBN: 978-3-11-019398-5. DOI: 10.1515/9783110193985.278. URL: https://www.degruyter.com/document/doi/10.1515/9783110193985.278/html (visited on 07/22/2024).

Schönfeld, Helmut (1990). "East Low German." In: *The Dialects of Modern German*. Routledge. ISBN: 978-1-315-00177-7.

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*. ACL 2020. Online: Association for Computational Linguistics, pp. 7881–7892. DOI: 10.18653/v1/2020.acl-main.704. URL: https://aclanthology.org/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)*. ACL 2016. Berlin, Germany: Association for Computational Linguistics, pp. 86–96. DOI: 10.18653/v1/P16-1009. URL: https://aclanthology.org/P16-1009 (visited on 06/27/2023).

Snover, Matthew, Bonnie Dorr, Rich Schwartz, Linnea Micciulla, and John Makhoul (Aug. 8, 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*. AMTA 2006. Cambridge, Massachusetts, USA: Association for Machine Translation in the Americas, pp. 223–231. URL: https://aclanthology.org/2006.amta-papers.25 (visited on 07/08/2023).

Solano, Rolando Coto, Sally Akevai Nicholas, and Samantha Wray (Dec. 2018). "Development of Natural Language Processing Tools for Cook Islands Māori." In: *Proceedings of the Australasian Language Technology Association Workshop 2018*. ALTA 2018. Ed. by Sunghwan Mac Kim and Xiuzhen (Jenny) Zhang. Dunedin, New Zealand, pp. 26–33. URL: https://aclanthology.org/U18-1003 (visited on 07/08/2024).

Spangenberg, Karl (1990). "Thuringian." In: *The Dialects of Modern German*. Routledge. ISBN: 978-1-315-00177-7.

Srivastava, Aarohi and David Chiang (May 2023). "Fine-Tuning BERT with Character-Level Noise for Zero-Shot Transfer to Dialects and Closely-Related Languages." In: *Tenth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2023)*. VarDial 2023. Ed. by Yves Scherrer, Tommi Jauhiainen, Nikola Ljubešić, Preslav Nakov, Jörg Tiedemann, and Marcos Zampieri. Dubrovnik, Croatia: Association for Computational Linguistics, pp. 152–162. DOI: 10.18653/v1/2023.vardial-1.16. URL: https://aclanthology.org/2023.vardial-1.16 (visited on 07/08/2024).

Stellmacher, Dieter (June 26, 2017). *Niederdeutsch: Formen und Forschungen*. De Gruyter. ISBN: 978-3-11-092068-0. DOI: 10.1515/9783110920680. URL: https://www.degruyter.com/document/doi/10.1515/9783110920680/html (visited on 07/22/2024).

Sun, Jiao, Thibault Sellam, Elizabeth Clark, Tu Vu, Timothy Dozat, Dan Garrette, Aditya Siddhant, Jacob Eisenstein, and Sebastian Gehrmann (July 2023). "Dialect-Robust Evaluation of Generated Text." In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. ACL 2023. Ed. by Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki. Toronto, Canada: Association for Computational Linguistics, pp. 6010–6028. DOI: 10.18653/v1/2023.acl-long.331. URL: https://aclanthology.org/2023.acl-long.331 (visited on 07/08/2024).

Sánchez-Cartagena, Víctor M., Miquel Esplà-Gomis, Juan Antonio Pérez-Ortiz, and Felipe Sánchez-Martínez (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*. EMNLP 2021. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, pp. 8502–8516. DOI: 10.18653/v1/2021.emnlp-main.669. URL: https://aclanthology.org/2021.emnlp-main.669 (visited on 06/27/2023).

Tavadze, Givi (2019). "Spreading of the Kurdish Language Dialects and Writing Systems Used in the Middle East." In: 13.1. URL: https://www.semanticscholar.org/paper/Spreading-of-the-Kurdish-Language-Dialects-and-Used-Bagrationi/01e4b41fe3a08dea68b2413914c4e46330d2c85e.

Tiedemann, Jörg (2009). "News from OPUS - a collection of multilingual parallel corpora with tools and interfaces." In: *Recent advances in natural language processing*. Ed. by N. Nicolov, K. Bontcheva, G. Angelova, and R. Mitkov. Vol. V, pp. 237–248.

Tien, Ha Nguyen, Dat Nguyen Huu, Huong Le Thanh, Vinh Nguyen Van, and Minh Nguyen Quang (Nov. 2021). "KC4Align: Improving Sentence Alignment Method for Low-resource Language Pairs." In: *Proceedings of the 35th Pacific Asia Conference on Language, Information and Computation*. PACLIC 2021. Ed. by Kaibao Hu, Jong-Bok Kim, Chengqing Zong, and Emmanuele Chersoni. Shanghai, China: Association for Computational Lingustics, pp. 354–363. URL: https://aclanthology.org/2021.paclic-1.38 (visited on 07/22/2024).

Tripathi, Sneha (2010). "Approaches to Machine Translation." In: URL: https://www.academia.edu/39920933/Approaches_to_machine_translation (visited on 07/25/2023).

Tsujii, Junichi (Dec. 23, 2021). "Natural Language Processing and Computational Linguistics." In: *Computational Linguistics* 47.4, pp. 707–727. ISSN: 0891-2017. DOI: 10.1162/coli_a_00420. URL: https://doi.org/10.1162/coli_a_00420 (visited on 08/12/2023).

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin (Dec. 5, 2017). *Attention Is All You Need*. DOI: 10.48550/arXiv.1706.03762. arXiv: 1706.03762 [cs]. URL: http://arxiv.org/abs/1706.03762 (visited on 04/29/2023). Pre-published.

Vries, Wietse de, Martijn Bartelds, Malvina Nissim, and Martijn Wieling (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]. URL: http://arxiv.org/abs/2105.02855 (visited on 07/15/2023).

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).

Vulić, Ivan and Marie-Francine Moens (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. URL: https://aclanthology.org/P15-2118 (visited on 12/13/2023).

Waldendorf, Jonas, Alexandra Birch, Barry Hadow, and Antonio Valerio Micele Barone (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).

Wan, Yu, Baosong Yang, Derek F. Wong, Lidia S. Chao, Haihua Du, and Ben C. H. Ao (Apr. 3, 2020). "Unsupervised Neural Dialect Translation with Commonality and Diversity Modeling." In: *Proceedings of the AAAI Conference on Artificial Intelligence* 34.05 (05), pp. 9130–9137. ISSN: 2374-3468. DOI: 10.1609/aaai.v34i05.6448. URL: https://ojs.aaai.org/index.php/AAAI/article/view/6448 (visited on 07/08/2024).

Wang, Alex, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman (May 1, 2019). *SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems*. Version 1. DOI: 10.48550/arXiv.1905.00537. arXiv: 1905.00537 [cs]. URL: http://arxiv.org/abs/1905.00537 (visited on 07/29/2024). Pre-published.

Wang, Alex, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman (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*. 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. URL: http://aclweb.org/anthology/W18-5446 (visited on 07/08/2023).

Wang, Xinpeng, Cheng Fan, and Maximilian Frantzen (2021). *Training Domain Specific Multilingually Aligned Word Embeddings*. Munich: Technical University of Munich. URL: https://xinpeng-wang.github.io/pdfs/nlp_final_report.pdf.

Wiesinger, Peter (1990). "The Central and Southern Bavarian Dialects in Bavaria and Austria." In: *The Dialects of Modern German*. Routledge. ISBN: 978-1-315-00177-7.

Xia, Mengzhou, Xiang Kong, Antonios Anastasopoulos, and Graham Neubig (July 2019). "Generalized Data Augmentation for Low-Resource Translation." In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. ACL 2019. Florence, Italy: Association for Computational Linguistics, pp. 5786–5796. DOI: 10.18653/v1/P19-1579. URL: https://aclanthology.org/P19-1579 (visited on 06/27/2023).

Xie, Ziang, Sida I. Wang, Jiwei Li, Daniel Lévy, Aiming Nie, Dan Jurafsky, and Andrew Y. Ng (Mar. 7, 2017). *Data Noising as Smoothing in Neural Network Language Models*. DOI: 10.48550/arXiv.1703.02573. arXiv: 1703.02573 [cs]. URL: http://arxiv.org/abs/1703.02573 (visited on 06/27/2023). Pre-published.

Yamamoto, Kaoru, Yuji Matsumoto, and Mihoko Kitamura (2001). "A Comparative Study on Translation Units for Bilingual Lexicon Extraction." In: *Proceedings of the Workshop on Data-driven Methods in Machine Translation* -. The Workshop. Vol. 14. Toulouse, France: Association for Computational Linguistics, pp. 1–8. DOI: 10.3115/1118037.1118049. URL: http://portal.acm.org/citation.cfm?doid=1118037.1118049 (visited on 07/19/2024).

Ye, Jiacheng, Jiahui Gao, Qintong Li, Hang Xu, Jiangtao Feng, Zhiyong Wu, Tao Yu, and Lingpeng Kong (Dec. 2022). "ZeroGen: Efficient Zero-shot Learning via Dataset Generation." In: *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. EMNLP 2022. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, pp. 11653–11669. URL: https://aclanthology.org/2022.emnlp-main.801 (visited on 07/10/2023).

Zhang, Zhirui, Shujie Liu, Mu Li, Ming Zhou, and Enhong Chen (Mar. 1, 2018). *Joint Training for Neural Machine Translation Models with Monolingual Data*. DOI: 10.48550/arXiv.1803.00353. arXiv: 1803.00353 [cs]. URL: http://arxiv.org/abs/1803.00353 (visited on 06/27/2023). Pre-published.

Ziems, Caleb, Jiaao Chen, Camille Harris, Jessica Anderson, and Diyi Yang (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)*. ACL 2022. Dublin, Ireland: Association for Computational Linguistics, pp. 3701–3720. DOI: 10.18653/v1/2022.acl-long.258. URL: https://aclanthology.org/2022.acl-long.258 (visited on 07/04/2023).

Ziems, Caleb, William Held, Jingfeng Yang, Jwala Dhamala, Rahul Gupta, and Diyi Yang (May 29, 2023). *Multi-VALUE: A Framework for Cross-Dialectal English NLP*. DOI: 10.48550/arXiv.2212.08011. arXiv: 2212.08011 [cs]. URL: http://arxiv.org/abs/2212.08011 (visited on 07/04/2023). Pre-published.

Üstün, Ahmet, Alexandre Bérard, Laurent Besacier, and Matthias Gallé (Oct. 20, 2021). *Multilingual Unsupervised Neural Machine Translation with Denoising Adapters*. arXiv: 2110.10472 [cs]. URL: http://arxiv.org/abs/2110.10472 (visited on 06/04/2023). Pre-published.

A Appendix: German Varieties

A.1 The German Language

German language is a West Germanic language with a significant presence across Western and Central Europe. Its status as an official or co-official language extends beyond Germany to several other countries and regions, including Austria, Switzerland, Liechtenstein, and the Italian province of South Tyrol. German also holds official status in Luxembourg and Belgium, and is recognized as a national language in Namibia. The reach of the German language extends further, with notable German-speaking communities found in parts of France (Alsace), Czech Republic (North Bohemia), Poland (Upper Silesia), Slovakia (Košice Region, Spiš, and Hauerland), and Hungary (Sopron)[1]. This wide geographic distribution, spanning multiple countries and cultures, contributes to the rich diversity of German language varieties.

A key feature of the German language landscape is the use of Standard German as a written standard across German-speaking countries. This often results in situations of diglossia, particularly pronounced in Switzerland (Ammon, 2011). While Standard German primarily draws from Middle and Upper German dialects, it has evolved over time, incorporating features from various dialectal sources (Polenz, 2011).

The performance of differently sized NLLB models translating between English and German (see Figure A.2).

A.2 The Dialects of German

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 A.3). These linguistic features can be shared by multiple dialects (see Figure A.4) and encountering one such feature in a text does not

[1] https://en.wikipedia.org/wiki/German_language

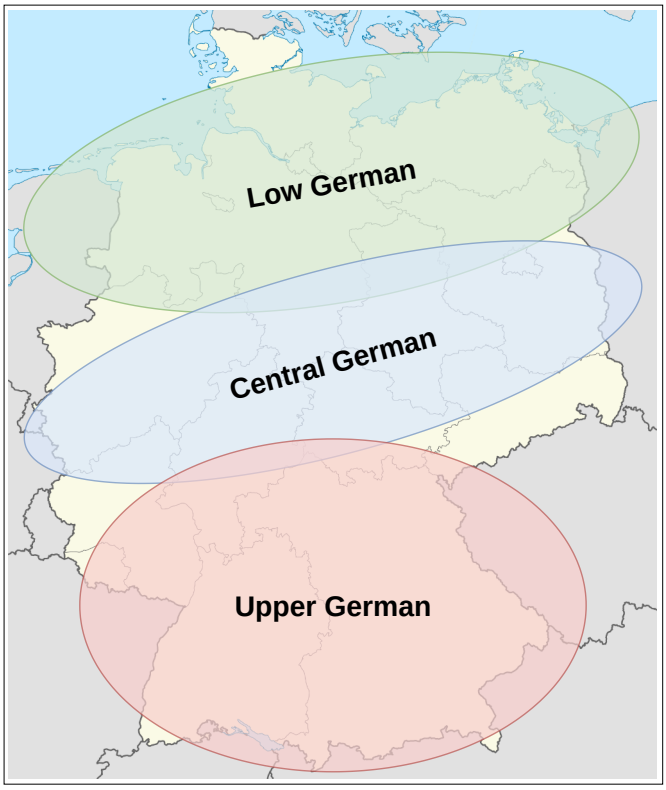


FIGURE A.1: German is usually divided into three divisions. These can roughly be placed as shown here, with Upper German in the south and Lower German in the north of Germany.

Reference Data	**English sentence from the NLLB dev data**	
	On Monday, scientists from the Stanford University School of Medicine announced the invention of a new diagnostic tool that can sort cells by type: a tiny printable chip that can be manufactured using standard inkjet printers for possibly about one U.S. cent each.	
	(Aligned) German sentence from the NLLB dev data	
	Am Montag haben die Wisenschaftler der Stanford University School of Medicine die Erfindung eines neuen Diagnosetools bekanntgegeben, mit dem Zellen nach ihrem Typ sortiert werden können: ein winziger, ausdruckbarer Chip, der für jeweils etwa einen US-Cent mit Standard-Tintenstrahldruckern hergestellt werden kann.	
Translated English to German	**dist-600M**	**German**
	Am Montag haben die Wissenschaftler der Stanford University School of Medicine die Erfindung eines neuen Diagnosetools bekanntgegeben, mit dem Zellen nach ihrem Typ sortiert werden können: ein winziger, ausdruckbarer Chip, der für Juwelen etwa ein US-Cent mit Standard-Tintenstrahldruckern hergestellt werden können.	
	dist-1.3B	**German**
	Am Montag gaben Wissenschaftler der Stanford University School of Medicine die Erfindung eines neuen Diagnosetools bekannt, mit dem die Zellen nach ihrem Typ sortiert werden können: ein winziger, ausdrucksfähigerer Chip, der für Juwelen wie einen US-Cent mit Standard-Tintenstrahldruckern hergestellt werden kann.	
	1.3B	**German**
	Am Montag haben die Wissenschaftler der Stanford University School of Medicine die Erfindung eines neuen Diagnosetools bekannt gegeben, mit dem Zellen nach ihrem Typ sortiert werden können: ein winziger, ausdruckbarer Chip, der für Juwelen etwa einen US-Cent mit Standard-Tintenstrahldruckern hergestellt werden kann.	
	3.3B	**German**
	Am Montag haben die Wissenschaftler der Stanford University School of Medicine die Erfindung eines neuen Diagnosetools bekanntgegeben, mit dem Zellen nach ihrem Typ sortiert werden können: ein winziger, ausdruckbarer Chip, der für jeweils etwa einen US-Cent mit Standard-Tintenstrahldruckern hergestellt werden kann.	

FIGURE A.2: Performance of differently sized NLLB models for English to German translation, with differences shown on a randomly selected example sentence from the NLLB-Devset.

necessarily guarantee the text to be of a specific dialect. However, these considerations are more of an idealistic nature, since the reality of data availability sets considerably harsher constraints than pure neglect could achieve.

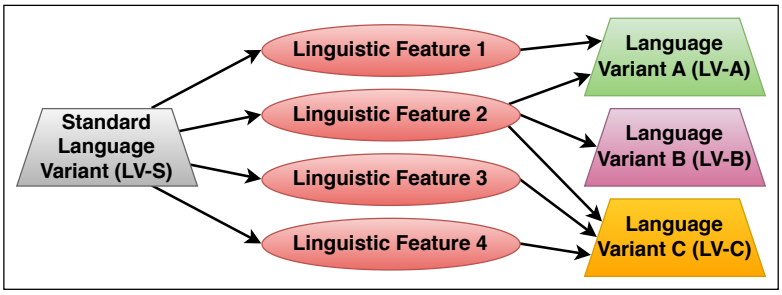


FIGURE A.3: The standard variant of a language, modified by certain linguistic features, takes on the form of a language variation- or dialect.

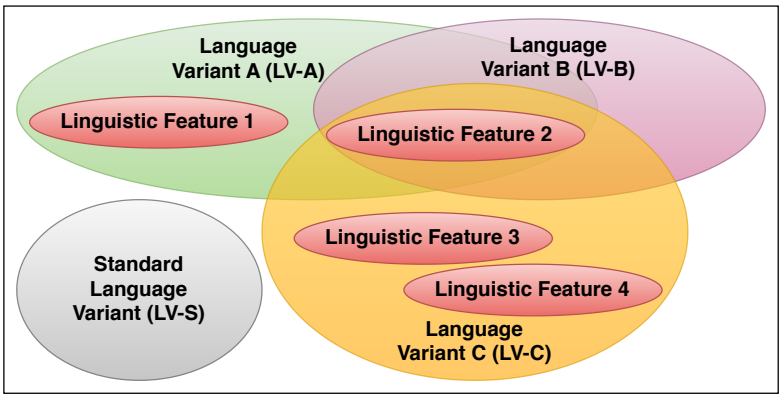


FIGURE A.4: Dialects are not monolithic i.e., being discretely distinguishable classes with clearly defined boarders, but can have fluent transitions and overlap between each other.

A.2.1 German Varieties

The German language (Mattheier and Wiesinger, 1994) can roughly be separated into Lower German (spoken in the north of Germany), Middle German (spoken in the center of Germany), and Upper German (spoken in the south of Germany) as shown in Figure A.1. It is important to note that the dialects discussed below form a continuum, and the boundaries between them are often fuzzy (Auer, 2011), which is why many transitional dialects exist that share features of multiple groups. The German language exhibits a complex dialectal landscape (see Figure A.5 for an aggregation of dialect locations), highlighting the challenges in processing diverse linguistic data. This diversity can be broadly categorized into several major groups, each with its own subdivisions:

- Low German (Stellmacher, 2017), which can be further divided into:
  - West Low German (Durrell, 1990) comprising of the dialects Westphalian, Eastphalian, North Low Saxon
  - East Low German (Schönfeld, 1990) comprising of the dialects Mecklenburgisch-Vorpommersch, Brandenburgisch, Mittelpommersch
- Middle German (Schmidt, 2008)
  - West Middle German comprising of the dialects Ripuarian, Moselle Franconian, Luxembourgish, Rhine Franconian (Green, 1990), Hessian (Durrell and Davies, 1990)

  - East Middle German comprising of the dialects Thuringian (Spangenberg, 1990), Upper Saxon (Bergmann, 1990), North Upper Saxon (Goltz and Walker, 1990), Lusatian dialects

- Franconian dialects which are often considered transitional between Upper German and Middle German:
  - East Franconian (Rowley, 1990a) spoken in parts of Bavaria and Baden-Württemberg
  - South Franconian: spoken in northern Baden-Württemberg
  - Central Franconian (Newton, 1990)

- West Upper German, or simply Alemannic
  - Low Alemannic (Philipp and Bothorel-Witz, 1990) comprising of the dialects Upper Rhine Alemannic and Alemán Coloniero (spoken in Venezuela)
  - Swabian
  - Lake Constance Alemannic, sometimes considered to be part of Low Alemannic, can also be seen as a transitional dialect, closer to High Alemannic with some Swabian mixed in.
  - High Alemannic (Russ, 1990a) comprising of the dialects Bernese German, Vorarlberg German, Zurich German, and Liechtensteinisch.
  - Highest Alemannic (Salzmann and Schaden, 2019) comprising of the dialects Walliser German and Walser German.

- East Upper German, or simply Bavarian
  - Northern Bavarian (Rowley, 1990b)
  - Northern Central Bavarian
  - Central Bavarian
  - Central Southern Bavarian (Wiesinger, 1990)
  - Southern Bavarian

- Swiss German, while part of the Alemannic group, deserves special mention due to its status in Switzerland. It's used as the spoken language in most situations in German-speaking Switzerland It has significant differences from Standard German in vocabulary, pronunciation, and grammar

- Yiddish (Jacobs, 2005) While not typically classified as a German dialect, Yiddish is a Germanic language that developed from Middle High German and is historically associated with Ashkenazi Jewish communities.

This extensive dialectal variation presents significant challenges for natural language processing tasks, including MT. The diversity in vocabulary, grammar, and pronunciation across these dialects underscores the complexity of developing robust language models capable of accurately processing and translating German in all its varieties. This linguistic landscape emphasizes the need for sophisticated approaches in CL to address the nuances of dialectal variations in German and other languages with similar diversity.

Along the previously discussed notion of dialects being part of a language, our approach requires language varieties that are related enough for detecting, codifying, and

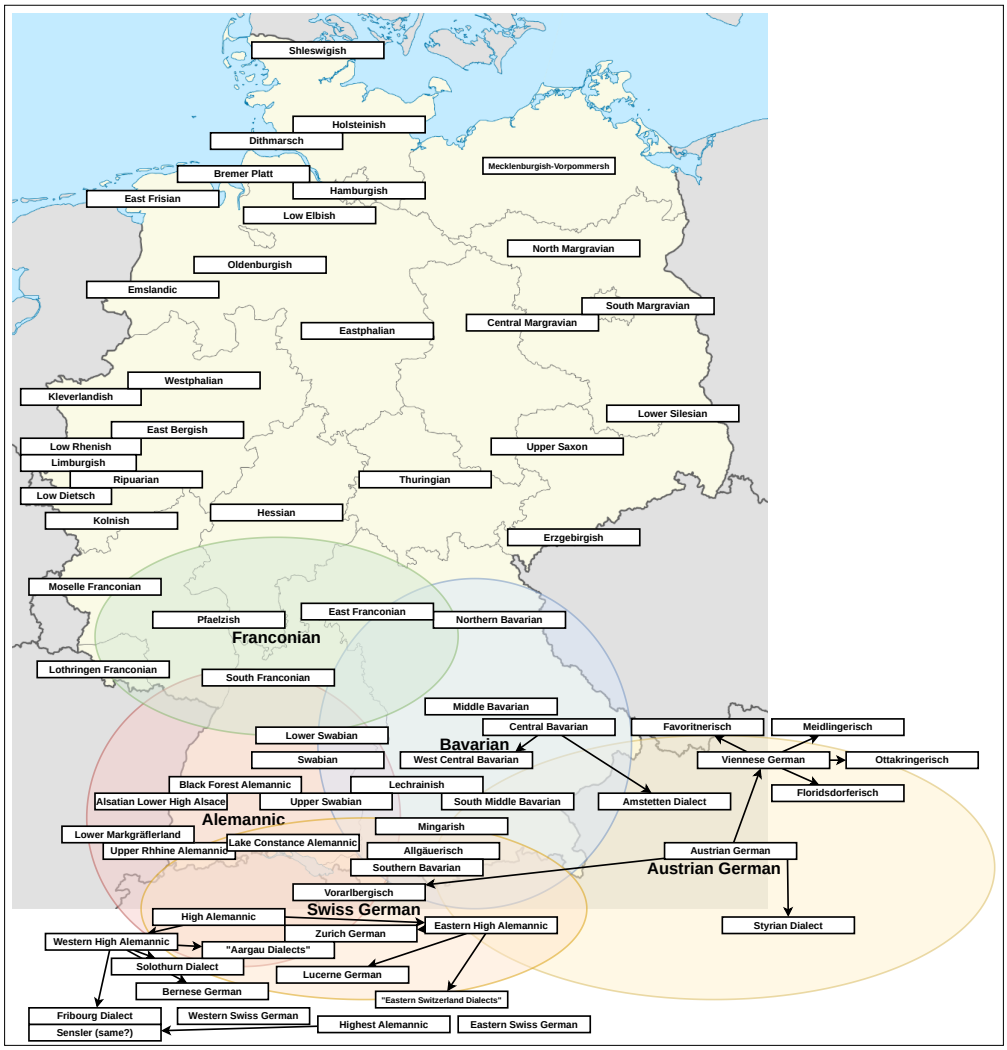


FIGURE A.5: Estimated and often rough location of German dialects across Germany and its surrounding, with a focus on the ones found in the south. These locations have been collected over a longer time period from various Wikipedia language pages in combination of educated guesses by the author of this thesis.

applying rules that transform text between them, but still so different, that this process does not become completely trivial. The following motivates Alemannic and Bavarian as sensible choices for dialects that, in combination with the language German might benefit from this line of work. For the Standard German language there exists a very large Wikipedia with almost 3,000,000 articles[2]. The Upper German dialect groups Alemannic and Bavarian each have a much smaller sized Wikipedia[3][4] with a bit over 30,000 and 27,000 articles respectively. Worth mentioning is the existence of a third German variety Wikipedia which is for Ripuarian with close to 3,000 articles[5] online.

[2] https://de.wikipedia.org/wiki/Deutschsprachige_Wikipedia
[3] https://als.wikipedia.org/wiki/Wikipedia:Houptsyte
[4] https://bar.wikipedia.org/wiki/Wikipedia:Hoamseitn
[5] https://ksh.wikipedia.org/wiki/Wikipedia:Houpsigk

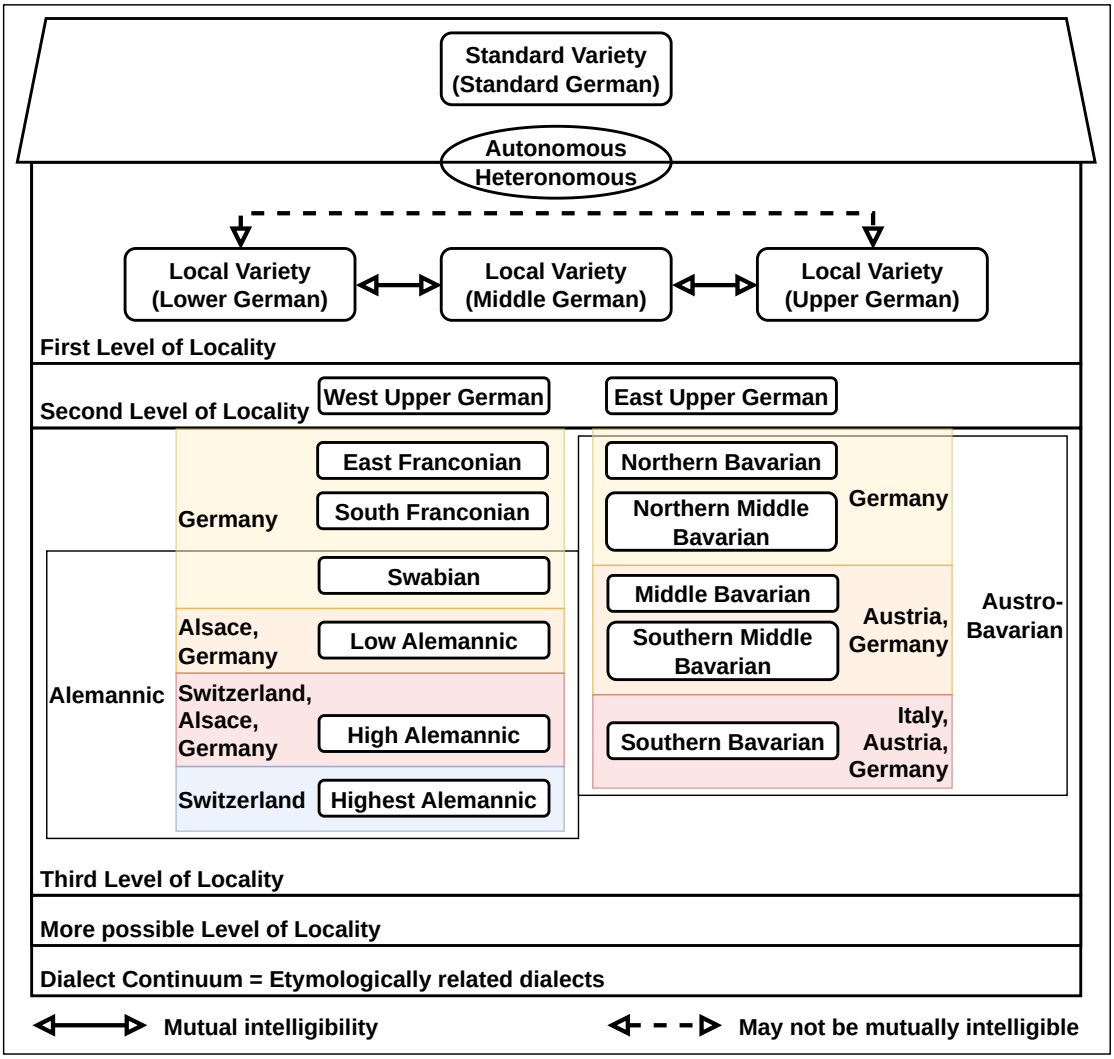


FIGURE A.6: House of dialects with additional floors for sub-dialects showing the two most dominant Upper German dialect groups. This figure was created based on the ideas explored in (Bird, 2022; Scherrer, 2012).

A.2.2 Alemannic Dialects

(Lambrecht, Schneider, and Waibel, 2022) explored MT from Standard German into Alemannic dialects based on cleaned and dialect-specific filtered text data from Wikipedia articles.

Some examples of Alemannic characteristics which clearly differentiate it from German and in part already hint at the nature of possible replacement rules:

- The diminutive is used frequently in all Alemannic dialects. Northern and eastern dialects use the suffix -le; western varieties (e.g. northern Alsace) uses the suffix -el /l/; southern dialects use the suffix -li (Standard German suffix -lein or -chen). As in standard German, these suffixes cause umlaut. Depending on dialect, 'little house' may be Heisle, Hiisel, Hüüsle, Hüüsli or Hiisli (Standard German Häuslein or Häuschen). Some varieties have plural diminutives in -ler, -la or -lich

- Northern variants of Alemannic (Swabian and Low Alemannic), like Standard German, pronounce ch as a uvular or velar [χ] or [x] (Ach-Laut) after back vowels (a,

o, u) and as a palatal [ç] consonant (Ich-Laut) elsewhere. High Alemannic, Lake Constance Alemannic and Highest Alemannic dialects exclusively use the Ach-Laut

- In most Alemannic dialects, the past participle of the verb meaning to be (sein in standard German, with past participle gewesen) derives from a form akin to gesein (gsi, gsìnn, gsei etc.)
- The use of 'sch' instead of 's' before 't', 'p', and 'w' in many words (e.g., 'Fescht' instead of 'Fest')
- The tendency to round front vowels, especially in Swabian (Russ, 1990b) (e.g., 'Füeß' instead of 'Füße')

A.2.3 Bavarian Dialects

As done for Alemannic, some examples of Bavarian characteristics, which differentiate it from German and give an impression of how perturbations might change text from one variety to another:

- Bavarian usually has case inflection only for the article. With very few exceptions, nouns are not inflected for case
- The simple past tense is very rare in Bavarian and has been retained for only a few verbs, including 'to be' and 'to want'. In general, the perfect is used to express past time
- Bavarian features verbal inflection for several moods such as indicative, subjunctive, imperative and optative
- The use of 'a' instead of 'ei' in many words (e.g., 'zwa' instead of 'zwei')
- The tendency to use 'oa' instead of 'ei' in some words (e.g., 'hoaß' instead of 'heiß')
- The use of the prefix 'da-' instead of 'er-' in many verbs

B

Appendix: Sub-Dialectal Wikidump Filtering

B.1 Wikidump Processing

Lambrecht, Schneider, and Waibel (2022) report a considerable performance improvement just by data being split into dialect groups. They mention spelling inconsistencies still being present, but now decreased in their number. An approach to cleaning and filtering the available wikidump data of a language (see Figure B.1) has the potential to improve results by reducing the noise found in the data due to mixing many sub-dialects. While reasonable, the idea of training classifier for each dialect (Lambrecht, Schneider, and Waibel, 2022) is beyond the scope of this thesis. To filter text according to sub-dialects, currently, the step labeled **Learn to Classify Sentences** is substituted by a schema directly based on the observer word frequencies.

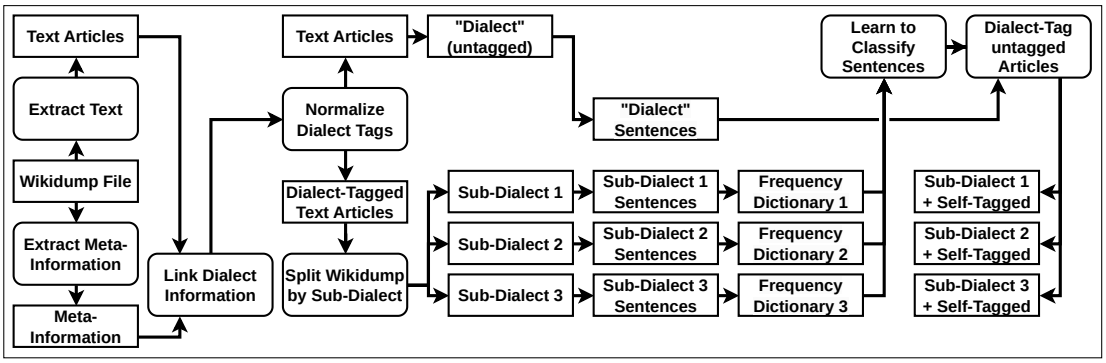


Figure B.1: Filtering wikidump data by sub-dialect tags.

The dialectal distribution of the Bavarian wikidump data shows the effect and potential of enriching the data by newly classifying texts based on observed word frequencies in already tagged articles. The cleaned data sorted by original tags (see Figure B.1) has many noisy identifier, which only slightly differ from each other or can otherwise be reasoned to be aggregated into fewer groups. Aggregating the provided dialect tags (see Figure B.3) into four main groups (and an additional default group simply called Bavarian, to capture those entries that have a strong overlap) provides a more comprehensive overview of the data and enables down-stream processing. During this step, all previously tagged sentences have had their classification been confirmed, or been

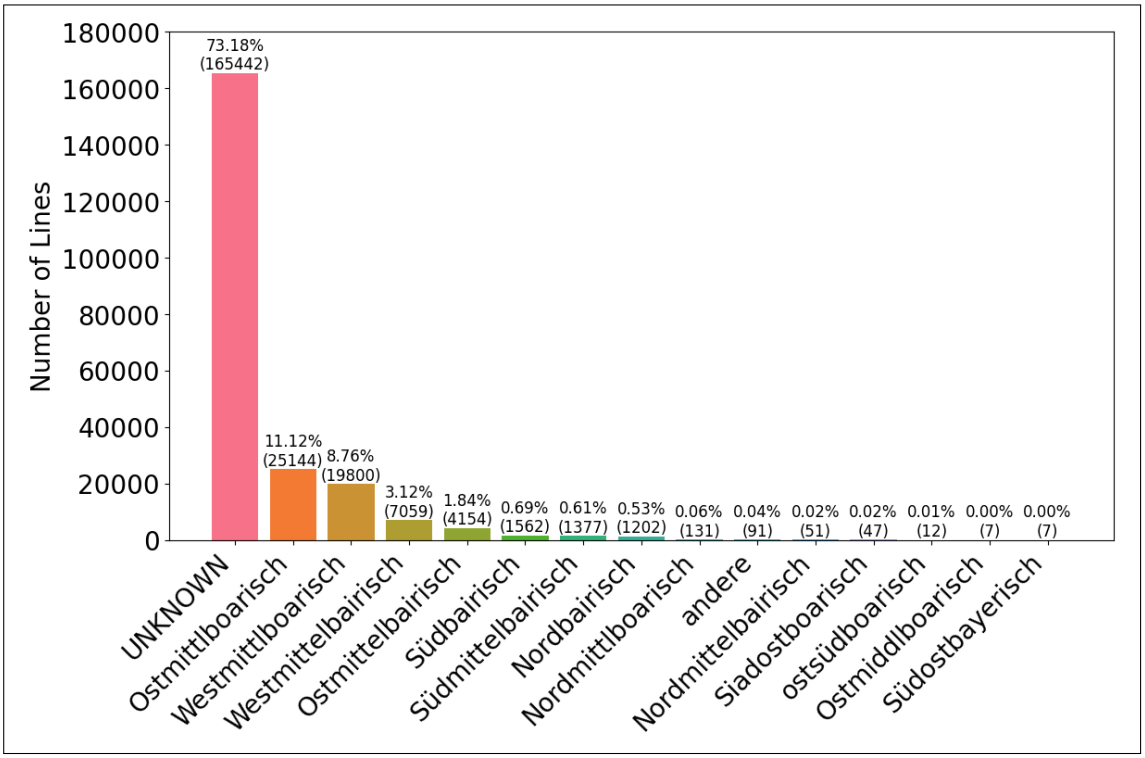


Figure B.2: Clean Bavarian wikidump data according to provided sub-dialect tags.

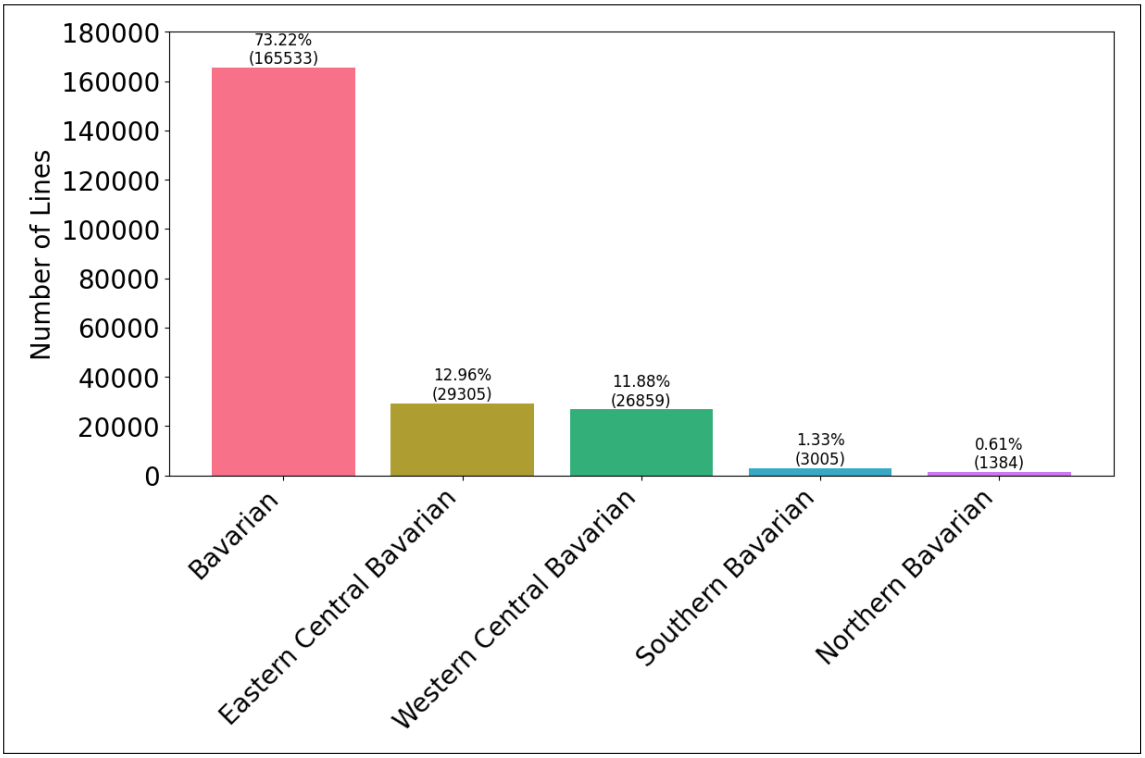


Figure B.3: Golden Bavarian wikidump data after normalizing sub-dialect tags into 4 groups + 1 general (overlapping) simply called Bavarian.

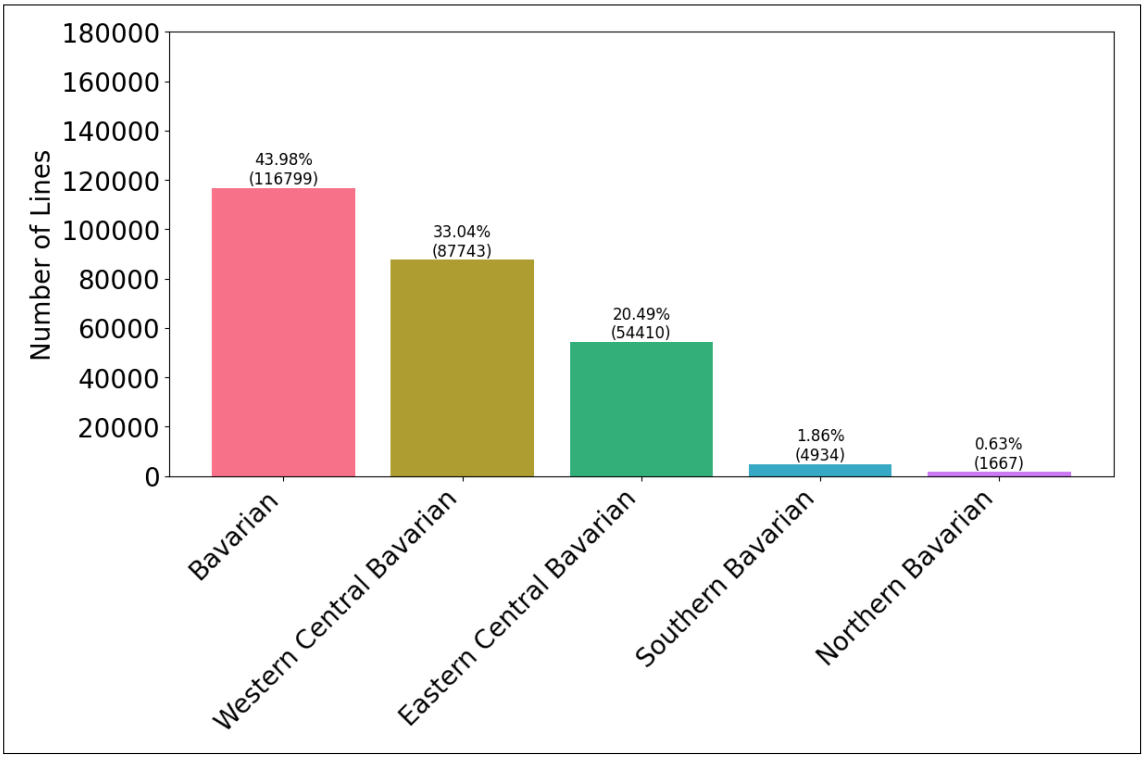


FIGURE B.4: Silver Bavarian wikidump data acquired by classifying previously untagged entries based on golden dialect-tagged word frequencies.

removed otherwise. For classification, a straight-forward approach was applied, based on counting the words in each sentence for each of the aggregated dialect-groups, which limitations will be discussed in Section 6.2.2. Finally, we have the newly classified sentences, shifting the balance away from the default group and especially benefiting the Western Central Bavarian dialect group (see Figure B.4).

Selected sample of words taken from the overlap of Bavarian DialectBLI data and the aggregated dialect groups from wikidump data is shown in Table B.1.

Dialect Group	**#words in wikidump**	**#words in DBLI (unique)**	**Example words**
Northern Bavarian	2,015	355,539 (576)	easchte, leitn, gnumma, gleichn, jeds, joa, foigende, fia
Eastern Central Bavarian	84,359	831,530 (11,353)	sej, sands, untaschiedlich, örtlichn, doa, singvegl
Southern Bavarian	9,117	385,423 (1,785)	grot, só, nit, wead, karpatn, wiesn, laid, iwa, aufm
Western Central Bavarian	136,428	906,954 (16,907)	isn, vona, tua, uf, moan, eihgricht, vabindungsstroß

TABLE B.1: Exploring Bavarian parts of DialectBLI data based on derived wikidump dialect-specific word lists.

Glossary

African American Vernacular English A dialect of English spoken primarily by African Americans in the United States. It has distinct grammatical features and vocabulary. 4, 23

Alemannic A major dialect group of German spoken in parts of Germany, Switzerland, Austria, and Liechtenstein. It includes several sub-dialects such as Swiss German and Swabian. 26, 28, 30, 31, 40, 41, 65–68

Alemán Coloniero A sub-dialect of Low Alemannic spoken in Venezuela. It's a unique dialect preserved by German-speaking immigrants and their descendants. 65

American English The variety of English primarily spoken in the United States. It has its own distinct vocabulary, spelling, and pronunciation. 5

Arabic A Semitic language spoken in various dialects across the Middle East and North Africa. It's the liturgical language of Islam and one of the most widely spoken languages in the world. 3, 6, 7

Ardalanî (Ardalani, Sanandajî, Sanandaji, Sanayî, Sanayi, Senayî, Senayi, Sine'i, Sine'î, Sine) A sub-dialect of Central Kurdish spoken in the region of Sanandaj, Iran. It has distinct phonological and lexical features. 12, 13

Argentinean Spanish A variety of Spanish spoken in Argentina. It's characterized by unique pronunciation, vocabulary, and grammar. 5

Australian English The variety of English spoken in Australia. It has its own distinct accent, vocabulary, and idioms. 5

Barisal A variety of Bengali named after the Barisal region in Bangladesh. It has its own distinctive features in pronunciation and vocabulary. 12, 13

Bavarian A major dialect group of German spoken in Bavaria, Austria, and South Tyrol. It's divided into Northern, Central, and Southern Bavarian sub-dialects. vii, 6, 8, 25–27, 29–32, 35–38, 40, 65, 66, 68

Belgian French The variety of French spoken in Belgium. It has some distinct vocabulary and pronunciation compared to Standard French. 5

Bengali An Indo-Aryan language primarily spoken in Bangladesh and the Indian state of West Bengal. It's known for its rich literary tradition. vi, 13, 31

Bernese German A sub-dialect of High Alemannic spoken in the canton of Bern, Switzerland. It's characterized by its unique pronunciation and vocabulary. 65

Brandenburgisch A sub-dialect of East Low German spoken in the Brandenburg region of Germany. It shares features with both Low German and Standard German. 64

British English The variety of English primarily spoken in the United Kingdom. It encompasses several regional accents and dialects. 5

Canadian French The variety of French spoken in Canada, primarily in Quebec. It has distinct vocabulary and pronunciation compared to European French. 5

Central Bavarian A dialect sub-group of Bavarian spoken in parts of Bavaria and most of Austria. It's the most widely spoken Bavarian dialect. vii, 8, 65

Central Franconian A sub-dialect of Franconian spoken in parts of western Germany. It includes Ripuarian and Moselle Franconian. 65

Central Kurdish A major dialect group of Kurdish, mainly spoken in Iraq and Iran. It's also known as Sorani. vi, 12, 13

Central Southern Bavarian A dialect sub-group of Bavarian spoken in parts of Austria and northern Italy. It shares features with both Central and Southern Bavarian. 65

Chilean Spanish A variety of Spanish spoken in Chile. It's known for its unique vocabulary and pronunciation. 5

Chinese A group of language varieties spoken by the Han Chinese and many other ethnic groups in China. Mandarin is the most widely spoken variety. 14, 15, 31

Danube Bavarian A sub-dialect of Bavarian spoken along the Danube River in Austria and Germany. It shares features with both Central and Southern Bavarian. 12

Dhakaiya A variety of Bengali named after Dhaka, the capital of Bangladesh. It has distinctive features in pronunciation and vocabulary. 12, 13

East Franconian A sub-dialect of Franconian spoken in parts of Bavaria and Thuringia. It's transitional between Central and South Franconian. 65

East Low German A dialect sub-group of Low Saxon spoken in northeastern Germany. It includes Mecklenburgisch-Vorpommersch and Brandenburgisch. 64

East Middle German A dialect sub-group of Middle German spoken in eastern Germany. It includes Upper Saxon, Thuringian, and Lusatian. 65

East Upper German An alternative denomination for Bavarian, emphasizing its geographic location within the Upper German dialect group. 65

Eastphalian A sub-dialect of West Low German spoken in parts of Lower Saxony and Saxony-Anhalt. It's known for its distinctive vocabulary and pronunciation. 64

English An Indo-European language originating in England, now spoken globally. It's the most widely spoken language in the world by total number of speakers. vii, 2, 5, 7, 8, 12, 15, 20, 22–24, 26, 30, 32, 36–39, 62, 63

Farsi Also known as Persian, it's the official language of Iran, Afghanistan, and Tajikistan. It's an Indo-Iranian language with a rich literary tradition. 7

Franconian A major dialect group of German spoken in parts of western and central Germany. It includes several sub-dialects like Rhine Franconian and East Franconian. 65

French A Romance language originating in France, now spoken worldwide. It's an official language in 29 countries and widely used in international diplomacy. 5, 7

German A West Germanic language, written in Latin script and mainly spoken in Germany, Austria, and Switzerland. It's the most widely spoken native language in the European Union. vi, vii, 7, 8, 12, 21, 24–26, 29–31, 35–40, 44, 62, 63, 65–68

Hessian A sub-dialect of West Middle German spoken in the state of Hesse, Germany. It has distinctive vocabulary and pronunciation. 64

Hewlêrî (Hewleri, Hewlêr, Hewler) A sub-dialect of Central Kurdish spoken in the region of Erbil (Hewlêr) in Iraq. It has unique phonological and lexical features. 12, 13

High Alemannic A dialect sub-group of Alemannic spoken in parts of Switzerland and neighboring regions. It includes Swiss German dialects. 65, 68

Highest Alemannic A dialect sub-group of Alemannic spoken in the southern parts of Switzerland and in some alpine regions. It's known for preserving many archaic features of Old High German. 65, 68

Hindi An Indo-Aryan language spoken primarily in India. It's one of the official languages of India and is written in the Devanagari script. 5

Indian English A group of English dialects spoken in India. It has distinct features influenced by Indian languages. 5

Italian A Romance language primarily spoken in Italy. It's known for its musical quality and significant influence on Western culture. 6

Jessore A variety of Bengali named after the Jessore region in Bangladesh. It has its own distinctive features in pronunciation and vocabulary. 12, 13

Khulna A variety of Bengali named after the Khulna region in Bangladesh. It has unique linguistic features compared to Standard Bengali. 12, 13

Kobanî (Kobani) A sub-dialect of Northern Kurdish spoken in western Kurdistan, north Syria. It's named after the city of Kobanî. 30

Kurdish An Indo-Iranian language spoken by Kurds in Western Asia. It has several major dialect groups including Northern Kurdish (Kurmanji) and Central Kurdish (Sorani). 7, 30

Kurmanjî (Kurmanji) Also known as Northern Kurdish, it's the most widely spoken Kurdish dialect. It's used in Turkey, Syria, and parts of Iraq and Iran. 23, 30

Kushtia A variety of Bengali named after the Kushtia region in Bangladesh. It has distinctive features in pronunciation and vocabulary. 12, 13

Lake Constance Alemannic A sub-dialect of Alemannic spoken around Lake Constance. It has features of both Low Alemannic and High Alemannic, with some Swabian influence. 41, 65, 68

Liechtensteinisch A sub-dialect of High Alemannic spoken in Liechtenstein. It's closely related to Swiss German dialects. 65

Ligurian A Gallo-Italic language mainly spoken in Liguria, northern Italy. It's considered either a separate language or a dialect of Italian. 21

Low Alemannic A dialect sub-group of Alemannic spoken in parts of southwestern Germany and Alsace. It includes dialects like Upper Rhine Alemannic. 41, 65, 67

Low Saxon Also called Low German, it's a major dialect group of German spoken in northern Germany and eastern Netherlands. It's closely related to Dutch and Frisian. 64

Lusatian A sub-dialect of East Middle German spoken in Lusatia, a region in eastern Germany. It has been influenced by contact with Sorbian languages. 65

Luxembourgish A sub-dialect of West Middle German that has gained official language status in Luxembourg. It has influences from both German and French. 64

Mauritian Creole A French-based creole language spoken in Mauritius. It's the most widely spoken language on the island. 7, 30

Mecklenburgisch-Vorpommersch A sub-dialect of East Low German spoken in Mecklenburg-Vorpommern, Germany. It has distinctive vocabulary and pronunciation. 64

Mexican Spanish A variety of Spanish spoken in Mexico. It has unique vocabulary and pronunciation influenced by indigenous languages. 5

Middle German A dialect group of German spoken in central Germany. It's divided into West Middle German and East Middle German. 62, 64, 65

Middle High German The form of German spoken in the High Middle Ages, from about 1050 to 1350. It's the ancestor of modern Standard German and many German dialects. 65

Mittelpommersch A sub-dialect of East Low German formerly spoken in Central Pomerania. Many of its speakers were displaced after World War II. 64

Moselle Franconian A sub-dialect of West Middle German spoken along the Moselle River in Germany, Luxembourg, and France. It's closely related to Luxembourgish. 64

Mukriyanî (Mukriyani, Mukrî, Mukri) A sub-dialect of Central Kurdish spoken in the region of Mahabad, Iran. It has distinct phonological and lexical features. 12, 13

Myanmar (Burmese) The official language of Myanmar (formerly Burma). It's a Sino-Tibetan language written in the Burmese script. 19

North Low Saxon A sub-dialect of West Low German spoken in northern Germany. It's closely related to Frisian and Dutch. 64

North Upper Saxon A sub-dialect of East Middle German spoken in parts of Saxony and Saxony-Anhalt. It's transitional between Upper Saxon and Low German. 65

Northern Bavarian A dialect sub-group of Bavarian spoken in Upper Palatinate and parts of Upper Franconia. It has distinctive phonological features. 8, 65

Northern Central Bavarian A transitional dialect between Central Bavarian and Northern Bavarian. It's spoken in parts of Upper Bavaria and Lower Bavaria. 65

Rhine Franconian A sub-dialect of West Middle German spoken along the Rhine River. It includes Palatine German and Hessian. 64

Ripuarian A sub-dialect of West Middle German spoken around Cologne and Aachen. It's closely related to Moselle Franconian. 64, 66

Saterland Frisian A dialect of Frisian spoken in Saterland, Lower Saxony, Germany. It's the last surviving dialect of East Frisian. 9

Saxon A major dialect group of German spoken in Saxony and surrounding areas. It includes Upper Saxon and North Upper Saxon. 12

Scottish Gaelic A Celtic language spoken mainly in the Scottish Highlands and the Hebrides. It's closely related to Irish and Manx. 18

Soranî (Sorani) A sub-dialect of Central Kurdish, often used as a name for Central Kurdish itself. It's mainly spoken in Iraq and Iran. 30

South Franconian A sub-dialect of Franconian spoken in northern Baden-Württemberg. It's transitional between Central and Upper German dialects. 65

Southern Bavarian A dialect sub-group of Bavarian spoken in southern Bavaria, Austria, and South Tyrol. It has distinctive phonological features. 8, 65

Spanish A Romance language originating in Spain, now spoken worldwide. It's the world's second-most spoken native language. 5

Standard American English The dialect of English considered standard in the United States. It's used in formal contexts and media. 4, 23

Standard Bengali The standardized form of Bengali, based on the dialect spoken in West Bengal, India. It's used in formal contexts and media. 12, 13

Standard German The standardized version of German used in formal contexts, education, and media. It's based on Middle and Upper German dialects. 8, 12, 21, 27, 32, 62, 65–67

Swabian A sub-dialect of Alemannic spoken in Swabia, southwestern Germany. It's known for its distinctive pronunciation and vocabulary. 26, 65, 67

Swiss German The variety of Alemannic German spoken in Switzerland. It encompasses several dialects and is used in everyday communication. 18, 21, 65

Thuringian A sub-dialect of East Middle German spoken in Thuringia, Germany. It has features transitional between Upper and Central German. 65

Turkish A Turkic language primarily spoken in Turkey and Cyprus. It's known for its agglutinative structure and vowel harmony. 7

Upper German A major dialect group of German spoken in southern Germany, Austria, Switzerland, and South Tyrol. It includes Alemannic and Bavarian. 62, 65–67

Upper Rhine Alemannic A sub-dialect of Low Alemannic spoken along the Upper Rhine in Germany and Alsace. It has features transitional to High Alemannic. 41, 65

Upper Saxon A sub-dialect of East Middle German spoken in much of Saxony. It has significant influence on the pronunciation of Standard German. 65

Vorarlberg German A sub-dialect of High Alemannic spoken in Vorarlberg, Austria. It's closely related to Swiss German dialects. 65

Walliser German A sub-dialect of High Alemannic spoken in Valais, Switzerland. It's known for preserving many archaic features. 65

Walser German A sub-dialect of High Alemannic spoken by the Walser people in parts of Switzerland, Italy, and Austria. It's derived from Walliser German. 65

West Low German A dialect sub-group of Low Saxon spoken in northwestern Germany and northeastern Netherlands. It includes North Low Saxon and Westphalian. 64

West Middle German A dialect sub-group of Middle German spoken in western Germany. It includes Moselle Franconian, Ripuarian, and Luxembourgish. 64

West Upper German An alternative denomination for Alemannic, emphasizing its geographic location within the Upper German dialect group. 65

Westphalian A sub-dialect of West Low German spoken in Westphalia and parts of Lower Saxony. It has distinctive vocabulary and pronunciation. 64

Yiddish A Germanic language historically spoken by Ashkenazi Jews. It developed from Middle High German with significant Hebrew and Aramaic influences. 65

Zurich German A sub-dialect of High Alemannic spoken in and around Zurich, Switzerland. It's one of the most prominent Swiss German dialects. 65

Declaration of Authorship

Eidestattliche Erklärung

Hiermit versichere ich, Christian Schuler, an Eides statt, dass ich die vorliegende Arbeit im Masterstudiengang Informatik selbstständig verfasst und keine anderen als die angegebenen Hilfsmittel – insbesondere keine im Quellenverzeichnis nicht benannten Internet-Quellen – benutzt habe. Alle Stellen, die wörtlich oder sinngemäß aus Veröffentlichungen entnommen wurden, sind als solche kenntlich gemacht. Ich versichere weiterhin, dass ich die Arbeit vorher nicht in einem anderen Prüfungsverfahren eingereicht habe.

Statutory Declaration

I, Christian Schuler, hereby certify under oath that I wrote this thesis in the Master's program in Computer Science independently and that I did not use any resources other than those specified - in particular no Internet sources not named in the list of sources. All passages that have been taken literally or essentially from publications are identified as such. I further certify that I have not previously submitted the work to another examination procedure.

Signature
Christian Schuler