DISSERTATIONES
INFORMATICAE
UNIVERSITATIS
TARTUENSIS
65

ANDRE TÄTTAR

Multilingual Machine Translation for Under-Resourced Languages

DISSERTATIONES INFORMATICAE UNIVERSITATIS TARTUENSIS
65

DISSERTATIONES INFORMATICAE UNIVERSITATIS TARTUENSIS
65

ANDRE TÄTTAR

Multilingual Machine Translation for Under-Resourced Languages

UNIVERSITY OF TARTU
Press
1632

Institute of Computer Science, Faculty of Science and Technology, University of Tartu, Estonia.

Dissertation has been accepted for the commencement of the degree of Doctor of Philosophy (PhD) in Computer Science on April 1, 2025 by the Council of the Institute of Computer Science, University of Tartu.

Supervisor

Prof.	Mark Fišel University of Tartu Estonia

Opponents

Prof.	Jörg Tiedemann University of Helsinki Finland
Dr.	David Vilar Torres Google Berlin Germany

The public defense will take place on May 23, 2025 at 10:15 in Narva Rd. 18-2048.

The publication of this dissertation was financed by the Institute of Computer Science, University of Tartu.

ISSN 2613-5906 (print) ISSN 2806-2345 (pdf)
ISBN 978-9916-27-862-8 (print) ISBN 978-9916-27-863-5 (pdf)

University of Tartu Press
http://www.tyk.ee/

To my family and friends

ABSTRACT

Imagine a language family spoken by over 20 million people across Europe and North Asia. This isn't English, Mandarin, or Spanish – it's the Finno-Ugric language family, encompassing national languages like Hungarian ($\sim$13 million speakers), Finnish ($\sim$5.8 million), and Estonian ($\sim$1.2 million). But the richness of this family extends far beyond these widely spoken members. There are 42 languages in total, many with speakers numbering in the tens of thousands or even hundreds. Sadly, these lesser-known languages face a decline, at risk of disappearing altogether.

This dissertation tackles this challenge by exploring multilingual neural machine translation (NMT) specifically designed to address the data scarcity these Finno-Ugric languages face. My work aims to pay attention to these smaller languages and provide open-source machine translation models for their communities, demonstrating the importance of their languages and showing that researchers are working on solutions for them. Additionally, I focus on creating a state-of-the-art Estonian-specific translation model that could compete with Google Translate and DeepL.

The study evaluates various NMT strategies, like the universal model for leveraging higher resource language for lower resource models, making models more language-specific using the modular approach and using large pre-trained models to ultimately improve translation quality across a wide range of languages, ranging from resource-rich to low-resource scenarios. We show that the back-translation technique remains effective, even in settings with minimal resources.

This thesis addresses the significant challenges and proposes solutions for low-resource language translation. It provides valuable insights into optimizing Neural Machine Translation (NMT) systems for languages with extremely limited resources. The research evaluates the impact of these optimized systems on minor languages and their broader implications in the field of natural language processing (NLP).

The work introduces novel methodologies, such as fine-tuning pre-trained models for previously unseen languages and leveraging knowledge-sharing techniques among closely related languages. These approaches enable substantial advancements in multilingual NMT, particularly for zero-resource language pairs. Notably, this research created the first Finno-Ugric multilingual models, initially covering 5 languages and expanding to 23 languages over the course of the thesis. This includes many endangered languages, such as Livonian, which currently has just one native speaker and forty L2 speakers. This demonstrates the effectiveness of using large pre-trained models as a foundation for training with very limited parallel data.

Beyond aiding communication within Finno-Ugric communities, these findings offer a foundation for significant advancements in NMT for resource-scarce languages globally, demonstrating the potential for adaptation to other languages.

CONTENTS

List of original publications **12**

Preface **15**

1. Introduction **16**
1.1. Introduction 16
1.2. Motivation 16
1.3. Research methodology and goals 17
1.4. Significance and Impact 18
1.5. Societal Impact of the Research 19
1.6. Structure of the thesis 22

2. Background **23**
2.1. Core NMT Concepts 23
2.1.1. Transformers 25
2.2. NMT Vocabulary 26
2.2.1. Byte-Pair Encoding (BPE) 27
2.3. Data Concepts 28
2.3.1. Data Processing and Filtering 29
2.4. Data Augmentation 29
2.5. MT Evaluation 30
2.6. Transfer Learning 33
2.6.1. Finetuning 33
2.7. Universal NMT Overview 34
2.8. Modular NMT 37
2.9. Using pre-trained multilingual models 38

3. Multilingual shared NMT models **39**
3.1. University of Tartu's Multilingual Multi-domain WMT19 News Translation Shared Task Submission 39
3.2. Extremely low-resource machine translation for closely related languages 41

4. Modular NMT **45**
4.1. Open and Competitive Multilingual Neural Machine Translation in Production 45
4.2. Multilingual Neural Machine Translation With the Right Amount of Sharing 49

5. Large-pretrained Models enable high-quality low-resource NMT **54**
5.1. Cross-lingual Transfer from Large Multilingual Translation Models to Unseen Under-resourced Languages 55
5.2. Teaching unseen low-resource languages to large translation models 58
5.3. Machine Translation for Low-resource Finno-Ugric Languages . . 59

6. Conclusion **64**

Bibliography **66**

Acknowledgements **75**

Sisukokkuvõte (Summary in Estonian) **76**

Publications **79**

Curriculum Vitae **163**

Elulookirjeldus (Curriculum Vitae in Estonian) **164**

LIST OF FIGURES

1. Diagram showing (a) a set of single directional models where no parameter sharing across language pairs is possible, and (b) multilingual many-to-many universally shared model, where only a single encoder and decoder are shared across all language pairs. 23
2. Diagram showing (a) a set of single directional models where no parameter sharing across language pairs is possible, and (b) multilingual modular model, where encoders and decoders are shared across all language pairs. 37
3. German↔Estonian translation BLEU scores. MTee refers to our model developed in the paper. 47
4. English↔Estonian translation BLEU scores. 48
5. Russian↔Estonian translation BLEU scores. 48
6. Different granularities of the modular architecture. *roa* – Romance; *gem* – Germanic; *tgt lang* – Target language token added to indicate the language of the output sentence. 50
7. Different types of encoder layer sharing in the modular architecture. Note that the width of layers in the figure does not correspond to the actual width but rather reflects the sharing extent, i.e., all layers in the encoder have the same width dimension. *U* – universal, *G* – Germanic, *R* – Romance. 51

LIST OF TABLES

1. Examples of machine-translated sentences using TartuNLP's online system. The first sentence is translated from Estonian to English, the second from English to Estonian with first sentence translation used as source, and the third is an example to show different word choices for similar meanings. 16
2. Example of BPE tokenization. Here, "Tere" remains unchanged as it's likely frequent, while "maailm" is broken down into "maa@@", which continues on, ("maa" meaning "land") and "ilm" ("ilm" suffix changes the word into "world"). 27
3. Dataset sizes (number of parallel sentences) after filtering in thousands(K). **NEWS** contained Rapid2019, Rapid2016, EESC, dev datasets from previous years, EMEA2016, ECB2017, news (from CzEng) and News-commentary corpuses. **OFF** comprised parts of the CzEng corpus and Europarl. **SUBS** were from OpenSubtitles. **OTHER** was anything else. 39
4. Results of the multilingual baseline model and the fine-tuned model. 40
5. Parallel data sets (in sentence pairs). ET - Estonian, FI - Finnish, VRO - Võro, SME - North Sámi, SMA - South Sámi. Further details can be found in the paper (Tars et al. (2021)). 42
6. Monolingual data sets after preliminary cleaning (in sentences). . . 43
7. BLEU scores. (*) - trained without pre-trained weights, *BT* - back-translation data set, *FT* - forward-translation data set, $BLEU_{low}$ - average BLEU score on low-resource language pairs (excluding ET-FI and FI-ET), **bold** - best BLEU score for a language pair. Check the original paper for a complete table. 44
8. BLEU and CHRF scores for transfer learning and fine-tuning experiments. 44
9. Parallel data split size. K means thousands, and M means millions of sentences per domain. 46
10. Monolingual data size in thousands (K) or millions (M) of sentences per domain. 47
11. Dataset sizes in thousands of sentence pairs for each language pair. 50
12. Average test set BLEU scores per language pair resource, all refers to average score over all resource settings. EG - encoder layer shared within language group, EA - encoder layer shared between all languages. Best score(s) per resource (column) in bold. 52
13. Average test set BLEU scores per language pair resource, all refers to average score over all resource settings. BLEU scores are given for experiments with encoder layer sharing between all languages in the language modular model. Best score(s) per resource (column) in bold. 53

14. Parallel data numbers before and after filtering (in sentence pairs) are covered in columns raw and filtered. Validation and test set sentence counts are the other two columns. 55
15. BLEU scores with test data from Tars et al., 2021. “418M" and “1.2B" refer to models trained with all data. “not-EN" refers to a model trained without any English data. “only-(VRO/SM*)" refers to a model trained only on those specific languages and pairs. *prev_best* refers to best results by Tars et al. (2021). **bold** - best BLEU score for a language pair. 56
16. BLEU scores on our new test set and liv4ever test set. “418M" and “1.2B" refer to models trained with all data. “not-EN" refers to model trained without any English data. “only-(VRO/LIV)" refers to models trained only on that specific language group data. *prev_best* refers to results by Rikters et al. (2022) on their fine-tuned models (without back-translation). **bold** - best BLEU score for a language pair. 57
17. Monolingual data numbers, with expanded data, filtering, and language coverage compared to Table:6. Check the paper for more details (Tars et al. (2022b)). 58
18. Experiment results on BLEU. “1.2B” and “418M” refer to models trained with all original parallel data. “bt1” is trained on parallel + first back-translation iteration data, “bt2” on parallel + second back-translation iteration data. “only-LIV” - only data between ET-EN-LV-LIV languages was used for training. “tuned on LIV” refers to the “bt2” model that was tuned on ET-EN-LV-LIV original parallel data. The last row represents previously best results for EN-LIV-EN by (Rikters et al., 2022). 59
19. The collected parallel corpus with Russian. The figures in the table are in thousands of sentences. 60
20. The collected monolingual corpus of the low-resource languages. The figures in the table are in thousands of sentences. 61
21. Average chrF++ results for all experiments across different language pair clusters on FLORES benchmarks. **Bold** - highest score per grouping. all-all - contains BT data from every language pair. all-all-10 - contains 10% of BT data used in all-all. L-H - contains BT data from each low-resource(L) language to each high-resource(H) language and vice versa. L-rH - contains BT data from each language to its related high-resource language(rH) + high-resource languages it had parallel data with and vice versa. 63

LIST OF ORIGINAL PUBLICATIONS

Publications included in the thesis

I Tättar, A., Korotkova, E., & Fishel, M. (2019). University of Tartu's multilingual multi-domain WMT19 news translation shared task submission. *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, 382–385. https://doi.org/10.18653/v1/W19-5342

II Tars, M., Tättar, A., & Fišel, M. (2021). Extremely low-resource machine translation for closely related languages. *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, 41–52.

III Tättar, A., Purason, T., Kuulmets, H.-A., Luhtaru, A., Ratsep, L., Tars, M., Pinnis, M., Bergmanis, T., & Fishel, M. (2022). Open and competitive multilingual neural machine translation in production. *Baltic Journal of Modern Computing*, *10*(3), 422–434. https://doi.org/10.22364/bjmc.2022.10.3.15

IV Tars, M., Purason, T., & Tättar, A. (2022a). Teaching unseen low-resource languages to large translation models. *Proceedings of the Seventh Conference on Machine Translation (WMT)*, 375–380.

V Yankovskaya, L., Tars, M., Tättar, A., & Fishel, M. (2023). Machine translation for low-resource Finno-Ugric languages. *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, 762–771.

VI Purason, T., & Tättar, A. (2022). Multilingual neural machine translation with the right amount of sharing. *Proceedings of the 23rd Annual Conference of the European Association for Machine Translation*, 91–100.

VII Tars, M., Tattar, A., & Fishel, M. (2022b). Cross-lingual transfer from large multilingual translation models to unseen under-resourced languages. *Baltic Journal of Modern Computing*, *10*(3), 435–446. https://doi.org/10.22364/bjmc.2022.10.3.16

The summary of my contributions can be categorized into three main categories:

- **Lead Author Contributions**
  - **Publication I**: Trained and fine-tuned the multilingual machine translation model. Collected, filtered, and preprocessed data. Conducted experiments with back-translation to create synthetic parallel data. Analyzed results and wrote the paper.
  - **Publication III**: Main author and the technical coordinator of the entire governmental project. Led the design and implementation of all NMT systems and experiments. Analyzed results and led the writing process. Main contributor to the article. I consider this to be the biggest contribution and achievement of my academic career.

- **Supervisory role**: Managed to start and lead a small machine translation team at the start of work on the **Publication III**, consisting of me and two university employees who were also students at the time. I lead the team and oversaw their work and helped them become independent researchers and advance their academic careers. This applies to **Publications II, IV, VI, and VII**. For **Publications IV** and **VI**, I was the last author. Supervising helped me become a better manager and mentor. Specifically, I organized daily work, including daily and/or weekly meetings, and supported the team members in their day-to-day work with any questions they had. I collaborated with the team daily on the idea, research, development, experiments, interpreting the results, and co-writing the paper. I would like to highlight Publication II, which was originally my idea and initiated the group's efforts in Finno-Ugric focused machine translation.
- **Collaboration effects** I participated in a wide range of research in the machine translation field which has broadened my expertise, with many publications not being part of this thesis, but I would like to bring out one publication, which I included in the thesis. I had mainly an advisory role and a small supervisory role for the second author for **Publication: V**. I included this work since it directly builds upon the work of this thesis.

Publications not included in the thesis

VIII Bawden, R., Zhang, B., Tättar, A., & Post, M. (2020a). ParBLEU: Augmenting metrics with automatic paraphrases for the WMT'20 metrics shared task. *Proceedings of the Fifth Conference on Machine Translation*, 887–894. https://aclanthology.org/2020.wmt-1.98

IX Bawden, R., Zhang, B., Yankovskaya, L., Tättar, A., & Post, M. (2020b). A study in improving BLEU reference coverage with diverse automatic paraphrasing. *Findings of the Association for Computational Linguistics: EMNLP 2020*, 918–932. https://doi.org/10.18653/v1/2020.findings-emnlp.82

X Bergmanis, T., Pinnis, M., Rozis, R., Šlapiņš, J., Šics, V., Bernāne, B., Pužulis, G., Titomers, E., Tättar, A., Purason, T., Kuulmets, H.-A., Luhtaru, A., Rätsep, L., Tars, M., Laumets-Tättar, A., & Fishel, M. (2022). MTee: Open machine translation platform for Estonian government. *Proceedings of the 23rd Annual Conference of the European Association for Machine Translation*, 309–310. https://aclanthology.org/2022.eamt-1.44

XI Del, M., Tättar, A., & Fishel, M. (2018). Phrase-based unsupervised machine translation with compositional phrase embeddings. *Proceedings of the Third Conference on Machine Translation*, 365–371. http://www.aclweb.org/anthology/W18-64034

XII Ingel, A., Shahroudi, N., Kängsepp, M., Tättar, A., Komisarenko, V., & Kull, M. (2020). Correlated daily time series and forecasting in the m4 com-

petition [M4 Competition]. *International Journal of Forecasting*, *36*(1), 121–128. https://doi.org/https://doi.org/10.1016/j.ijforecast.2019.02.018

XIII Kuulmets, H.-A., Tattar, A., & Fishel, M. (2022). Estonian language understanding: A case study on the copa task. *Baltic Journal of Modern Computing*, *10*(3), 470–480. https://doi.org/10.22364/bjmc.2022.10.3.19

XIV Purason, T., Tättar, A., & Fishel, M. (2024). Mixing and matching: Combining independently trained translation model components. In R. Vázquez, T. Mickus, J. Tiedemann, I. Vulić, & A. Üstün (Eds.), *Proceedings of the 1st workshop on modular and open multilingual nlp (moomin 2024)* (pp. 44–56). Association for Computational Linguistics. https://aclanthology.org/2024.moomin-1.5

XV Tättar, A., & Fishel, M. (2017). Bleu2vec: The painfully familiar metric on continuous vector space steroids. *Proceedings of the Second Conference on Machine Translation, Volume 2: Shared Task Papers*, 619–622. http://www.aclweb.org/anthology/W17-4771

XVI Yankovskaya, E., Tättar, A., & Fishel, M. (2018). Quality estimation with force-decoded attention and cross-lingual embeddings. *Proceedings of the Third Conference on Machine Translation*, 829–834. http://www.aclweb.org/anthology/W18-64094

XVII Yankovskaya, E., Tättar, A., & Fishel, M. (2019). Quality estimation and translation metrics via pre-trained word and sentence embeddings. *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, 103–107. http://www.aclweb.org/anthology/W19-5410

PREFACE

Estonia, a small and beautiful country in Europe, boasts a unique linguistic heritage as Estonians speak the Estonian language, which belongs to the Finno-Ugric language family — one of the smallest language families in the world, famous for its intriguing and distinctive features.

My grandparents were all from South Estonia, and their native language was the Võru language instead of the official Estonian Language. When I was growing up, they mostly spoke a mix of Estonian and Võru, which still affects my language usage. I have always thought that speaking and listening to the Võru language brings me joy and gives me a fascination with the diversity of languages and cultures - so I am incredibly lucky that during my Ph.D. studies, I got to impact the Estonian and the Võru languages positively.

As I pursued a master's degree in Computer Science, I was introduced to the wonderful world of Natural Language Processing and Machine Translation by my supervisor Mark Fišel. What started as a simple course project for Machine Learning grew into a long and fruitful cooperation and successful research track in the field of Machine Translation.

Throughout the Ph.D. journey, my commitment to breaking down language barriers and empowering underrepresented languages (mainly in the Finno-Ugric family) only grew stronger. I am deeply passionate about contributing to advancements in natural language processing and artificial intelligence by developing efficient and versatile multilingual neural machine translation systems. My aspiration is to bring high-quality research to the end-users who might feel like their language has been left out of the developments of NLP.

As I share the discoveries and insights from my Ph.D. research, I hope to make meaningful contributions to Estonians, the Estonian language, and the Finno-Ugric language family in general. I sincerely hope that you enjoy reading this thesis.

1. INTRODUCTION

1.1. Introduction

Machine translation has been a rapidly evolving and vibrant field that has brought significant scientific advancements to natural language processing and artificial intelligence in general. The transformer model by Vaswani et al. (2017), highly popular due to GPT-like and large language models currently, grew out of machine translation research. The commencement of my doctoral studies coincided with the pivotal and groundbreaking release of the transformer model, and this thesis will cover the usage of transformer models for multilingual neural machine translation in low to high-resource scenarios. In this introduction, I will briefly give an overview of this thesis, covering the motivation, methodology, significance, and objectives.

Machine Translation (MT) refers to the use of software to translate text or speech from one language into another. The goal of MT is to automatically produce high-quality translations that are equivalent to those generated by human translators. Table 1 illustrates the nuances of languages and translation systems. The translations are created with TartuNLP's online system[1].

Source Sentence	**Translated Sentence**
Masintõlge on keerukas ja huvitav uurimisvaldkond.	Machine translation is a complex and interesting field of research.
Machine translation is a complex and exciting field of research.	Masintõlge on keeruline ja huvitav uurimisvaldkond.
Machine translation is a sophisticated and fascinating research field.	Masintõlge on keerukas ja põnev uurimisvaldkond.

Table 1: Examples of machine-translated sentences using TartuNLP's online system. The first sentence is translated from Estonian to English, the second from English to Estonian with first sentence translation used as source, and the third is an example to show different word choices for similar meanings.

1.2. Motivation

The primary motivation for multilingual models is the scarcity of parallel data, with low computational resources being a secondary consideration. Machine translation models are expensive to train and expensive to host. The main driver for this was that the standard approach to machine translation was single-directional models prior to (Johnson et al., 2017), which meant that a complete system with seven languages meant training and hosting 42 models (all combinations, 7×6), one for each translation direction (see Figure:1). This, however, is extremely

[1] https://neurotolge.ee/

expensive, does not scale well and requires a lot of monitoring. Alternatively, our objective has been to train and maintain a single multilingual system with the same capacity as a single-directional model. An additional motivation for our work has been that languages often share similar traits (e.g., languages belonging to the same language family) - so it makes sense to investigate the ability of the models to learn multilingual representations of similar concepts to improve the translation service's overall performance. Our final main motivation is directly tied to the previous point - low-resource translation quality - where the goal is to use any data to improve the performance of machine translation for language directions where very little data is available.

A common theme in this thesis is that most of the experiments are Estonian-centric because the overall goal has been to further push the quality of Estonian machine translation. Estonian is very interesting because, combined with other languages, we can get a few high-resource language pairs (English-Estonian), a few medium-resource language pairs (Estonian-Czech), and a lot of low (Estonian-Arabic) or zero-resource language pairs (Estonian-Northern Sámi).

Beyond the Estonian language, the Finno-Ugric language proposes a unique challenge with resource constraints. Languages can exhibit significant variation in dialects, writing systems (orthography), and even within geographically close regions. For example, the Võro language, spoken in southeastern Estonia, has multiple orthographic variations depending on location. This research strives to develop a multilingual NMT system that can address resource scarcity. Additionally, many Finno-Ugric languages like the Komi or Mari use the Cyrillic alphabet instead of the Latin transcript, which makes generalization more challenging.

All these motivations lead to the overall theme of the research conducted, where we try to find a single multilingual system that is efficient (in terms of data and computational resources) while achieving competitive translation quality.

1.3. Research methodology and goals

The methodology, which has been consistent over the course of this PhD dissertation, can be summarized as follows:

- Develop a multilingual neural machine translation system that efficiently handles low, medium, and high-resource scenarios while achieving competitive translation quality.
- Investigate the ability of the multilingual model to learn and generalize across languages to improve translation performance.
- Address the challenge of improving translation quality for zero- or low-resource languages by effectively utilizing any additional data, primarily via synthetic data creation.
- Develop and deploy usable systems for users by making the final versions of the models available to the general public via an interface.

1.4. Significance and Impact

The significance of this research lies in its potential to advance the field of machine translation and contribute to advancements in natural language processing and artificial intelligence. By developing a robust and efficient multilingual neural machine translation system, this research addresses several key challenges and opens up new possibilities for language translation in various resource scenarios.

1. Optimizing Resources for Multilingual Machine Translation. One of the primary contributions of this research is the development of a single multilingual neural machine translation system that efficiently handles low, medium, and high-resource scenarios. Traditional approaches to machine translation, with single-directional models for each language pair, are resource-intensive and impractical, especially for multilingual settings, as shown by Johnson et al. (2017). By proposing a unified system that can handle multiple languages with reduced resource requirements, this research significantly advances the field in terms of cost-effectiveness, scalability, and ease of monitoring.

2. Enhancing Cross-Lingual Generalization. Another crucial aspect addressed in this research is the investigation of the multilingual model's ability to learn and generalize across languages. By exploring the potential of multilingual representations, this study seeks to improve the overall performance of the translation service. The findings could lead to more effective cross-lingual transfer techniques, enabling the model to leverage knowledge from high-resource language pairs to enhance translation quality for low-resource and zero-resource language pairs. Additionally, complementary techniques such as knowledge distillation and data augmentation are employed to enhance the model's capabilities further, providing a comprehensive approach to improving translation quality.

3. Bridging the Gap Between Research, Low-Resource Languages, and Real-World Accessibility. Zero- and low-resource language pairs remain a major challenge in machine translation (Araabi et al., 2020; Haddow et al., 2022). Traditional models often struggle to produce satisfactory results due to limited training data. This research tackles this issue head-on by developing innovative approaches that effectively utilize all available data to improve translation quality in these scenarios. This work aims to empower linguistically marginalized communities, demonstrating the feasibility of translation even in zero-resource settings like Võro $\leftrightarrow$ Northern Sámi pair.

Furthermore, this research emphasizes practicality and accessibility by developing user-friendly systems. The goal of pushing the boundaries of Estonian/Finno-Ugric-centric machine translation is to benefit the Estonian-speaking community directly while serving as a model for similar efforts in other underrepresented languages. Accessible translation services have the power to break down communication barriers, fostering cross-cultural understanding and supporting linguistic diversity (Marzotto, 2019). One obvious benefit would be any crisis situation in communities, where people might need to communicate quickly, which would

be nearly impossible if there are no translators to be found, like the mission of Translators without Borders.[2]

4. Advancing the State of Machine Translation. The research objectives outlined in this dissertation collectively contribute to advancing the state of machine translation. By pushing the boundaries of translation quality across different resource scenarios, exploring the potential of multilingual representations, and addressing the challenges of low-resource translation, this work stands to make a significant impact on the field's theoretical foundations and practical applications.

The research objectives collectively contribute to advancing the state of machine translation by:

- Improving Cost-Efficiency and Scalability: The proposed multilingual neural machine translation system has the potential to reduce the costs associated with training and hosting multiple models for various language pairs, making machine translation more financially viable for a broader range of applications. The training costs are further lowered by using pre-trained models to fine-tune the models.
- Cross-Lingual Generalization: The findings on multilingual representations could lead to breakthroughs in transfer learning techniques such as knowledge distillation, using pre-trained models or data augmentation, benefiting other natural language processing tasks beyond machine translation and potentially improving the performance of various AI systems.
- Empowering Low-Resource Languages: By addressing the challenges of zero- or low-resource translation, this research makes machine translation accessible for underrepresented languages, thereby preserving cultural heritage. This is crucial for enabling communication between communities, such as during the *Hõimupäevad* event[3], which celebrates Finno-Ugric culture and languages. Additionally, machine translation aids in preserving endangered or near-extinct languages, such as Livonian, with only one native speaker, and several Sámi languages, which have only a few hundred speakers.
- Developing user-focused Applications: The focus on developing usable systems through research ensures that the benefits of machine translation are more accessible to end-users, promoting seamless communication across linguistic barriers.

1.5. Societal Impact of the Research

While this thesis primarily focuses on the technical and scientific aspects of multilingual neural machine translation (NMT), it is important to highlight the real-world impact that this research has had on communities, particularly those speak-

[2] https://translatorswithoutborders.org/

[3] https://fennougria.ee/

ing under-resourced Finno-Ugric languages. This section does not rely on formal scientific evaluation but aims to show the non-academic value derived from this work.

The translation systems developed as part of this research have attracted attention from both media and online communities. For instance, the University of Tartu's machine translation engine has been featured in several news outlets, including Postimees (2023), Baltic Times (2023), YLE (2023), Study in Estonia (2023), and University of Tartu (2023). These articles highlight the significance of this work for language preservation and technology, emphasizing its practical value and contribution to cultural sustainability—especially for languages at risk of decline.

Beyond media attention, the reception from language communities has been overwhelmingly positive. On platforms such as Reddit's *r/finnougric*, the system has received enthusiastic feedback, becoming the second most upvoted post of all time in that community. Users expressed excitement about the availability of translation tools for languages that have historically lacked digital resources. Comments like, "I never thought I'd see my native language supported by a machine translation system," and "This is a huge step for our linguistic identity," reflect the deep personal and cultural significance of this work.[4]

In addition to media coverage and online feedback, key figures within language communities have recognized the importance of this work. Sulev Iva, PhD — Lecturer in South Estonian Language and Culture at the University of Tartu and Research Fellow at the Võro Institute—remarked:

> "In collaboration with the University of Tartu's machine translation team, a Võro machine translation application based on neural networks (neurotolge.ee) has been developed... The Võro neural translation has certainly significantly improved language technology support for Võro and made it more accessible to learners and enthusiasts. The current quality of the translation is such that with light editing, it is already quite usable, at least in the Võro-Estonian-Võro direction, truly saving the translator's time and effort.
>
> For the Võro language, as well as for many other small Finno-Ugric languages, having its own machine translation is extremely important. When last year, languages like Latgalian and about a hundred other small nations received their own Google Translate, it made the Võro people somewhat envious. However, we had our own machine translation of much better quality already years before Google provided it to the Latgalians and many other small nations."

[4] https://www.reddit.com/r/finnougric/comments/11g8bt8/automatic_translation_for_23_finnougric_languages/

For clarity, here is the original quote in Estonian:

> Kuuntüün TÜ massintõlkõ tüürühmäga om valmis saanuq närvivõrkõ meetodi perrä luud võro massintõlkõ rakõndus (neurotolge.ee). Tuu arõndamisõga om mitmit kõrdo parõmbas lännüq eesti-võro-eesti massintõlkõ tõlkmiskvalitiit ni om luud võimalus võro keelest ja võro kiilde massintõlkiq ka inglüse, soomõ, läti ja pall'osit tõisi kiili, ka vähämbit soomõ-ugri kiili. Valmissaanu neurotõlkõrakõndusõ olõmi integriirnüq ka miiq võro võrgosõnaraamatuhe synaq.org. Tuud niimuudu, et nii eesti-võro ku võro-eesti tsihin saa tarvitaja umalõ mitmõsõnalidsõlõ otsmisõlõ noq kõgõ automaatsõhe ka massintõlkõvastussõ. Võro neurotõlgõq om kimmähe kõvastõ parõmbas tennüq võro keele keeleteknoloogilist tukõ ni toonuq võro keele opjilõ ja huviliisilõ inämb kätte. Tõlkõ parhillanõ kvalitiit om sääne, et ku tuu kergehe üle toimõndaq, om tuu joba peris häste pruugitav vähämbält võro-eesti-võro tsihin, ni hoit joba tõtõstõ kokko ümbrepandja aigo ja vaiva.
>
> Võro keele, nigu arvadaq ka pall'odõ tõisi soomõ-ugri väikeisi kiili jaos om uma massinatõlkõ olõmanolõminõ väega tähtsä. Ku minevä aastaga saiq umakeelidse Google'i tõlkõ näütüses latgallasõq ja viil sadakund väikeist rahvast, kinkal tuud inne es olõq, sis tuu tekk' võrokõisi muidoki veidükese kadõhõs, a tuu-iist oll' meil peris uma ja pall'o parõmba kvaliteediga massintõlgõq olõman joba aastit inne tuud, ku tuu saiq Google'ist latgallasõq ja pall'oq tõsõq väikuq rahvaq.

In recognition of these contributions, the work presented in this thesis — particularly the developments covered in Chapter 4.1 related to the **MTee** system—has received formal recognition. It was honored with the **"Eesti Keeletegu 2021"** award, which acknowledges significant achievements in promoting and advancing Estonian language.[5] Additionally, the University of Tartu awarded the team the **Keeletegu Award** in 2021 for outstanding efforts in combining language and technology through research and development in machine translation and neural speech technologies.[6] These awards reflect the broader societal value of the work, highlighting its role in supporting language preservation and technological development.

In personal interactions, both within Estonia and internationally, the MT systems have been met with warm and positive reactions. Some people have been genuinely touched—occasionally even moved to tears — upon seeing their languages represented in modern technology. Although these impressions come from my own experiences at conferences and language events, I felt it was important to share this side of the story.

[5] https://www.emakeeleselts.ee/koik-teated/selgusid-aasta-keeleteod/

[6] https://ut.ee/et/sisu/aastaauhinnad

While not relying on scientific data, this feedback offers insight into the practical relevance and real-world value of the research. It serves as a reminder that beyond technical metrics and models, the ultimate goal of language technology is to ehlp humans connect and support cultural preservation. For me, this has been the most meaningful part - being able to give something back to the communities.

1.6. Structure of the thesis

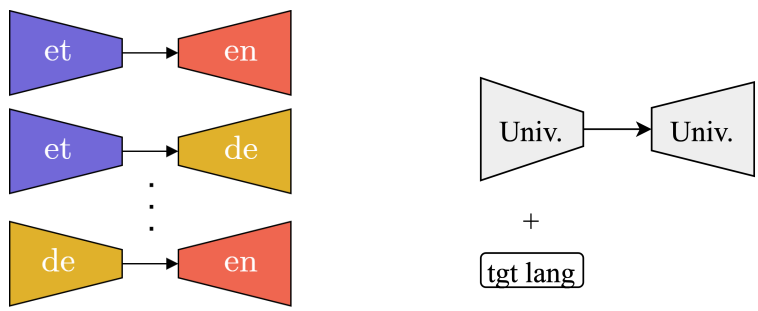
The structure of this dissertation is organized as follows:

1. Chapter 1 provides an introduction to the topic and outlines the research objectives and questions.
2. Chapter 2 introduces core concepts in neural machine translation, with a focus on multilingual models, setting the stage for the contributions presented in subsequent chapters.
3. Chapter 3 presents two papers (contributions I and II) on universal models, investigating parameter-sharing techniques for universal multilingual models, with a focus on improving translation quality using knowledge-sharing and mitigating performance disparities between languages.
4. Chapter 4 presents two papers (contributions III and VI) covering the modular NMT architecture, focusing on all resource scenarios of multilinguality. This chapter proposes a new method of doing NMT and includes two papers exploring its effectiveness for low-resource and zero-resource language pairs.
5. Chapter 5 presents three papers (contributions IV, V, and VII) exploring the use of large pre-trained language models for translation, demonstrating how to fine-tune them for specific language pairs. Three papers in this chapter showcase the significant performance gains achievable with this approach.
6. Chapter 6 concludes the dissertation, summarizing key findings and offering suggestions for future research.

2. BACKGROUND

The background section provides a brief overview of the relevant concepts and our research trajectory over the course of the PhD period. This chapter is split into multiple sections, starting with a short glossary of terms relevant to understanding the rest of the thesis. Followed by deeper definitions for core concepts and, finally, a deeper dive into chapter-specific stories.

This chapter covers the relevant concepts needed to understand the seven articles combined into three chapters. Traditional machine translation models followed a simple architecture: one model for one translation direction, or in other words, one model for one task (see Figure 1.a). In contrast, this thesis focuses on multilingual models, which handle multiple translation directions within a single system, effectively handling multiple tasks at once. The third chapter covers universal neural machine translation models, where a single model is used for multilingual translation (see Figure 1.b). This is followed by the chapter covering the modular approach by Lyu et al. (2020), proposing a model with language-specific (modular) components within a multilingual system. Finally, the last chapter explores the recent trend of leveraging pre-trained language models (Devlin et al., 2019) for improved performance and training efficiency in multilingual NMT.



(a) A set of single directional models (b) Shared many-to-many model

Figure 1: Diagram showing (a) a set of single directional models where no parameter sharing across language pairs is possible, and (b) multilingual many-to-many universally shared model, where only a single encoder and decoder are shared across all language pairs.

2.1. Core NMT Concepts

This section gives a very brief explanation of the terms used in the thesis.

Artificial Neuron is a fundamental computational unit in a neural network that mimics the function of a biological neuron. It receives multiple inputs, applies weights to them, and processes the weighted sum through an activation function to generate an output. This output can then be transmitted to other neurons in

subsequent layers, allowing the network to learn complex patterns and make predictions by training the weights of each connection during training (Goodfellow et al., 2016).

Neural Networks are complex mathematical systems inspired by biological neural networks. They process information through interconnected layers of artificial neurons, enabling them to learn patterns from data and perform various tasks such as classification, prediction, and natural language processing tasks, including translation (Goodfellow et al., 2016).

Encoder, introduced by Cho et al. (2014), is a component that transforms the input text (source language) into a numerical representation (embedding) that captures its meaning. The encoder acts like the first part of the "translator's brain" within a neural machine translation system. It processes the input sentence in the original language and then "analyzes its meaning" by encoding the natural sentence into a numerical form that the AI system can understand. This encoded information is then passed to the decoder to generate the final translation.

Decoder, also introduced as part of the encoder-decoder framework by Cho et al. (2014), functions as the second half of the "translator's brain" within a neural machine translation system. It receives the encoded understanding of the source sentence from the encoder and generates the translated text (target language) word by word. To enhance the translation process, the decoder often uses an attention mechanism (Bahdanau et al., 2014), allowing it to focus on specific parts of the encoded input as it generates each word.

Encoder–Decoder framework, introduced by Cho et al. (2014), is designed to transform an input sequence into an output sequence. In these models, the encoder processes the input sequence and compresses it into a fixed-length vector representation, which captures the essence of the input. The decoder then takes this representation and generates the output sequence, producing one token (word or character) at a time. Here, a 'sequence' typically refers to a series of tokens, such as subword units or words, which together form a sentence. This approach, also known as **sequence-to-sequence (seq2seq) model** and popularized by Sutskever et al. (2014), has revolutionized fields such as machine translation through its ability to handle sequences of varying lengths and complex dependencies.

Beam Search is a heuristic search algorithm used in sequence-to-sequence models (Koehn, 2009), especially during the decoding phase of neural machine translation (NMT) systems. Unlike greedy search, which selects the most likely word at each step without considering alternatives, beam search keeps multiple promising partial hypotheses (beams) and prunes less likely ones. Maintaining and evaluating these hypotheses reduces the risk of committing too early to a suboptimal translation. This approach leads to more fluent and accurate translations. Additionally, beam search is more memory and speed efficient than naive exhaustive search, as it limits the number of hypotheses considered at each step, significantly reducing computational complexity and memory usage.

Attention is a mechanism that allows the decoder to dynamically focus on the most relevant parts of the input sentence as it generates each word of the translation. Introduced by Bahdanau et al. (2014), this mechanism significantly improved translation quality in NMT systems. Instead of relying on a single fixed-length vector representation for the entire input sequence, attention allows the decoder to calculate attention weights, which represent the relative importance of each word in the source sentence for the current translation step. It works like a spotlight, enabling the decoder to better understand the context of the whole input sentence when generating each translated word.

2.1.1. Transformers

Transformer, introduced by Vaswani et al. (2017) in their groundbreaking paper "Attention Is All You Need," ignited a revolution in the field of natural language processing (NLP) and profoundly reshaped the landscape of machine translation. This novel architecture discarded the recurrent components that dominated earlier models and instead relied entirely on a powerful mechanism called self-attention. With this shift, the Transformer achieved significant leaps in translation quality, training efficiency, and the ability to understand complex linguistic relationships and long-term dependencies.

At its core, the Transformer consists of an encoder and a decoder. The encoder processes the input text sequence, employing multiple layers of self-attention to model relationships between input tokens and create a contextualized numerical representation of the entire input. The decoder then generates the translation token-by-token, relying on two key attention mechanisms:

1. **Self-attention within the decoder** This allows the model to consider previously generated words when producing the next token.
2. **Cross-attention to the encoder output** As in earlier attention-based RNN models (Bahdanau et al., 2014), the decoder selectively attends to the encoder's output to focus on the most relevant parts of the input while generating each word. However, Transformers implement this attention differently, using scaled dot-product attention, which enables better parallelization and computational efficiency.

The innovative design of the Transformer brought substantial improvements over its predecessors (Bahdanau et al., 2014):

- Higher Translation Quality: The Transformer consistently outperforms previous approaches (statistical machine translation(SMT), recurrent neural network-based approaches), leading to more fluent, accurate, and natural-sounding translations (Barrault et al., 2019).
- Training Efficiency: Self-attention allows for highly efficient parallelization, accelerating the training process and enabling the model to learn from larger datasets in significantly less time.

- Improved handling of long-range dependencies: The Transformer excels at capturing relationships between inputs that are far apart in a sentence/text, which is vital for providing consistency and contextual awareness.

The impact of the Transformer architecture extends far beyond its initial success in machine translation. Its core principles have diverged into two influential branches within NLP. Firstly, the encoder component's power for understanding language has led to models like BERT by Devlin et al. (2019), revolutionizing natural language understanding (NLU) tasks. Secondly, the decoder component's focus on language generation has powered models like GPT by Radford et al. (2018), driving the development of increasingly sophisticated conversational agents and large language models. This widespread influence underscores the profound and ongoing transformation the Transformer has brought about in our ability to process and interact with language using computers.

2.2. NMT Vocabulary

One of the most fascinating things about language is its infinite potential. New words and phrases are constantly born. Additionally, morphologically rich languages like Estonian or German use a process of compounding to create words with an exponentially increasing number of possibilities. Unfortunately, this conflicts with the limits of computational systems designed to understand and translate text. In NMT, the term "vocabulary" refers to the predefined (limited) set of tokens the model can process, whereas in natural language, vocabulary is inherently open-ended. Tokenization, which involves splitting sentences into smaller units called tokens (such as words), is crucial in this context because it significantly affects how well the translation model can handle natural language vocabulary.

Memory constraints pose a significant challenge in NMT. Early models often relied on fixed vocabularies, typically around 32,000 words. While seemingly large, these proved insufficient to capture the richness of natural language. Increasing vocabulary size helped reduce the number of out-of-vocabulary (OOV) words but led to slower training and larger memory requirements. Conversely, smaller vocabularies improved speed and efficiency but harmed the ability to generalize and translate accurately. However, as long as tokenization was word-based, handling previously unseen words remained a major challenge, as the model could not generate tokens that were not explicitly present in its fixed vocabulary. Finding the right balance was difficult but crucial.

2.2.1. Byte-Pair Encoding (BPE)

To address the vocabulary issue, Sennrich et al. (2016c) introduced Byte-Pair Encoding (BPE). This technique provides a flexible way to represent words, overcoming the problems of fixed vocabularies. Instead of treating words as tokens, BPE breaks words into smaller components called subwords. These subwords can range from individual characters, to common letter combinations or even the most common words.

The BPE algorithm is quite straightforward - let's imagine training a BPE model on Estonian text data. Initially, each word is split into characters. The algorithm iteratively combines the most frequent characters together (for example, the characters $'l'$ and $'e'$ can be combined into one $'le'$ token), repeating the process until the predefined vocabulary size is reached, resulting in a mix of words and subword tokens.

For instance, the phrase "Tere maailm" might be tokenized by BPE as shown in Table:2. The `@@` symbol is a commonly used notation to indicate that the token is part of a subword that continues.

Input	Tere, maailm
Output	Tere, maa@@ ilm

Table 2: Example of BPE tokenization. Here, "Tere" remains unchanged as it's likely frequent, while "maailm" is broken down into "maa@@", which continues on, ("maa" meaning "land") and "ilm" ("ilm" suffix changes the word into "world").

Benefits of BPE for NMT:

1. Eliminates OOV words/tokens: BPE eliminates the issue of unseen words (OOV tokens) by breaking them into subwords present in the training data. The only exceptions are completely unseen characters or rare symbols (e.g., from different encodings, rare/foreign names). To handle these cases, models can use byte-level backoff, where unknown characters are represented using byte-based tokens, ensuring the input remains processable.
2. Customizable vocabulary efficiency: The vocabulary size can be controlled, striking a balance between translation quality and computational efficiency.
3. Vocabulary generalization: BPE enables an open vocabulary approach. Even if a word is not seen during training, BPE can still represent it using known subword units, allowing the model to generate translations to improve model robustness.

While BPE has proven transformative, variations have emerged to offer even greater flexibility, with SentencePiece (SPM) (Kudo et al., 2018b) being the most impactful. Like BPE, SentencePiece segments words into subword units, but it differs in two key aspects. First, SentencePiece does not require pre-tokenization into words, meaning it treats input text as a continuous stream of characters rather

than relying on predefined word boundaries. This makes it better suited for languages without clear word boundaries or with complex morphology. Second, instead of merging the most frequent character pairs iteratively like BPE, SentencePiece uses a statistical unigram model, which selects the most probable subword segmentation rather than purely relying on frequency. These refinements make SentencePiece highly versatile, and it has become a popular choice for tokenization in multilingual NMT systems and large language models.

2.3. Data Concepts

Machine translation is fundamentally driven by two core elements: the data used for training and the model's design/architecture. This section covers basic data-related concepts referenced throughout this thesis. Importantly, the quantity, quality, and characteristics of training data significantly impact the performance of the final system.

Parallel Corpora. A parallel corpus consists of aligned text in two or more languages. For example, an English sentence and its Estonian translation would form one data point, often called a sentence pair.

Language pair. A language pair refers to the specific combination of source and target languages an NMT system translates between (e.g., English-to-Estonian, Estonian-to-French). Language pairs can impact the model performance due to linguistic similarities/dissimilarities between the languages - so some models might require more data to learn a sufficiently good model.

Monolingual corpora. Monolingual corpora contain text in a single language. While they are less crucial for supervised NMT training compared to parallel corpora, they are valuable for supplementary data augmentation techniques. Since monolingual data is easier to find than parallel data, they are often used in machine translation model training processes.

Data Domain. The domain of the training data refers to its topic or field (news, legal documents, technical manuals, etc.). NMT systems perform best when the training data's domain aligns closely with the domain of the text they will be translating for real-world use.

Resource levels of data. Languages or language pairs with extensive, accessible text collections are considered high-resource; for example, English is regarded the highest resource language. It is important to note that there are no universally agreed-upon thresholds for these definitions - different studies use different criteria. For the purpose of this thesis, I define a language pair as low-resource if it contains fewer than 100K parallel sentence pairs, and high-resource if it exceeds 1M parallel sentence pairs. This classification, which is widely accepted in the community, aligns with the guidelines proposed by Joshi et al. (2020). Furthermore, language pairs with fewer than 10K parallel sentences are classified as very low-resource.

Zero-shot. Zero-shot or zero-shot translation in the context of this thesis refers to language pairs without any parallel data between them, but usually, at least some data exists for the language pair. For example, Võro and Finnish do not have parallel data, but models can learn to translate both via generalization when a multilingual model learns to translate Võro $\leftrightarrow$ Estonian and Finnish $\leftrightarrow$ Estonian. Zero-shot does not mean that there are no data, usually at least some monolingual and minimal parallel corpora exists for some language pair.

2.3.1. Data Processing and Filtering

Data processing and filtering are important steps in preparing high-quality parallel corpora for machine translation. In this thesis, we applied standard data filtering techniques commonly used in multilingual NMT, including language identification, deduplication, length-based filtering, and alignment checking to remove noisy or misaligned sentence pairs. These approaches are consistent with the best practices outlined in the WMT 2020 Shared Task on Parallel Corpus Filtering and Alignment (Koehn et al., 2020), which demonstrated the importance of systematic data cleaning for improving model performance. Furthermore, competitive systems such as the WMT20 submissions of OPPO (Shi et al., 2020) highlight the effectiveness of data preprocessing pipelines, where filtering strategies contributed significantly to achieving top results in multiple language pairs.

2.4. Data Augmentation

The availability of large, high-quality datasets is often a limiting factor in NMT, particularly for low-resource languages. Data augmentation techniques help to mitigate that problem by expanding existing training data synthetically, aiming to improve model performance and robustness. The most used data augmentation technique is called back-translation.

Back-translation (BT) by Sennrich et al. (2016a) is a powerful method for generating synthetic parallel data. In back-translation, you usually need two trained models (source $\rightarrow$ target and target $\rightarrow$ source), which can be trained on limited parallel data. Additionally, let's say that you want to improve the source $\rightarrow$ target model, which we call M_{st}, which also requires the source $\rightarrow$ target model, which we call M_{ts}. The algorithm then is very simple yet powerful:

1. Translate from target (monolingual data) to source (called the synthetic source) using the using the M_{ts} model.
2. Create synthetic parallel data: gather all translations paired with the monolingual targets to generate a new synthetical dataset, consisting of pairs (synth source, target).
3. Combine the parallel data with back-translated data, which consists of original parallel data and synthetic parallel data, to train a new (iteration) of model M_{st}.

The benefits are that back-translation increases the training data, which is especially useful in low-resource settings. It also acts as a generalization method, introducing slight imperfections, thus making the trained model less sensitive to noise. Data quality and domain alignment are still important, and often, monolingual in-domain data can be found, so back-translation can be used to make domain-tuned models. One drawback of back-translation is that BT introduces an error propagation risk, meaning that errors in model translations (synthetic sources) can be reinforced in the synthetic data, which might lead to model degradations.

In multilingual models, iterative back-translation offers a distinct advantage: a single unified model can generate synthetic data and subsequently refine its own performance, whereas bilingual setups require two separate models (one for each translation direction). Repeating this process over multiple iterations often yields a better model.

A similar, simpler data augmentation method is forward-translation, where monolingual data is used as the source and the M_{st} model is used to create training data in the format (source, synthetic target), which doesn't require the M_{ts} model. However, it typically results in smaller performance gains compared to back-translation (Burlot et al., 2018). This is because the synthetic target sentences generated by forward-translation can be (much) noisier, and models trained on such data may suffer from reduced performance due to the lack of natural, human-generated target sentences.

2.5. MT Evaluation

Evaluating the quality of machine translation systems is a complex but essential task. Several metrics exist, each with its strengths and limitations. This thesis focuses on two widely used automated surface-based metrics: BLEU and CHRF. Surface-based metrics, like BLEU, evaluate the quality of machine translation by comparing the translated text to one or more reference translations based on n-gram overlap, focusing on the surface forms of words rather than their underlying meaning. Importantly, they are typically computed over an entire corpus rather than on individual sentence pairs, since sentence-level scores can be (very) unreliable. Both are scores from 0 to 1 (or 0 to 100 as percentages), but the scores always depend on test sets, and BLEU scores on different language pairs or domains cannot be compared. This means that, mostly without a baseline to compare to, these numbers (BLEU and CHRF) don't tell much about the quality of the model.

BLEU. (Papineni et al., 2002) measures the similarity between machine generated translations and human-produced reference translations. BLEU calculates precision based on how many n-grams in the machine translation also appear in the reference, also called the n-gram overlap. It includes a brevity penalty to prevent short translations from scoring artificially high. It is easy to calculate and correlates reasonably well with human judgments. Some of its main limitations

are that it looks for exact word matches, penalizes semantically correct but differently phrased translations, and struggles with morphologically rich languages.

CHRF. (Popović, 2015) is a similar metric, comparing a translation to a reference, but operates on the character level. It calculates an F-score (combination of precision and recall) based on character n-gram matches. CHRF is also easy to calculate, handles morphological differences better, and can be tuned to prioritize precision or recall depending on the use case. It also correlates reasonably well with human judgments and correlates slightly better with humans than BLEU. A popular extension of CHRF is CHRF++ (Popović, 2017), which includes both character n-grams and word n-grams in its calculation.

Here are some examples of translations and their scores with popular translation metrics BLEU and ChrF++.

Ex 1 **Source:** Masintõlge on keerukas ja huvitav uurimisvaldkond.

Reference: Machine translation is a complex and interesting research field.

Hypothesis: Machine translation is a complex and interesting field of research.

BLEU: 62.39

ChrF++: 88.21

Comment: The hypothesis matches the reference well, leading to fairly high scores. *research field* and *field of research* have some overlap, which increases the score.

Ex 2 **Source:** Masintõlge on keerukas ja huvitav uurimisvaldkond.

Reference: Machine translation is a complex and interesting research field.

Hypothesis: Machine translation is a sophisticated and fascinating research field.

BLEU: 39.28

ChrF++: 67.31

Comment: The hypothesis has different wording but a similar meaning, resulting in lower scores compared to the previous example. This example demonstrates a key limitation of surface-based metrics like BLEU and ChrF++, which rely on exact word (or character) matches. Although the hypothesis accurately conveys the meaning of the reference, differences in word choice (e.g., “sophisticated” vs. “complex” and “fascinating” vs. “interesting”) result in lower scores. This limitation highlights how surface-based metrics can penalize semantically similar translations that use synonyms or alternate phrasings.

Ex 3 **Source:** Machine translation is a sophisticated and fascinating research field.

Reference: Masintõlge on keerukas ja põnev uurimisvaldkond.

Hypothesis: Masintõlge on keerukas ja põnev uurimisvaldkond.

BLEU: 100.00

ChrF++: 100.00

Comment: The hypothesis perfectly matches the reference, leading to the highest score.

Ex 4 **Source:** Machine translation is a complex and exciting field of research.

Reference: Masintõlge on keerukas ja põnev uurimisvaldkond.

Hypothesis: Masintõlge on keeruline ja huvitav uurimisvaldkond.

BLEU: 19.64

ChrF++: 73.03

Comment: The hypothesis has some differences in wording compared to the reference, resulting in low scores for both BLEU and ChrF++, but the reference and hypothesis have the same meaning. This example highlights the weakness of surface-based metrics, which fail to consider synonyms. This example also shows how BLEU score is very low (lowest of the 4 examples), but ChrF++, which scores on the character level, does not penalize as much, leading to fairly high ChrF++ score.

Limitations of surface-based metrics. are that there has not been a single best metric so far in translation, and all can be situational. Beyond surface-level metrics, researchers also use other metrics, such as COMET, which goes beyond simple word matching to assess factors like meaning preservation and fluency. COMET (Rei et al., 2020) (Crosslingual Optimized Metric for Evaluation of Translation) is a neural-based metric that evaluates translation quality by considering semantic adequacy (that is, how well the translation conveys the meaning of the source text) and fluency (that is, how natural and grammatically correct the output is). Human judgments, while being the best metrics for assessing translation quality, are too expensive to use and do not scale at all.

The original BLEU and CHRF papers report high correlations with human judgments. However, Freitag et al. (2023) showed that these correlations can vary depending on language pairs and domains and that other metrics may perform (much) better, such as the abovementioned COMET. However, BLEU and CHRF remain widely used because of their simplicity, efficiency, and ease of interpretation, especially when comparing the system performance across multiple experiments.

In this thesis, COMET will not be used, as it was only recently introduced during the course of this PhD and had limited language support at the time, lacking models for many of the lower-resourced Finno-Ugric languages relevant to this research.

2.6. Transfer Learning

Traditional approaches to training NMT systems often require large amounts of parallel data for each language pair. This data scarcity poses a significant challenge, limiting the development of high-quality translation systems for many languages. Transfer learning (Goodfellow et al., 2016) offers a powerful solution by leveraging knowledge acquired from previous tasks or (trained) models. This approach reduces the need for extensive parallel data, accelerates training, and boosts translation quality. Transfer learning makes NMT more accessible and feasible for a wider range of languages and data scenarios. This thesis focuses on leveraging transfer learning methods to improve the performance of NMT in low-resource settings.

2.6.1. Finetuning

Finetuning (Erhan et al., 2010) involves taking a (pre-)trained model and updating specific portions of its parameters using task-specific data to tune the model to the target task. Because this process leverages the knowledge and representations learned during pre-training, fine-tuning is a form of transfer learning. It allows the model to adapt to new tasks or domains with relatively little additional data, which is especially beneficial in low-resource scenarios. There are multiple ways to apply finetuning in NMT:

- Domain Adaptation: A common scenario involves finetuning a general-purpose NMT model on in-domain data (e.g., technical manuals, legal documents, spoken text) (Luong et al., 2015). This allows the model to understand and model the target domain's specialized terminology and stylistic conventions, leading to more accurate translations within that domain.
- Finetuning multilingual models for specific tasks: Multilingual NMT models, pre-trained on data from multiple languages, provide a strong basis for language-pair-specific training. Finetuning on even a relatively small amount of parallel data for a new language pair often allows the model to adapt its parameters for the new task, improving its ability to translate specifically between those languages, as shown by Johnson et al. (2017).
- Handling Low-Resource Scenarios: Finetuning is crucial for low-resource languages. Starting with a model pre-trained on data from related or higher-resource languages provides a valuable knowledge base. Finetuning with limited in-domain data can significantly boost performance compared to training from scratch. Research has actually shown that any pre-trained model as a starting point for training can be beneficial, as shown by Kocmi et al. (2018).

Finetuning usually follows a simple routine by continuing to train from some state of an existing model - usually, it is the last or best-performing checkpoint of a trained model. This allows you to resume training, adapting the model to

the new task or domain. It's crucial to carefully manage hyperparameters like learning rates during this continuation phase, often using a lower learning rate to prevent overwriting pre-existing knowledge (also known as the catastrophic forgetting problem (Kirkpatrick et al., 2016)) while still allowing for adaptation. There are many distinct ways of doing fine-tuning, such as selecting different datasets, doing iterative back-translation and fine-tuning, gradual unfreezing of (parts) the model, or selective layer updating.

The effectiveness of finetuning for new language pairs often increases when there is lexical similarity (shared vocabulary) between the languages involved. Closely related languages frequently have overlapping words and grammatical structures. This enables smoother knowledge transfer during finetuning. Multilingual models pre-trained on data from multiple languages develop representations that reflect these cross-lingual relationships. When the target language of finetuning is similar to those in the pre-training data, the model benefits from shared representations and an advantageous starting point. This facilitates faster learning, reduces data requirements, and often leads to overall better translation performance. This phenomenon is crucial for this thesis.

Lexical similarity plays a key role in knowledge transfer within multilingual models. A model trained on multiple languages with a shared vocabulary develops internal representations capable of capturing these commonalities. This ability is called parameter sharing, which leverages these shared representations by reusing model components across languages. Parameter sharing leads to zero-shot translation, where a multilingual model can surprisingly translate between language pairs which were not seen during training Johnson et al. (2017). This arises from the model learning to map similar concepts across languages, even if those pairs haven't been directly linked. For example, a multilingual model trained on English $\leftrightarrow$ Estonian and English $\leftrightarrow$ Finnish might translate Estonian-Finnish despite never seeing that specific pair at training time. While beneficial, parameter sharing has its complexities. Multilingual models need a way to distinguish between languages. Typically, special tokens are added at the beginning of the input to signal the target language. However, sharing can also cause problems, such as the model mistakenly producing tokens of the wrong target language (Chen et al., 2023; Zhang et al., 2020), usually confused in favor of more common languages like English - a challenge experienced especially in the early systems. Ongoing research focuses on techniques to mitigate this, such as tailored modular components for specific languages within the larger multilingual architecture.

2.7. Universal NMT Overview

The third chapter focuses on Universal Neural Machine Translation models (Gu et al., 2018), which have been a standard approach for multilingual machine translation after the Recurrent Neural Networks(RNN-s, Cho et al. (2014)) replaced statistical methods. Universal NMT models refer to a single model's ability to

handle multiple language translations by sharing parameters, and they can also be called universally shared NMT models or shared multilingual NMT models. To be more concise, the core concept around Universal NMT is that a single model with shared encoder-decoder parameters, RNN or transformer-based, is used to train an NMT model with multiple input and output languages. Input and output languages can differ. The universal system has its advantages and disadvantages. The advantages of the Universal NMT model are as follows:

- Parameter sharing - Multilingual models share parameters for all the different languages, which can lead to more efficient use of resources and lower the number of models needed for multiple language pairs. Ideally, a single model is required for a system of N languages when using the universal approach, while N*(N-1) models are required for single-directional models.
- Eliminates Pivoting - The previous standard for multilingual systems before universal models were pivoting (Kauers et al., 2002), meaning that low-resourced translation directions used more popular models for translation; for example, Estonian into German was done by pivoting over English, so Estonian $\rightarrow$ English $\rightarrow$ German. The pivoting approach has two main problems - (1) it requires two translations to be done, which is inefficient, and (2) it introduces an error propagation problem, where 1st translation's problems are propagated to the 2nd translation.
- Low-resource performance - Multilingual models can improve translation quality for low-resource languages by leveraging knowledge from high-resource languages (Aharoni et al., 2019). This transfer learning can help improve the performance of languages with limited parallel data.
- Zero-shot ability - Multilingual models can perform zero-shot translation (Johnson et al., 2017), meaning they can translate between language pairs not seen during training. This can be beneficial for language pairs with little or no parallel data available.

The disadvantages of the Universal NMT models compared to single-directional models are as follows:

- Complexity: Multilingual models can be more complex than unidirectional models as they need to handle multiple languages and their specific characteristics. This can lead to increased training time and computational requirements. For example, one problem is with vocabulary in different alphabets like Latin, Cyrillic, and Chinese in one model - these languages make sharing the vocabulary a hard task because the model needs to generalize. Additionally, it is harder to fine-tune multilingual models, especially for single directions or domains, as this can introduce undesirable results like "catastrophic forgetting" (Kirkpatrick et al., 2016).
- Interference: When training a multilingual model, there is a risk of interference between languages, which can negatively affect translation quality. This can happen when the model confuses the features of different

languages or when the optimization process prioritizes one language over another. A problem in multilingual systems is that the translation is in the wrong language. This issue has been widely observed in multilingual NMT systems, where overlapping linguistic representations can cause the model to produce output in an unintended language (Chen et al., 2023; Dabre et al., 2020; Zhang et al., 2020).

- Scalability: As the number of languages increases, maintaining model performance becomes more challenging. When the number of languages grows but the model size stays the same, capacity limitations can become an issue (Arivazhagan et al., 2019). With limited capacity, translation quality may drop, especially for low-resource languages. One way to address this is by increasing the model size, but this also raises training costs and resource requirements. Finding the right balance between model capacity and the number of parameters is challenging, as a model that is too small may underperform, while an unnecessarily large model wastes resources.
- Dataset imbalances: Data distribution is often problematic when thinking about training multilingual systems. Usually, language pairs, with one side being English, have more abundant data compared to non-English language pairs. For example, English-Estonian has 60M sentence pairs in Opus (Tiedemann, 2012) while Arabic-Estonian has 5.5M sentence pairs.[1]

Universally shared multilingual NMT models have their advantages and disadvantages, but given resource constraints, both for training and inference, then this approach is optimal for us. The main reason is that Estonian is usually considered a low-resource language (except for English-Estonian), making knowledge transfer in multilingual systems a desirable property for other translation directions, such as Arabic-Estonian. Additionally, given that these multilingual models are more complex to train, it has been an interesting research task to investigate.

The research described in this dissertation focuses on multilingual models. The first multilingual models were available for public usage on TartuNLP's website[2]. Chapter 3 contains the first work of our group on training universally shared NMT models. This chapter covers two approaches: (1) a multilingual multi-domain shared task paper for WMT19[3] (Contribution I) and (2) a multilingual NMT for extremely low-resource Finno-Ugric languages (Contribution II).

One option for improving multilingual models is the use of source factors (Koehn et al., 2007a; Sennrich et al., 2016b), which provide additional information about input tokens. They were introduced to NMT by Sennrich et al. (2016b) and initially contained linguistic details such as part-of-speech tags, lemmas, subword information, morphological features, or syntactic dependencies. These factors are appended to token embeddings, allowing the model to capture linguistic

[1] According to: https://opus.nlpl.eu/

[2] https://translate.ut.ee

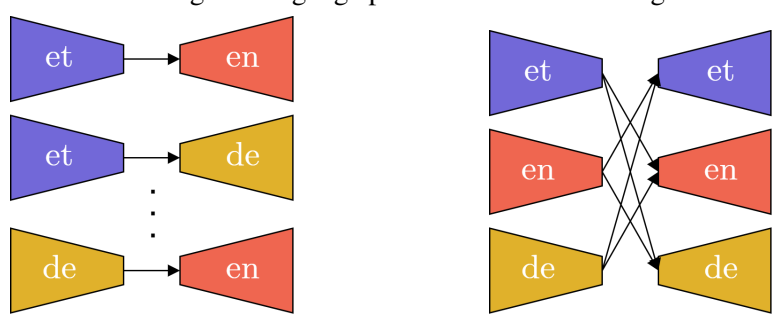
[3] https://statmt.org/wmt19/

structures that might otherwise be lost. In multilingual settings, they can help handle word variations in morphologically rich languages and reduce cases where the model outputs the wrong language in many-to-many setups. However, a drawback is that they rely on high-quality linguistic annotations that may not always be available, especially for low-resource languages. In this thesis, instead of using source factors for morphology, we use them to explicitly indicate the desired output language, helping the model produce translations in the correct target language.

2.8. Modular NMT

The motivation for the fourth chapter stems from the growing significance of multilingual neural machine translation (NMT) in tackling the challenges associated with low-resource languages and high-resource translation scenarios. Universal NMT models (see Figure:1b) and traditional single-directional NMT (Figure:2a) models each have their strengths and weaknesses, and striking the right balance between sharing information across languages might be crucial for success. Arivazhagan et al. (2019) showed that bilingual models often outperform universal models in high-resource settings, while universal models excel in very multilingual scenarios. This motivates our exploration of modular approaches that aim to capture the best of both worlds.

To address these challenges, we investigate various knowledge-sharing approaches in this chapter. Firstly, we investigate the Modular (shown in Figure:2b) approach, which combines language-specific encoders/decoders in one system; then, we propose a novel approach that combines universal, language-group-specific, and language-specific modules to overcome the limitations of both universal and language-specific models. By exploring different granularities of the Modular architecture and layer sharing, we aim to enhance the overall translation quality for a diverse range of language pairs and resource settings.



(a) A set of single directional models (b) Multilingual modular model

Figure 2: Diagram showing (a) a set of single directional models where no parameter sharing across language pairs is possible, and (b) multilingual modular model, where encoders and decoders are shared across all language pairs.

Our experiments on multilingual datasets demonstrate that the proposed models outperform purely universal and language-specific approaches. This chapter presents the motivation, methodology, and results of our investigation, offering insights into the potential benefits of sharing layers among all languages to boost translation quality in multilingual neural machine translation models.

2.9. Using pre-trained multilingual models

The overall trend of NLP has been the emergence of pre-trained large language models (LLM-s), often called foundation models like GPT-3 (Brown et al., 2020), GPT-4 (OpenAI, 2023), M2M (Fan et al., 2021), NLLB (NLLB Team et al., 2022), BART (Lewis et al., 2020), and so on. So far, we covered the effect of sharing layers and knowledge inside the models but still trained them from scratch. In section 5, however, we want to go a step further and investigate the effective usage of large, pre-trained models for multilingual machine translation. We will focus on the M2M model and the NLLB model. The main motivation is to lower training times, improve the qualitative performance of the model, and investigate the ability to teach new languages for translation to the pre-trained model.

This chapter's main investigation is done for low-resource Finno-Ugric languages, which is an interesting problem due to very scarce training data. Additionally, most Finno-Ugric languages are missing from the models mentioned above, so appending them to the models is an interesting research question. Most of these languages are not present in large pre-trained models right now. Meaning that these new capabilities have to be introduced to these models. The main motivation for all of these papers is to improve overall Finno-Ugric machine translation by any means. The work in the final chapter has been mostly focused on investigating general research developments (large pre-trained models), transfer learning, and investigating various subsets of the full dataset for smaller, more specific experimental results. All of the papers in this chapter have similar limitations:

- Very limited parallel data - many directions with 0 data, few with just hundreds of sentences
- The domain of the data is not optimal for translation - lots of parallel resources are from Biblical origins, which leads to domain mismatch, as the translation system is used for general/news text mostly.
- Very limited monolingual data - a lot of work has been done for languages with little parallel data but lots of monolingual data. My work focuses on challenges with almost no monolingual data.

3. MULTILINGUAL SHARED NMT MODELS

The field of multilingual neural machine translation (NMT) offers powerful solutions, particularly for low-resource languages, which often lack high-quality translation systems. However, challenges remain in creating universal models that effectively balance high- and low-resource performance. This section delves into two key contributions of this thesis. Firstly, we present multilingual NMT systems that offer valuable insights into the complexities and limitations of balancing performance across a diverse set of languages. Secondly, we investigate methods like synthetic data generation (back-translation) and fine-tuning to optimize translation quality. Our experimental results and open-sourced models provide a strong baseline for future work for the later two chapters.

3.1. University of Tartu's Multilingual Multi-domain WMT19 News Translation Shared Task Submission

This paper presents the University of Tartu's WMT'19[1] news translation submission (Tättar et al., 2019), exploring the potential of a multilingual, multi-domain NMT model in a competitive shared task. Our approach involved a multilingual, multi-domain NMT model trained on a substantial dataset (see Table:3). We researched fine-tuning a jointly-trained multilingual model for specific language pairs within the competition's constrained shared task. Our architecture utilized the self-attention encoder-decoder framework.

	CZ-EN	DE-EN	DE-FR	EN-FI	EN-LT	TOTAL
NEWS	253	5985	4372	2656	1803	17352
OFF	11462	1798	1687	1726	615	17288
SUBS	37251	-	-	-	-	37251
OTHER	10932	34458	7585	4013	1291	58279
TOTAL	62180	42241	13644	8395	3709	130170

Table 3: Dataset sizes (number of parallel sentences) after filtering in thousands(K). **NEWS** contained Rapid2019, Rapid2016, EESC, dev datasets from previous years, EMEA2016, ECB2017, news (from CzEng) and News-commentary corpuses. **OFF** comprised parts of the CzEng corpus and Europarl. **SUBS** were from OpenSubtitles. **OTHER** was anything else.

Our model utilized all available parallel data the WMT'19 constrained task provided, resulting in a diverse dataset of 130.17M sentence pairs (see Table:3 for a detailed breakdown). To further enhance training, we utilized back-translation to create synthetic parallel data.

Data preparation involved downloading, cleaning, and filtering the data using standard techniques like length-based filtering and language identification. We

[1] https://statmt.org/wmt19/

then applied tokenization using a shared byte-pair encoding (BPE) model, trained on a large multilingual corpus with balanced representation of high- and low-resource languages.

The final fine-tuning step consisted of two parts: back-translation, where the shared multilingual pre-trained baseline model was used to translate the monolingual data for itself, resulting in synthetic training pairs (translated source, original monolingual target). Secondly, the model was fine-tuned on data consisting of in-domain synthetic news data to get a fine-tuned model, which is fine-tuned for language and the domain - this whole process is called the back-translation method.

	Baseline	Fine-tune
EN-CS	22.8	-
DE-EN	29.9	-
EN-DE	39.6	-
DE-FR	32.4	30.7
EN-FI	18.6	-
EN-LT	12.7	-
FI-EN	22.1	24.8
FR-DE	25.9	-
LT-EN	24.5	25.3

Table 4: Results of the multilingual baseline model and the fine-tuned model.

The results of the paper are presented in Table:4. The baseline model had quite modest results, but it was the only multilingual submission in 2019 WMT shared translation task (Barrault et al., 2019) competing with state-of-the-art language-specific models. The fine-tuning model, however, broke the model and had one critical error - it started generating English output only, such that the EN $\rightarrow$ LT model, for example, produced mostly English output. There could be multiple reasons for that phenomenon and we suggested that the problems might be in the training data - the filtering didn't filter out sentence pairs that had sentences in the wrong language. The wrong language problem was even worse for synthetic data because the data in wrong language errors were compounded by having training samples consist of the wrong synthetic source - target sentence pairs.

In conclusion, the paper described the multilingual multi-domain neural machine translation approach that can be trained on a mixture of different language pairs and text domains. The results are modest, primarily due to the challenges in fine-tuning and error propagation.

It's important to note that our system faced strong competition in the WMT'19 task from established, language-specific models. Additionally, as one of the earliest multilingual Transformer-based submissions, our work highlights the challenges and potential of these approaches in a competitive setting. Overall, we ranked close to the bottom, beating just a few online systems.

Reflecting on this research, it provided valuable insights and taught us critical lessons for future endeavors and applications.

- One hypothesis is that the main issue with fine-tuning stemmed from not freezing parts of the model, especially the source-factor embeddings. This might have introduced undesirable effects, or perhaps we did not use optimal hyperparameters, leading to complications during fine-tuning. At that time, the process of multilingual fine-tuning was still an open research question.
- Another potential issue could have been checkpoint ensembling, as done by Sennrich et al. (2017), which might have adversely affected the model's performance through suboptimal averaging of weights (or source factors). The primary problem appeared to be the model's confusion regarding target languages. Additionally, data imbalance might have made the issue worse. A possible solution could have involved fixing certain parameters during fine-tuning, such as the source factors, to stabilize performance.
- The tokenization process using MosesTokenizer and truecasing might have also been suboptimal. Removing these steps and adding punctuation normalization could have improved our results—lessons we adopted in our subsequent systems.

Based on what we learned from this research, we used more balanced datasets during fine-tuning to include all languages. We also tried partial parameter freezing for the factor embeddings while fine-tuning but did not observe any clear benefits. However, when we switched to the widely used Fairseq library and abandoned the source factors entirely, we no longer experienced the issues mentioned above.

3.2. Extremely low-resource machine translation for closely related languages

This paper covers an effective approach to achieving good very low-resource neural machine translation (NMT) quality by leveraging multilingually trained universal (shared) models and synthetic parallel corpora creation using the back-translation method. The work focuses on closely related languages from the Finno-Ugric language family - two high-resource languages - Estonian and Finnish, and three very low-resource languages, Võro, North Sámi, and South Sámi. The main contributions of this paper are (1) experimental results and baseline scores for new language pairs, some of which never had a proper MT system before or had very poorly performing systems; (2) additional data collected for the five languages; (3) open-sourced models and freely available models hosted by the University of Tartu Language Technology group.

Language pair	Before cleaning	After cleaning	Eliminated
ET-FI	3 566 826	2 646 922	919 904
ET-VRO	30 816	30 502	314
FI-SME	109 852	35 426	74 426
FI-SMA	3098	2895	203
SME-SMA	23 746	21 557	2189
Overall	3 734 338	2 737 302	997 036

Table 5: Parallel data sets (in sentence pairs). ET - Estonian, FI - Finnish, VRO - Võro, SME - North Sámi, SMA - South Sámi. Further details can be found in the paper (Tars et al. (2021)).

The dataset used for the paper consists of parallel and monolingual data for the five languages, with Estonian-Finnish being the high-resource language pair; all other language pairs were low, extremely low-resource, or didn't have data at all, meaning that we also explored the zero-shot capabilities with the model. The parallel data sizes are given in Table:5, where in total, 2.74M unique sentence pairs are used for training the 5-lingual model after the cleaning, and Et-Fi makes up around 97% of that data. The dataset is doubled during the training because we train in both directions - resulting in a total of 5.48M training samples.

We unified various data sources and formats in the preprocessing stage to create a cohesive data set. We performed a uniqueness check to eliminate repeated sentence pairs, significantly reducing the number of sentence pairs for the Finnish-North Sámi language pair. Monolingual data preprocessing involved extracting sentences from texts, removing empty lines, and addressing alignment issues.

Regarding monolingual data, we collected two sets of data separately, with Estonian and Finnish having the most available data. We used the down-sampling technique, whereby only a subset of the large corpora (ET,FI) was used, to balance the amount of data for low-resource languages, making Võro the most prominent language in the data set. This down-sampling is a form of data sampling, commonly applied in multilingual NMT to mitigate data imbalance issues and ensure more balanced training across languages (Arivazhagan et al., 2019). Although the amount of data for North and South Sámi was still relatively low, a bigger monolingual low-resource dataset was an improvement because this enables more extensive synthetic corpora creation via back-translation. The dataset is described in Table:6.

Language	First set	Second set	All
Estonian (ET)	100 000	25 000	125 000
Finnish (FI)	100 000	25 000	125 000
Võro (VRO)	162 807	5290	168 097
North Sámi (SME)	33 964	6057	40 021
South Sámi (SMA)	55 088	5377	60 465

Table 6: Monolingual data sets after preliminary cleaning (in sentences).

We used the Sockeye framework (Hieber et al., 2017) for model development with the source-side factors, where the target language token was added to each input token. During training, we used the shared vocabulary approach, tokenized by the byte-pair encoding algorithm using the SentencePiece framework (Kudo et al., 2018a). The tokenization model was separately trained on all the training data merged together. Our experiments consisted of developing a multilingual baseline model, which had five source languages and five target languages, totaling 20 translation directions. Implementation details can be found in the paper.

After the baselines were trained, we did back-translation and transfer learning experiments. We carried out two separate synthetic parallel dataset creation iterations via the back-translation approach; we also trained and fine-tuned models on both iterations. We experimented with fine-tuning the multilingual baseline model on ET-VRO parallel data and a transfer learning experiment, initializing the ET-VRO model with ET-FI baseline model weights.

The results show that multilingual NMT with shared encoders and decoders works well for very low-resource language translation visible in Table:7, where Multilingual outperforms Baselines by 7.6 BLEU points. Using back-translation for low-resource NMT is vital for achieving the best results, which can be further improved by transfer learning and fine-tuning - with two back-translation iterations and training, we can achieve +1.6 BLEU points or +9.2 improvement over the baseline approach. The dataset augmentation approach shows how we can improve low-resource translation further via back-translation while not losing quality on the high-resource (ET - FI language pair).

We additionally present the results for transfer learning and fine-tuning approaches in a single translation direction, as shown in Table:8. The baseline Estonian-Võro model achieved a BLEU score of 14.6. When utilizing Estonian-Finnish model weights to train the Estonian-Võro model, the score improved to 26.5 BLEU. The multilingual model fine-tuned on Estonian-Võro achieved the highest score of 27.6 BLEU. While this approach yielded the best results in terms of BLEU score, the slightly lower-performing multilingual model (with BLEU 26.2) offers the advantage of being a single system capable of handling all 20 language directions, providing greater resource efficiency.

Overall, the paper exceeded our expectations with the quality of the translations and we were happy with the results, receiving lots of feedback from the community who used the system.

This paper concludes with a clear path forward, laying the foundation for multiple follow-up papers, which will be covered in the following chapters. The main future goals stemming from the paper were - (1) work with more Finno-Ugric languages and add unrelated high-resource languages that are geographically close, such as German, English, Norwegian or Swedish; (2) improve multilingual fine-tuning and find more parallel and monolingual data for the languages studied; (3) use better initializations like any pre-trained models as a starting point.

Model	ET-FI	FI-ET	ET-VRO	VRO-ET	SME-SMA	SMA-SME	$BLEU_{low}$
Baselines	32.0	29.4	14.6	17.5	8.3	9.1	14.6
Multilingual (ML)	30.9	29.5	23.8	29.6	19.8	19.8	22.2
+ BT1	**32.4**	29.9	25.2	29.4	20.3	20.0	23.0
+ BT1(*)	30.1	29.1	24.5	30.3	21.4	20.0	23.3
+ BT2	31.5	**30.2**	26.0	31.0	20.3	**21.0**	23.7
+ BT1 + BT2(*)	31.3	29.6	**26.2**	31.3	**21.6**	20.7	**23.8**

Table 7: BLEU scores. (*) - trained without pre-trained weights, *BT* - back-translation data set, *FT* - forward-translation data set, $BLEU_{low}$ - average BLEU score on low-resource language pairs (excluding ET-FI and FI-ET), **bold** - best BLEU score for a language pair. Check the original paper for a complete table.

Model	BLEU	CHRF
Estonian-Võro baseline	14.6	0.393
Estonian-Võro on Estonian-Finnish model weights	26.5	0.540
Multilingual model fine-tuned on Estonian-Võro	27.6	0.563

Table 8: BLEU and CHRF scores for transfer learning and fine-tuning experiments.

4. MODULAR NMT

Multilingual neural machine translation (NMT) offers the option of translation systems to handle multiple languages seamlessly. However, finding the optimal balance between knowledge sharing and specialization across languages remains a challenge. Universal models, where a single encoder and decoder are used for all languages, often lead to strong performance on high-resource language pairs but struggle with low-resource scenarios. Conversely, completely modular models, where each language has its own encoder-decoder pair, can excel in high-resource settings but may have difficulty leveraging knowledge from related languages to improve low-resource translation.

Building upon the universal NMT architecture explored in the previous chapter, this chapter examines optimal architectures with higher multilingual capacity that are both efficient and practical for deployment and scalability across diverse language pairs. We propose innovative approaches that strategically combine elements of universal and modular models. In the first contribution, the key focus is strictly on modular language-specific encoders-decoders, but they are jointly trained. In the second contribution, the key focus shifts to the encoder side by investigating different encoder layer-sharing methods while keeping the decoders language-specific. We introduce a tiered architecture that combines language-group-specific modules with universal knowledge sharing. Additionally, we thoroughly explore the impact of sharing the upper layers of the encoder between all languages.

Our experimental results offer valuable insights. We demonstrate that our proposed models outperform both purely universal and fully modular baselines, boosting BLEU scores in any resource scenario. This work contributes a promising step toward finding the right balance between knowledge sharing and specialization in multilingual NMT systems.

4.1. Open and Competitive Multilingual Neural Machine Translation in Production

This paper was part of a bigger project created to improve Estonian-centric machine translation. The project addressed six translation directions (Estonian $\leftrightarrow$ English, Estonian $\leftrightarrow$ German, Estonian $\leftrightarrow$ Russian) across five domains: general, legal, military, crisis texts, and finally, spoken language. The project results include openly distributed parallel and monolingual corpora for the relevant languages, open-source neural machine translation systems trained on them, and new public evaluation benchmarks.

Acknowledging the limitations of universal NMT architectures, particularly in handling morphologically rich languages like Estonian, we investigates the potential alternative - Modular architecture, with their language-specific encoders and decoders, offer the potential to leverage linguistic knowledge better and specialize

for particular domains or language pairs. This work aims to thoroughly evaluate the performance of modular models against the current state-of-the-art available models, with the goal of providing a new optimal MT model for Estonian.

Table 9: Parallel data split size. K means thousands, and M means millions of sentences per domain.

	General				**Crisis**		
	Train	**Valid**	**Test**		**Train**	**Valid**	**Test**
ET-DE	9.3M	1.5K	1.5K	ET-DE	21	0.4K	0.5K
ET-EN	20.4M	1.5K	1.5K	ET-EN	50K	0.7K	1.3K
ET-RU	5.6M	1.5K	1.5K	ET-RU	50K	0.9K	1K
	Legal				**Military**		
	Train	**Valid**	**Test**		**Train**	**Valid**	**Test**
ET-DE	3M	0.9K	1K	ET-DE	0.11M	0.9K	0.9K
ET-EN	3.3M	1.1K	2K	ET-EN	0.17M	0.9K	0.9K
ET-RU	50K	0.5K	1K	ET-RU	50K	0.9K	0.9K

For data collection, we collected all available public data for this project from open sources. Donors and industry partners donated some data for this project. We pre-processed the monolingual and parallel corpora by applying various tests, rules, and criteria to filter the data and ensure that only high-quality data is retained. After filtering, we also normalized the punctuation and whitespace in the data. The number of train, validation and test set sizes per domain and language pair is shown in Table:9. The table represents the amount of data left after filtering. The monolingual dataset sizes are shown in Table:10.

We compared two architectures: a set of single-directional models and a multilingual model with language-specific encoders and decoders, also known as a modular or modularized model. We found that the multilingual modular model outperforms the universal encoder-decoder models in terms of translation quality and is smaller and more convenient to deploy than a set of multiple single-directional models.

We used a modified Fairseq framework for our experiments [1] and models that follow the Transformer architecture base model (check paper for more details on parameters (Tättar et al. (2022))). Embedding layers are shared between the encoder and decoder for the same language. We trained our models in two phases, the first being the pre-training phase, where the training dataset consisted of parallel sentences and synthetic data, including back-translated sentences with a prepended back-translation token to separate the back-translated data from the rest of the data. The second phase is fine-tuning to obtain domain-specific models. Each domain-specific model builds upon the initial general model and continues on domain-specific data.

[1] https://github.com/TartuNLP/fairseq

Table 10: Monolingual data size in thousands (K) or millions (M) of sentences per domain.

	General	Military	Legal	Crisis
ET	50M	0.9M	0.5M	0.6M
EN	48.9M	1.5M	0.3M	10M
DE	49.3M	130K	0.6M	3.4M
RU	49.6M	8K	5.4M	142K

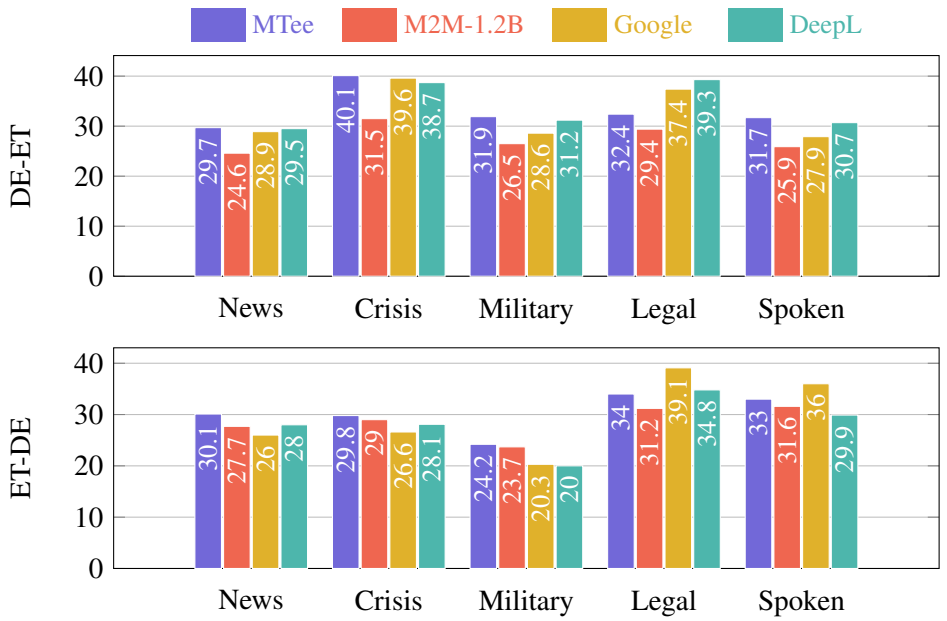


Figure 3: German↔Estonian translation BLEU scores. MTee refers to our model developed in the paper.

We also developed a domain detection model, which is a fine-tuned version of XLM-Roberta (Conneau et al. (2020)). Our domain detection model is designed to default to the general domain in case of uncertainty, increasing the model's precision for the crisis, legal and military domains with a trade-off of lower precision in the general domain. Synthetic parallel data was generated from monolingual corpora using back-translation. A modular model trained on parallel data and beam search with a beam width of 2 (normally for translation it is 4 to 6, lower value is used to speed up translations) was used to translate the corpora, resulting in 50 million new sentence pairs per translation direction used to train the final models.

We evaluated our MT systems using BLEU and CHRF metrics. Comparisons included our general and domain-tuned models (referred to as MTee), widely used online systems (Google Translate, DeepL), and the M2M multilingual model with

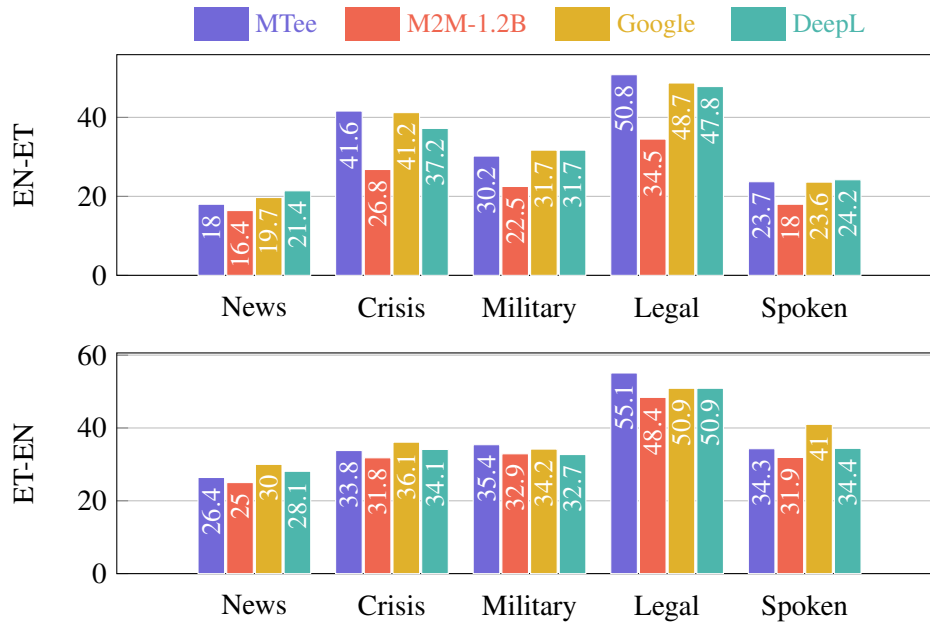


Figure 4: English↔Estonian translation BLEU scores.

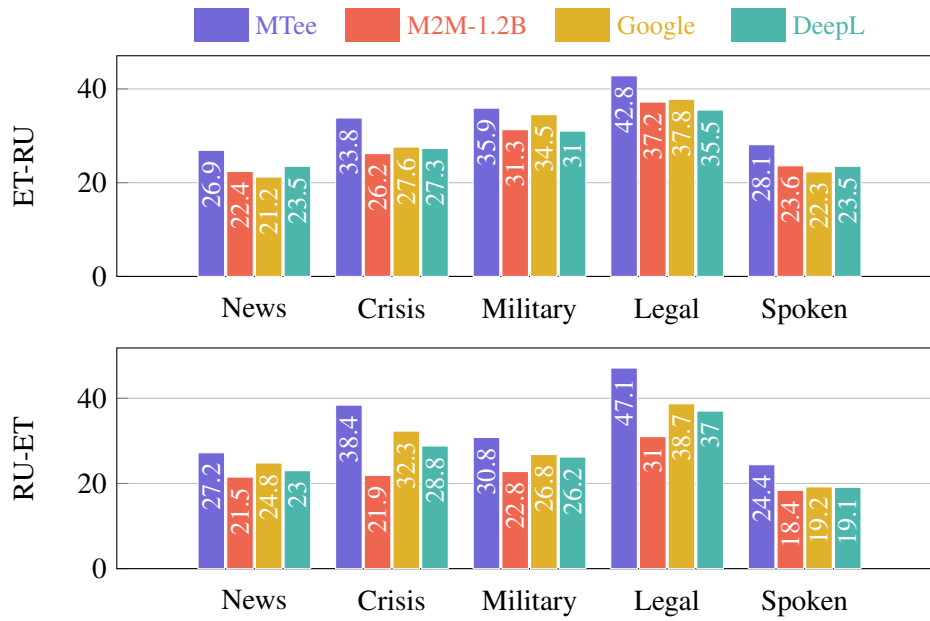


Figure 5: Russian↔Estonian translation BLEU scores.

1.2B parameters (Fan et al., 2021). Figures 3, 4, 5 illustrate the detailed results.

The highest overall gains over open-source and online systems were observed in Estonian → Russian translation, where our models achieved top scores on all ten benchmarks, as shown in Table:5. We also achieved outstanding results

in Estonian-German translation (Table:3), surpassing all benchmarks except in the legal domain and the EN → DE spoken benchmarks. For English → Estonian news translation shown in Table:4, our models underperformed compared to Google and DeepL. However, our domain-tuned models showed comparable or better performance and clearly outperformed other MT systems in the legal domain. It is worth mentioning that compared to the popular open-source M2M-1.2B model, we achieved better performance on all 30 benchmarks.

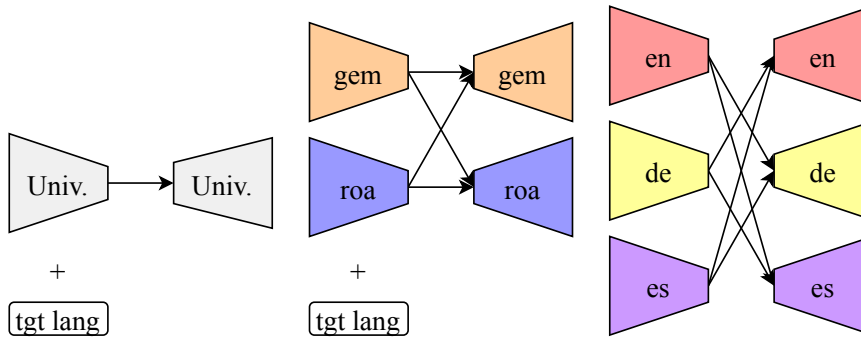
While this work primarily focused on text-based machine translation, we integrated existing Estonian speech recognition models with our MT systems, similar to the approach in (Olev et al., 2022). The main focus was ensuring compatibility between high-quality ASR outputs and our translation models, rather than developing specialized speech translation models from scratch. Dedicated speech translation models are an active area of research, with efforts such as the IWSLT conference[2] providing benchmarks for advancements in the field, including an Estonian-English shared task in 2025.

This project successfully developed competitive Estonian-centric open-source Neural Machine Translation (NMT) systems, significantly enhancing translation quality across several language pairs. Our systems generally outperformed existing freely available or open-source systems, with the best gains observed in the Estonian ↔ Russian and Estonian ↔ German language pairs. The English ↔ Estonian results showed potential for further improvement. The release of parallel and monolingual datasets and the open-source deployment code constitute a substantial contribution to the Estonian language community.

4.2. Multilingual Neural Machine Translation With the Right Amount of Sharing

Multilingual neural machine translation (NMT) has gained significant attention in recent years, particularly for zero- and few-shot translation scenarios - so far, this thesis has covered two approaches - the universal approach (Fig:6a) and the Modular approach (Fig:6c). Both come with trade-offs. A universal single encoder-decoder can be simpler to train and deploy while offering decent generalization for low-resource languages, yet it may run into model capacity problems for high-resource ones. In contrast, fully modular architectures reduce cross-lingual interference and often yield better performance on high-resource languages, but they are more complex to set up and can fail to generalize adequately to lower-resource settings. The results to back these claims up are in Table:12. In practice, the best solution likely lies somewhere in between, where some parameters are shared to leverage broader data, while specialized modules address particular languages or groups. To address these issues, we investigate various knowledge-sharing approaches on the encoder side while keeping the decoder language- or

[2]https://iwslt.org/2025/low-resource



(a) Universal model (b) Model with language-group-specific modules (c) Model with language-specific modules

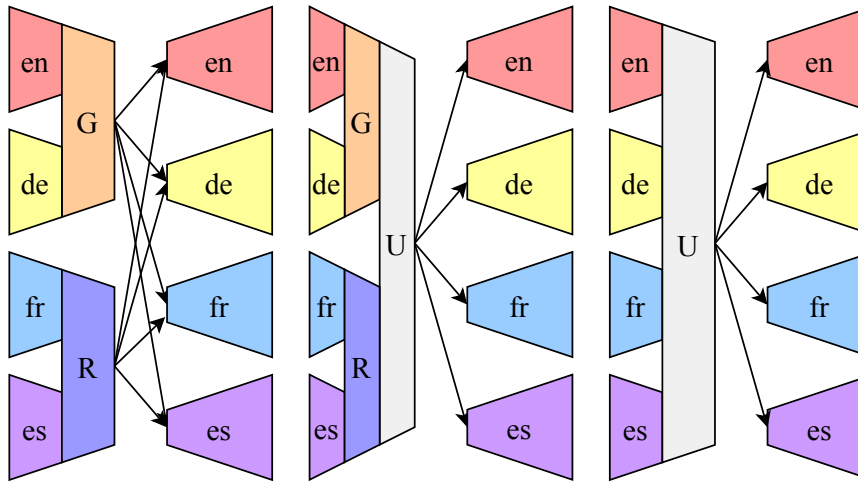
Figure 6: Different granularities of the modular architecture. *roa* – Romance; *gem* – Germanic; *tgt lang* – Target language token added to indicate the language of the output sentence.

language-group-specific. We propose a novel approach that combines universal, language-group-specific, and language-specific modules to overcome the shortcomings of both universal and language-specific models. Experiments on a multilingual dataset show that the proposed models achieve better results compared to purely universal and language-specific approaches, improving BLEU scores for zero-shot translations.

src	tgt						
	EN	DE	DA	FR	ES	PT	All
EN	–	1,000	100	1,000	500	10	2,610
DE	1,000	–	100	1,000	500	10	2,610
DA	100	100	–	10	0	0	210
FR	1,000	1,000	10	–	500	100	2,610
ES	500	500	0	500	–	100	1,600
PT	10	10	0	100	100	–	220
All	2,610	2,610	210	2,610	1,600	220	9,860

Table 11: Dataset sizes in thousands of sentence pairs for each language pair.

We tried to imitate real-world scenarios for the dataset where language pairs have varying amounts of data. The dataset consists of English, German, Danish, French, Spanish, and Portuguese parallel data, divided into Germanic (English, German, Danish) and Romance (French, Spanish, Portuguese) language groups. The training dataset is composed of 70% Europarl, 15% EMEA, and 15% JRC-Acquis, with different amounts of data for high-resource (1M), medium-resource(100K-500K), low-resource(10K), and zero-shot(0) language pairs shown



(a) Language group sharing

(b) Language group and universal sharing (tiered)

(c) Universal sharing (between all languages)

Figure 7: Different types of encoder layer sharing in the modular architecture. Note that the width of layers in the figure does not correspond to the actual width but rather reflects the sharing extent, i.e., all layers in the encoder have the same width dimension. *U* – universal, *G* – Germanic, *R* – Romance.

in Table:11. The test set consists of 2000 multi-parallel sentences for each language pair. For tokenisation, we use SentencePiece with the BPE algorithm, training separate models for each scenario—each individual language, each language group, and a unified model for the universal approach, with specific details in the paper. The dataset size is relatively small compared to state-of-the-art models. However, it is sufficient for drawing conclusions that can be applied at larger scales. The restrictions (dataset and model size) are mainly due to resource constraints.

We focus on improving the overall translation quality using different knowledge and layer-sharing methods. We propose various degrees of granularity (or specificity) of modules and layer sharing, including combining layers of various granularities into a tiered architecture. All models follow the transformer base architecture with six encoder and six decoder layers. The experiments we use follow different levels of granularity and can be summarized into the following categories:

- Baseline models - Universal (Fig:6a) and Modular system (Fig: 6c)
- Sharing the upper encoder layers grouped by language-groups (Fig: 7a)
- Tiered approach combining language-group-based and universally shared layers (2+2) (Fig:7b)
- Sharing the N upper encoder layers universally (Fig: 7c), which is much simpler way compared to language-group or tiered.

Architecture	Language pair resource					
	zero-shot	low	medium-low	medium-high	high	all
Universal	33.62	38.12	39.64	43.64	42.32	39.87
Group modular (GM)						
EA3–6	**35.03**	**39.48**	40.89	44.66	43.31	**41.06**
EA5–6	34.52	39.23	40.78	44.59	43.19	40.88
No sharing	33.76	38.90	40.75	44.60	43.32	40.73
Language modular (LM)						
EA3–6	34.73	38.79	**40.91**	44.68	43.36	40.90
EG3–4 EA5–6	34.57	38.61	40.76	**44.91**	**43.59**	40.90
EG 3–6	34.37	38.56	40.56	44.90	43.42	40.78
EA5–6	33.81	38.28	40.32	44.75	43.38	40.54
EG5 EA6	33.51	38.07	40.33	44.72	43.41	40.46
EG5–6	33.59	37.85	40.32	44.69	43.44	40.43
No sharing	32.14	37.19	39.92	44.74	43.50	40.02

Table 12: Average test set BLEU scores per language pair resource, all refers to average score over all resource settings. EG - encoder layer shared within language group, EA - encoder layer shared between all languages. Best score(s) per resource (column) in bold.

Enc. shared layer(s)	Language pair resource					
	zero-shot	low	medium-low	medium-high	high	all
No sharing	32.14	37.19	39.92	44.74	**43.50**	40.02
6	33.07	37.63	40.09	44.67	43.35	40.23
5–6	33.81	38.28	40.32	44.75	43.38	40.54
4–6	34.16	38.43	40.41	44.85	43.43	40.68
3–6	34.73	38.79	**40.91**	44.68	43.36	40.90
2–6	**34.97**	**39.03**	40.81	**44.94**	43.44	**41.03**
1–6	34.61	38.70	40.79	44.60	43.23	40.80

Table 13: Average test set BLEU scores per language pair resource, all refers to average score over all resource settings. BLEU scores are given for experiments with encoder layer sharing between all languages in the language modular model. Best score(s) per resource (column) in bold.

The main results show that the proposed models achieve higher translation quality than both baselines. The results are shown in Table:12. The modular model without any sharing performs worse on zero-shot and low-resource language pairs than the universal model but achieves better overall translation quality in medium-high and high-resource language pairs. Sharing two upper encoder layers in the modular model improves the low-resource MT quality while improving the high-resource MT quality that suffers in the universal NMT setting. The language-group-specific modular encoder and decoder architecture outperforms the universal model in all data scenarios, improving the modular architecture's zero-shot and low-resource translation quality essentially without harming the high-resource language translation quality. The best-performing model outperforms the baseline language modular model by 1.04 BLEU points and the universal model by 1.19 BLEU points on average.

We investigated the effect of sharing encoder layers in multilingual neural machine translation, shown in Table:13. We discovered that sharing more layers consistently increases BLEU scores until five upper encoder layers are shared. Sharing all six layers marginally reduces BLEU scores across all resource settings compared to sharing five layers. However, the modular model with shared encoder layers 2-6 achieves a BLEU score close to the best-performing model from our previous experiments.

Overall, our findings indicate that sharing multiple upper encoder layers can enhance translation quality in multilingual neural machine translation models while reducing the number of parameters compared to the modular approach. The biggest gains with sharing some layers, compared to baselines, come in the zero-shot performance, while the negative impact on high-resource languages is minimal. Ultimately, it represents a trade-off that can be strategically managed. This work showcases a novel approach combining features of both universal and modular architectures, achieving higher translation quality than both baselines.

5. LARGE-PRETRAINED MODELS ENABLE HIGH-QUALITY LOW-RESOURCE NMT

The vast majority of the world's 7,000+ languages are spoken by very small communities, with more than half having less than 10,000 speakers. This poses unique challenges for machine translation, as traditional techniques often struggle with such limited data - most of these languages have near-zero parallel data.

This chapter tackles the critical task of achieving accurate translation for languages with near-zero data, specifically focusing on Finno-Ugric languages. We present our exploration of optimal architectures and training strategies to address this challenge within the multilingual neural machine translation (NMT) framework. Our work is very relevant because we focus on extending existing state-of-the-art large pre-trained methods like NLLB, which supports 200 languages, and how to append new languages.

We investigate two key strategies to overcome the limitations of low-resource settings:

- Leveraging Transfer Learning: We explore how knowledge from large pre-trained multilingual (translation) models can be effectively transferred to Finno-Ugric languages, even those unseen during the original model training.
- Iterative Data Augmentation: We examine techniques like back-translation to create synthetic parallel data iteratively, significantly improving translation quality in the absence of naturally occurring parallel text.

Our research contributes to the development of more inclusive and powerful NMT systems. By enabling translation for under-resourced languages, we preserve linguistic diversity and hopefully aid cross-cultural understanding.

To summarize, this chapter presents three related approaches, all based on the central idea of 'using pre-trained models for translation.' Section 5.1 examines the viability of large pre-trained NMT systems for low-resource scenarios, including how to incorporate new characters into the models' vocabulary. Section 5.2 then focuses on fine-tuning for a single language to achieve optimal performance. Finally, Section 5.3 broadens the scope to a larger set of Finno-Ugric languages (mainly Cyrillic Finno-Ugric languages), illustrating how our multilingual approach can scale effectively across diverse low-resource settings.

lang-pair	raw	filtered	valid	test
Estonian-Võro	31 551	29 775	200	500
Finnish-North Sámi	77 710	62 837	200	500
FI-South Sámi	2913	2766	200	500
FI-Inari Sámi	10 639	9459	200	500
FI-Skolt Sámi	5769	2708	200	500
NO-South Sámi	17 388	15 702	200	500
NO-North Sámi	241 598	195 970	200	500
Norwegian-Lule Sámi	12 400	11 627	200	500
North Sámi-South Sámi	21 993	19 963	200	500
North Sámi-Lule Sámi	16 440	14 985	200	500
North Sámi-Inari Sámi	934	894	200	500
English-Livonian	617	280	586	856
Estonian-Livonian	14 261	12 887	586	856
Latvian-Livonian	11 732	10 763	586	856

Table 14: Parallel data numbers before and after filtering (in sentence pairs) are covered in columns raw and filtered. Validation and test set sentence counts are the other two columns.

5.1. Cross-lingual Transfer from Large Multilingual Translation Models to Unseen Under-resourced Languages

To address the critical need for extensive, high-quality data in NMT, while facing the challenges of low-resource language pairs, we present the first investigation into using large multilingual pre-trained models for Finno-Ugric languages. Our work demonstrates that leveraging such models for cross-lingual transfer learning is highly advantageous for this use case.

We perform cross-lingual transfer learning from these models to our target low-resource languages, which are unseen by the original pre-trained models. We experiment with seven low-resource Finno-Ugric languages: Livonian (LIV), Võro (VRO), North Sámi (SME), South Sámi (SMA), Inari Sámi (SMN), Lule Sámi (SMJ), and Skolt Sámi (SMS). We also include four high-resource languages that have the potential to help during training: Estonian, Finnish, Latvian, and English. The high-resource languages act as anchors for the pre-trained model we use as a transfer learning basis. We achieved state-of-the-art results on many translation directions on low-resource languages, and we show the feasibility and effectiveness of using pre-trained models for unseen (under-resourced) languages. We contribute to the open-source community by releasing our code here[1].

This paper builds upon our earlier dataset (Section 3.2, Table:5) and expands it to cover seven low-resource and four high-resource languages in total. We use only parallel data, including the Livonian dataset from Rikters et al., 2022, with

[1]https://github.com/TartuNLP/m2m-100-finetune

the updated dataset's size presented in Table:14. Recognizing issues within both our previous parallel data and the Livonian dataset, we meticulously addressed these shortcomings to create a new, more reliable benchmark test set. See the paper for details on data sources, identified problems, and the new test set.

We fine-tune the M2M-100 model by Fan et al. (2021) on our data using the Fairseq framework (Ott et al., 2019), changing only the learning rate and the number of update steps, which we adjust dynamically depending on data sizes. We evaluate our model on various test sets, including our new benchmark dataset created in this paper, WMT news test sets and Tatoeba test sets (Tiedemann, 2020), which are covered in Table:14.

To adapt the M2M-100 model for our Finno-Ugric use case, we needed to modify its original design, as it was not trained for our target languages. Specifically, we introduced new language ID tokens to signal our unseen languages to the model. Furthermore, we expanded the model's vocabulary to incorporate the unique characters and symbols present in our Finno-Ugric languages. These modifications were essential to prevent unknown symbol errors and enable effective cross-lingual transfer to our under-resourced languages.

Model	418M	not-EN	only-(VRO/SM*)	1.2B	*prev_best*
ET-VRO	25.72	25.73	**30.32**	26.03	*26.2*
VRO-ET	30.27	29.66	**33.95**	31.65	*31.7*
FI-SME	**38.45**	38.3	38.02	37.83	*32.3*
SME-FI	42.84	45.09	45.21	**45.83**	*37.5*
FI-SMA	17.82	20	22.56	**25.29**	*12.4*
SMA-FI	21.73	21.89	27.85	**29.44**	*10.9*
SME-SMA	33.54	33.74	34.93	**38.1**	*21.6*
SMA-SME	35.44	36.54	38.37	**42.97**	*21.0*

Table 15: BLEU scores with test data from Tars et al., 2021. "418M" and "1.2B" refer to models trained with all data. "not-EN" refers to a model trained without any English data. "only-(VRO/SM*)" refers to a model trained only on those specific languages and pairs. *prev_best* refers to best results by Tars et al. (2021). **bold** - best BLEU score for a language pair.

The results displayed in Table:15 demonstrate substantial advancements in machine translation performance across multiple language pairs. Notably, our models significantly outperform our own previous best results, in some cases by a staggering margin of +22 BLEU points and SMA-FI jumping from 10.9 to 29.44 BLEU. Interestingly, the VRO-ET and ET-VRO language pair does not achieve a jump in quality(418M and 1.2B versions), but when trained on a subset of the training data, the model performs better than the previous best. The most likely reason is that the inclusion of large amounts of back-translated data in the prior ET-VRO paper contributed (covered in Section 3.2) to its superior performance on this specific language pair.

Model	418M	not-EN	only-(VRO/LIV)	1.2B	*prev_best*
EN-LIV	10.64	2.51	8.97	**11.51**	*8.59*
LIV-EN	14.08	4.34	14.32	**15.85**	*14.69*
ET-LIV	13.97	13.7	13.87	**14.51**	*13.00*
LIV-ET	19.03	18.3	18.69	**19.62**	*17.76*
LV-LIV	13.64	13.43	13.89	**15.13**	*13.67*
LIV-LV	18.39	18.34	20.34	**20.47**	*17.55*

Table 16: BLEU scores on our new test set and liv4ever test set. “418M" and “1.2B" refer to models trained with all data. “not-EN" refers to model trained without any English data. “only-(VRO/LIV)" refers to models trained only on that specific language group data. *prev_best* refers to results by Rikters et al. (2022) on their fine-tuned models (without back-translation). **bold** - best BLEU score for a language pair.

Table:16 showcases impressive results for Livonian. Notably, we surpass the previous best results on English-Livonian translation, demonstrating the power of our approach in overcoming data scarcity challenges. Our methodology outperforms earlier work on Livonian by Rikters et al. (2022). Our model’s strength to generalize the English-Livonian shows a remarkable ability to learn EN-LIV and LIV-EN translation directions using just 280 training examples, with quality jumping from 2.51/4.34 ("not-EN" variant, which doesn’t basically learn) to 10.64/14.08 respectively on the model of the same size, meaning that very little data is required to get translations out.

Our approach leverages the power of pre-trained multilingual models trained on massive amounts of data. This provides a strong foundation for generalization across languages, making it highly effective for under-resourced scenarios. While the larger model (1.2B parameters) delivers superior translation quality, it comes with higher computational costs and memory demands. The smaller model (418M parameters) offers a more practical solution with reasonable performance, making it suitable for less powerful hardware.

In conclusion, this work demonstrates the feasibility and effectiveness of cross-lingual transfer learning from large multilingual pre-trained models for unseen, under-resourced languages. Notably, we demonstrate that significant gains in translation quality can be achieved even with less than 1000 sentences of parallel data. We introduce the first NMT models for Inari Sámi, Lule Sámi, and Skolt Sámi while achieving state-of-the-art results for other low-resource Finno-Ugric languages. Furthermore, we contribute to the field by creating a new benchmark dataset for Finno-Ugric language pairs that are publicly available under an open-source license.

5.2. Teaching unseen low-resource languages to large translation models

Building upon previous work with large pre-trained translation models, this paper investigates their adaptability for the very specific use case of English ↔ Livonian translation. Livonian, an unseen language for the original M2M-100 model, and a Finno-Ugric language with limited data present distinct challenges for machine translation.

Our approach centers on the hypothesis that the M2M-100 model's extensive multilingual knowledge can be effectively transferred to enhance its performance in translating the unseen language Livonian. Due to Livonian's linguistic similarities to Estonian and Latvian, we anticipate positive knowledge transfer from the M2M-100 model. Our approach consists of two main steps: (1) fine-tuning the model on the available Livonian dataset (Table:14) and (2) expanding this dataset through back-translation techniques. We aimed to improve the model specifically for English ↔ Livonian translation as part of the WMT22 shared task (Kocmi et al., 2022).

To further augment the parallel data, we employed back-translation, generating additional training examples by translating Livonian texts into English and then using these translations for further model training. We expanded and improved upon our previously used monolingual dataset from Section 3.1 (Table:6) to a improved version shown in Table:17.

Language	# Sentences
Võro	162807
Southern Sámi	55088
North Sámi	33964
Skolt Sámi	76685
Inari Sámi	122916
Lule Sámi	128180
Livonian	40329

Table 17: Monolingual data numbers, with expanded data, filtering, and language coverage compared to Table:6. Check the paper for more details (Tars et al. (2022a)).

Our results demonstrate a significant improvement over previous benchmarks for EN-LIV, with gains primarily stemming from our iterative back-translation and fine-tuning approach (Table:18). We hypothesize that the model developed a better general Finno-Ugric representation through two back-translation iterations on the broader dataset. Subsequent fine-tuning, specifically on Livonian data allowed for optimal results, surpassing those achieved by iterative Livonian-only fine-tuning. This suggests that closely related languages can provide valuable extra training data.

	1.2B Model		418M Model	
	EN-LIV	LIV-EN	EN-LIV	LIV-EN
Baseline	10.15	18.92	10.29	15.78
+ bt1	11.24	28.67	11.25	27.52
+ bt2	12.16	29.37	10.62	27.38
+ tuned on LIV	**15.19**	**31.06**	**12.83**	27.23
+ bt1 only-LIV	10.66	27.88	11.39	27.74
+ bt2 only-LIV	11.21	29.85	11.63	28.81
+ tuned on LIV	11.56	30.33	11.53	**29.27**
Previous Best	*11.03*	*19.01*	*11.03*	*19.01*

Table 18: Experiment results on BLEU. "1.2B" and "418M" refer to models trained with all original parallel data. "bt1" is trained on parallel + first back-translation iteration data, "bt2" on parallel + second back-translation iteration data. "only-LIV" - only data between ET-EN-LV-LIV languages was used for training. "tuned on LIV" refers to the "bt2" model that was tuned on ET-EN-LV-LIV original parallel data. The last row represents previously best results for EN-LIV-EN by (Rikters et al., 2022).

Our results in the WMT shared task (details in Kocmi et al. (2022)) were highly successful, placing us first in Liv $\rightarrow$ En and second in En $\rightarrow$ Liv according to automatic metrics (Comet, BLEU, and CHRF). While human evaluation data was limited for conclusive ranking and there were many ties, our systems still ranked in the top.

This study demonstrates the potential of adapting large pre-trained multilingual models for very low-resource languages. Our findings emphasize the effectiveness of transfer learning and data augmentation techniques. We show that fine-tuning pre-trained models like M2M-100 can significantly improve translation quality even with limited data and can be applied to very narrow and specific tasks.

However, this study focused solely on Livonian and did not investigate the impact that fine-tuning has on other language directions within the multilingual model. This presents an important area for future research to ensure optimal translation performance.

5.3. Machine Translation for Low-resource Finno-Ugric Languages

The paper tackles the challenge of machine translation for low-resource Finno-Ugric languages, focusing on enhancing NMT systems. The biggest goal of the paper was to include more Finno-Ugric languages in the translation system. Finno-Ugric languages are all from one language family, but many other languages influence them because of geography and closeness by having many sim-

ilar words for the same things. The main geographical influencers are Latvian, Norwegian and Russian. In the previous subchapters, we included Latvian and Norwegian in our training process, as discussed over the course of this PhD thesis. The core focus of this paper was to add additional languages that use the Cyrillic alphabet - due to that, we also included the Russian language to use it as a transfer learning basis. This paper covers many unique obstacles, like the scarcity of linguistic data for these languages. Our goal again is to create the first (multilingual) machine translation system, which uses the latest technology for many of the new languages and translation directions.

	parallel (Ru)		
	bible	others	total
Komi (KPV)	11	2	13
Komi-Permyak (KOI)	8	0.3	8.3
Udmurt (UDM)	30		30
Hill Mari (MRJ)	8		8
Meadow Mari (MHR)	9		9
Erzya (MYV)	11.5	0.9	12.4
Moksha (MDF)	11.5	1	12.5
Karelian (KRL)	10.5	7.7	18.2
Ludian (LUD)		10.5	10.5
Livvi Karelian (OLO)	11.9	4	15.9
Veps (VEP)	16.4	11.1	27.5
Mansi (MNS)	0.7		0.7
Khanty (KCA)	2		2

Table 19: The collected parallel corpus with Russian. The figures in the table are in thousands of sentences.

The main contribution of this paper was the addition of new languages. The added languages (compared to the previous research described in this thesis) were a high-resource Finno-Ugric language Hungarian and geographically linked high-resource Russian language; the added low-resource languages were all from the Finno-Ugric family: Komi(KPV), Udmurt(UDM), Erzya(MYV), Moksha(MDF), Livvi Karelian(OLO), Veps(VEP), Mansi(MNS) and Khanty(KCA). We also included the source (or domain) with the gathered datasets: wiki, bible, news, and others. A very detailed report of the whole dataset is covered in the paper (L. Yankovskaya et al. (2023)), and the sources are linked. The main table of parallel data is given in Table:19.

The next important goal was to create a multilingual benchmark for the low-resource languages. We chose the first 250 sentences from the FLORES200 dataset and extended the number of languages it covers. The languages for which

	mono				
	wiki	bible	news	others	total
Komi	18.4	4.5	38.3		61.2
Komi-Permyak	11.5	1.2			12.7
Udmurt	43.5	3.7	36		83.2
Hill Mari	49.5			14.6	64.1
Meadow Mari	141			109	251
Erzya	73.8		1.3	7.7	82.8
Moksha	8	3.9	3.9	0.3	16.1
Karelian		1.8		18.4	20.2
Ludian				5.3	5.3
Livvi Karelian			21	19.4	40.4
Veps	71.3	0.9	7.8	35.3	115.3
Võro				162	162
Livonian				40	40
South Sámi				55	55
North Sámi				34	34
Lule Sámi				128	128
Inari Sámi				123	123
Skolt Sámi				76.7	76.7
Mansi		0.8	10.3		11.1
Khanty		0.8	13.3		14.1

Table 20: The collected monolingual corpus of the low-resource languages. The figures in the table are in thousands of sentences.

we created the benchmark data are Komi, Udmurt, Hill Mari, Meadow Mari, Erzya, Moksha, Livonian, Mansi, and Livvi Karelian. We included these languages since we found translators for them. The data and languages covered are presented in Table:20.

One important note is that the FLORES200 dataset covers very specialized topics, which can be nearly impossible to translate accurately, even with bilingual translators with academic backgrounds. A suggestion would be to create large multilingual datasets with careful consideration of the sentences included in the benchmark. For example, consider sentence 382 in the FLORES200 dataset:

> "Mosasaurus was the apex predator of its time, so it feared nothing, except other mosasaurs. Its long jaws were studded with more than 70 razor-sharp teeth, along with an extra set in the roof of its mouth, meaning that there was no escape for anything that crossed its path."

Translating a sentence like this accurately into predominantly spoken Finno-Ugric languages poses significant challenges due to the complex biological terminology and specific historical context. In response to such challenges, the FLORES200 dataset extension included some adjustments, so please look for details in the paper.

In this study, we integrated new Finno-Ugric languages and Russian into our NMT system by experimenting with different data scenarios (presented in Table:21), focusing on optimizing the back-translation process and exploiting the transfer learning capabilities of the Russian language. We used a similar approach as before for the model base architecture - using a pre-trained large translation model M2M-100.

The results section highlights significant enhancements in the NMT system. Firstly - all experiments led to an increase over the baseline, which used only the parallel data available, meaning that synthetic data creation via back-translation was very effective, presented in Table:21. If we compare the baseline to the L-H or the L-rH approaches, then we can see almost a ten chrF++ point increase over all the language pairs.

Table:21c showcases the language-specific chrF++ score enhancements, where we can observe that the different training data scenarios heavily influence the low-resource languages we target. There are a couple of limitations to remember: the FLORES200 added benchmark set covers only a part of the languages, which should be extended further to include more Finno-Ugric languages. Most of the parallel data is biblical, so further data collection efforts will help a lot. Transliteration can be tried to help translation quality further.

This paper marks a significant advancement in NMT for low-resource Finno-Ugric languages, showcasing the successful incorporation of these languages into a robust translation system. For many translation directions, the system is the first to cover all these languages and is openly available on TartuNLP's website.[2] The findings have broad implications for the field of machine translation and language preservation, offering valuable insights for future research in the area of no-or-low-resource machine translation with very limited monolingual data.

[2]https://neurotolge.ee

	low-low	low-high	high-low	all pairs
baseline	18.7	24.0	20.7	23.7
all-all	20.2	36.5	19.1	28.5
all-all-10	25.9	34.3	24.1	30.3
L-H	26.6	**36.6**	25.8	**32.3**
L-rH	**27.2**	35.5	**26.8**	32.0

(a) low - low-resource languages, high - high-resource languages."-" indicates two-way translation directions between the languages.

	to-RU	to-EN	to-ET	to-FI	to-HU	to-LV	to-NO
baseline	19.6	25.6	26.6	25.1	22.0	24.3	24.8
all-all	42.3	39.8	**28.2**	36.8	35.2	37.6	**35.4**
all-all-10	39.2	37.5	27.8	34.7	32.4	35.2	33.6
L-H	**42.9**	**40.4**	27.8	**37.2**	**35.5**	**38.0**	34.6
L-rH	41.8	39.4	26.7	36.4	33.8	35.9	34.6

(b) to-* indicates translation directions from low-resource languages to the respective high-resource language.

	to-KPV	to-LIV	to-MDF	to-MHR	to-MNS	to-MRJ	to-MYV	to-OLO	to-UDM
baseline	15.9	28.4	22.1	21.3	12.2	19.9	22.9	22.7	21.0
all-all	15.9	26.0	18.2	24.4	12.4	15.2	16.7	21.2	21.5
all-all-10	22.3	28.6	25.2	28.3	13.7	22.1	23.1	25.3	28.1
L-H	24.6	29.5	27.0	**30.8**	14.3	23.8	24.4	**27.1**	31.2
L-rH	**26.4**	**29.7**	**28.5**	30.6	**16.1**	**26.2**	**25.2**	26.7	**31.6**

(c) to-* indicates translation directions from high-resource languages to the respective low-resource language.

Table 21: Average chrF++ results for all experiments across different language pair clusters on FLORES benchmarks. **Bold** - highest score per grouping. all-all - contains BT data from every language pair. all-all-10 - contains 10% of BT data used in all-all. L-H - contains BT data from each low-resource(L) language to each high-resource(H) language and vice versa. L-rH - contains BT data from each language to its related high-resource language(rH) + high-resource languages it had parallel data with and vice versa.

6. CONCLUSION

This thesis presents contributions to the field of multilingual neural machine translation (NMT), focusing on enhancing translation quality for low-resource languages using innovative approaches and methodologies. The findings highlight the effectiveness of multilingual models in handling diverse linguistic scenarios, from high-resource to extremely low-resource languages.

The primary contributions of this research include the development of new methodologies for integrating low-resource languages into NMT systems, the creation of multilingual benchmark datasets, and developing a competitive Estonian-centric translation system. These models, trained on publicly available data, have been deployed online as open-source tools, increasing accessibility for end-users. Key results show substantial improvements in translation quality, particularly for Finno-Ugric languages, through transfer learning and back-translation methods.

This thesis introduces novel techniques for fine-tuning large pre-trained models and leveraging knowledge transfer among related languages. It also explores architectural innovations in modular NMT, combining language-specific encoders and decoders with shared components. This has resulted in a tiered architecture that balances universal knowledge sharing with language-specific modules, improving performance across both high- and low-resource settings.

Challenges encountered during this research include data scarcity for low-resource languages and the complexity of fine-tuning large multilingual models. To address these, back-translation and synthetic data generation experiments proved effective, even in limited-resource environments. The importance of data preprocessing, model design, and optimization strategies became clear, with insights actively shared within the research group to support future projects.

The practical impact of this work extends beyond academic contributions. It provides high-quality machine translation tools for languages with limited digital resources, supports language preservation efforts, and improves cross-cultural communication. The publicly available translation services developed in this thesis aim to make NMT more accessible to a wider range of language communities.

Future research should focus on expanding the support for low-resource languages, collecting additional data, and refining data augmentation techniques. Integrating all Finno-Ugric languages into large language models (LLMs) could enable new applications beyond translation. Further exploration of model architectures and fine-tuning strategies will also be important, especially to improve efficiency without sacrificing performance.

From an engineering perspective, this research highlights the trade-offs between model size, performance, and resource efficiency. While large pre-trained models offer strong performance, smaller, fine-tuned models can deliver competitive results with lower resource demands. The practical deployment of these models provided insight into optimizing system performance, resource usage, and model monitoring.

In the long term, advancing machine translation for low-resource languages requires collaboration, open research, and a focus on practical applications. By addressing the challenges facing underrepresented languages, this work contributes to preserving linguistic diversity and improving global communication. The focus on building tools that are freely available and practically useful ensures that the impact of this research extends beyond academia, supporting language communities worldwide.

BIBLIOGRAPHY

Aharoni, R., Johnson, M., & Firat, O. (2019). Massively multilingual neural machine translation. In J. Burstein, C. Doran, & T. Solorio (Eds.), *Proceedings of the 2019 conference of the north American chapter of the association for computational linguistics: Human language technologies, volume 1 (long and short papers)* (pp. 3874–3884). Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1388

Araabi, A., & Monz, C. (2020). Optimizing transformer for low-resource neural machine translation. In D. Scott, N. Bel, & C. Zong (Eds.), *Proceedings of the 28th international conference on computational linguistics* (pp. 3429–3435). International Committee on Computational Linguistics. https://doi.org/10.18653/v1/2020.coling-main.304

Arivazhagan, N., Bapna, A., Firat, O., Lepikhin, D., Johnson, M., Krikun, M., Chen, M., Cao, Y., Foster, G., Cherry, C., Macherey, W., Chen, Z., & Wu, Y. (2019). *Massively multilingual neural machine translation in the wild: Findings and challenges*. https://doi.org/10.48550/arXiv.1907.05019

Bahdanau, D., Cho, K., & Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *ArXiv*, *1409*.

Baltic Times. (2023). *Ut's machine translation engine now supports 17 more finno-ugric languages* [Accessed: 12.02.2025]. https://www.baltictimes.com/ut_s_machine_translation_engine_now_supports_17_more_finno-ugric_languages/

Barrault, L., Bojar, O., Costa-jussà, M. R., Federmann, C., Fishel, M., Graham, Y., Haddow, B., Huck, M., Koehn, P., Malmasi, S., Monz, C., Müller, M., Pal, S., Post, M., & Zampieri, M. (2019). Findings of the 2019 conference on machine translation (WMT19). In O. Bojar, R. Chatterjee, C. Federmann, M. Fishel, Y. Graham, B. Haddow, M. Huck, A. J. Yepes, P. Koehn, A. Martins, C. Monz, M. Negri, A. Névéol, M. Neves, M. Post, M. Turchi, & K. Verspoor (Eds.), *Proceedings of the fourth conference on machine translation (volume 2: Shared task papers, day 1)* (pp. 1–61). Association for Computational Linguistics. https://doi.org/10.18653/v1/W19-5301

Bawden, R., Zhang, B., Tättar, A., & Post, M. (2020a). ParBLEU: Augmenting metrics with automatic paraphrases for the WMT'20 metrics shared task. *Proceedings of the Fifth Conference on Machine Translation*, 887–894. https://aclanthology.org/2020.wmt-1.98

Bawden, R., Zhang, B., Yankovskaya, L., Tättar, A., & Post, M. (2020b). A study in improving BLEU reference coverage with diverse automatic paraphrasing. *Findings of the Association for Computational Linguistics: EMNLP 2020*, 918–932. https://doi.org/10.18653/v1/2020.findings-emnlp.82

Behnke, M., Bogoychev, N., Aji, A. F., Heafield, K., Nail, G., Zhu, Q., Tchistiakova, S., van der Linde, J., Chen, P., Kashyap, S., & Grundkiewicz, R. (2021). Efficient machine translation with model pruning and quantiza-

tion. In L. Barrault, O. Bojar, F. Bougares, R. Chatterjee, M. R. Costa-jussa, C. Federmann, M. Fishel, A. Fraser, M. Freitag, Y. Graham, R. Grundkiewicz, P. Guzman, B. Haddow, M. Huck, A. J. Yepes, P. Koehn, T. Kocmi, A. Martins, M. Morishita, & C. Monz (Eds.), *Proceedings of the sixth conference on machine translation* (pp. 775–780). Association for Computational Linguistics. https://aclanthology.org/2021.wmt-1.74

Bergmanis, T., Pinnis, M., Rozis, R., Šlapiņš, J., Šics, V., Bernāne, B., Pužulis, G., Titomers, E., Tättar, A., Purason, T., Kuulmets, H.-A., Luhtaru, A., Rätsep, L., Tars, M., Laumets-Tättar, A., & Fishel, M. (2022). MTee: Open machine translation platform for Estonian government. *Proceedings of the 23rd Annual Conference of the European Association for Machine Translation*, 309–310. https://aclanthology.org/2022.eamt-1.44

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., … Amodei, D. (2020). Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, & H. Lin (Eds.), *Advances in neural information processing systems* (pp. 1877–1901). Curran Associates, Inc.

Burlot, F., & Yvon, F. (2018). Using monolingual data in neural machine translation: A systematic study. *Proceedings of the Third Conference on Machine Translation: Research Papers*, 144–155. https://doi.org/10.18653/v1/W18-6315

Chen, L., Ma, S., Zhang, D., Wei, F., & Chang, B. (2023). On the off-target problem of zero-shot multilingual neural machine translation. In A. Rogers, J. Boyd-Graber, & N. Okazaki (Eds.), *Findings of the association for computational linguistics: Acl 2023* (pp. 9542–9558). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.findings-acl.608

Cho, K., van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning phrase representations using RNN encoder–decoder for statistical machine translation. In A. Moschitti, B. Pang, & W. Daelemans (Eds.), *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (pp. 1724–1734). Association for Computational Linguistics. https://doi.org/10.3115/v1/D14-1179

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., & Stoyanov, V. (2020). Unsupervised cross-lingual representation learning at scale. In D. Jurafsky, J. Chai, N. Schluter, & J. Tetreault (Eds.), *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 8440–8451). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.747

Dabre, R., Chu, C., & Kunchukuttan, A. (2020). A survey of multilingual neural machine translation. *ACM Comput. Surv.*, *53*(5). https://doi.org/10.1145/3406095

Del, M., Tättar, A., & Fishel, M. (2018). Phrase-based unsupervised machine translation with compositional phrase embeddings. *Proceedings of the Third Conference on Machine Translation*, 365–371. http://www.aclweb.org/anthology/W18-64034

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In J. Burstein, C. Doran, & T. Solorio (Eds.), *Proceedings of the 2019 conference of the north American chapter of the association for computational linguistics: Human language technologies, volume 1 (long and short papers)* (pp. 4171–4186). Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1423

Erhan, D., Bengio, Y., Courville, A., Manzagol, P.-A., Vincent, P., & Bengio, S. (2010). Why does unsupervised pre-training help deep learning? *J. Mach. Learn. Res.*, *11*, 625–660.

Escolano, C., Costa-jussà, M. R., & Fonollosa, J. A. R. (2019). From bilingual to multilingual neural machine translation by incremental training. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, 236–242. https://doi.org/10.18653/v1/P19-2033

Escolano, C., Costa-jussà, M. R., Fonollosa, J. A. R., & Artetxe, M. (2021). Multilingual machine translation: Closing the gap between shared and language-specific encoder-decoders. In P. Merlo, J. Tiedemann, & R. Tsarfaty (Eds.), *Proceedings of the 16th conference of the european chapter of the association for computational linguistics: Main volume* (pp. 944–948). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.eacl-main.80

Fan, A., Bhosale, S., Schwenk, H., Ma, Z., El-Kishky, A., Goyal, S., Baines, M., Celebi, O., Wenzek, G., Chaudhary, V., Goyal, N., Birch, T., Liptchinsky, V., Edunov, S., Grave, E., Auli, M., & Joulin, A. (2021). Beyond english-centric multilingual machine translation. *J. Mach. Learn. Res.*, *22*(1).

Freitag, M., Mathur, N., Lo, C.-k., Avramidis, E., Rei, R., Thompson, B., Kocmi, T., Blain, F., Deutsch, D., Stewart, C., Zerva, C., Castilho, S., Lavie, A., & Foster, G. (2023). Results of WMT23 metrics shared task: Metrics might be guilty but references are not innocent. In P. Koehn, B. Haddow, T. Kocmi, & C. Monz (Eds.), *Proceedings of the eighth conference on machine translation* (pp. 578–628). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.wmt-1.51

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning* [http://www.deeplearningbook.org]. MIT Press.

Gu, J., Hassan, H., Devlin, J., & Li, V. O. (2018). Universal neural machine translation for extremely low resource languages. In M. Walker, H. Ji, & A. Stent (Eds.), *Proceedings of the 2018 conference of the north American chapter of the association for computational linguistics: Human language technologies, volume 1 (long papers)* (pp. 344–354). Association for Computational Linguistics. https://doi.org/10.18653/v1/N18-1032

Haddow, B., Bawden, R., Miceli Barone, A. V., Helcl, J., & Birch, A. (2022). Survey of low-resource machine translation. *Computational Linguistics*, *48*(3), 673–732. https://doi.org/10.1162/coli_a_00446

Hieber, F., Domhan, T., Denkowski, M., Vilar, D., Sokolov, A., Clifton, A., & Post, M. (2017). Sockeye: A toolkit for neural machine translation. *arXiv preprint arXiv:1712.05690*.

Ingel, A., Shahroudi, N., Kängsepp, M., Tättar, A., Komisarenko, V., & Kull, M. (2020). Correlated daily time series and forecasting in the m4 competition [M4 Competition]. *International Journal of Forecasting*, *36*(1), 121–128. https://doi.org/https://doi.org/10.1016/j.ijforecast.2019.02.018

Johnson, M., Schuster, M., Le, Q. V., Krikun, M., Wu, Y., Chen, Z., Thorat, N., Viégas, F., Wattenberg, M., Corrado, G., Hughes, M., & Dean, J. (2017). Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, *5*, 339–351. https://doi.org/10.1162/tacl_a_00065

Joshi, P., Santy, S., Budhiraja, A., Bali, K., & Choudhury, M. (2020). The state and fate of linguistic diversity and inclusion in the NLP world. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 6282–6293. https://doi.org/10.18653/v1/2020.acl-main.560

Kauers, M., Vogel, S., Fügen, C., & Waibel, A. (2002). Interlingua based statistical machine translation. https://doi.org/10.21437/ICSLP.2002-513

Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A., Milan, K., Quan, J., Ramalho, T., Grabska-Barwinska, A., Hassabis, D., Clopath, C., Kumaran, D., & Hadsell, R. (2016). Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, *114*. https://doi.org/10.1073/pnas.1611835114

Kocmi, T., & Bojar, O. (2018). Trivial transfer learning for low-resource neural machine translation. *CoRR*, *abs/1809.00357*. http://arxiv.org/abs/1809.00357

Kocmi, T., Bawden, R., Bojar, O., Dvorkovich, A., Federmann, C., Fishel, M., Gowda, T., Graham, Y., Grundkiewicz, R., Haddow, B., Knowles, R., Koehn, P., Monz, C., Morishita, M., Nagata, M., Nakazawa, T., NovÃ¡k, M., Popel, M., PopoviÄ‡, M., & Shmatova, M. (2022). Findings of the 2022 conference on machine translation (wmt22). *Proceedings of the Seventh Conference on Machine Translation*, 1–45. https://aclanthology.org/2022.wmt-1.1

Koehn, P. (2009). *Statistical machine translation*. Cambridge University Press.

Koehn, P., & Hoang, H. (2007a). Factored translation models. In J. Eisner (Ed.), *Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL)* (pp. 868–876). Association for Computational Linguistics. https://aclanthology.org/D07-1091

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., & Herbst, E. (2007b). Moses: Open source toolkit for statistical machine translation. In S. Ananiadou (Ed.), *Proceedings of the 45th annual meeting of the association for computational linguistics companion volume proceedings of the demo and poster sessions* (pp. 177–180). Association for Computational Linguistics. https://aclanthology.org/P07-2045

Koehn, P., Chaudhary, V., El-Kishky, A., Goyal, N., Chen, P.-J., & Guzmán, F. (2020). Findings of the WMT 2020 shared task on parallel corpus filtering and alignment. In L. Barrault, O. Bojar, F. Bougares, R. Chatterjee, M. R. Costa-jussà, C. Federmann, M. Fishel, A. Fraser, Y. Graham, P. Guzman, B. Haddow, M. Huck, A. J. Yepes, P. Koehn, A. Martins, M. Morishita, C. Monz, M. Nagata, T. Nakazawa, & M. Negri (Eds.), *Proceedings of the fifth conference on machine translation* (pp. 726–742). Association for Computational Linguistics. https://aclanthology.org/2020.wmt-1.78/

Kudo, T., & Richardson, J. (2018a). SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In E. Blanco & W. Lu (Eds.), *Proceedings of the 2018 conference on empirical methods in natural language processing: System demonstrations* (pp. 66–71). Association for Computational Linguistics. https://doi.org/10.18653/v1/D18-2012

Kudo, T., & Richardson, J. (2018b). SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In E. Blanco & W. Lu (Eds.), *Proceedings of the 2018 conference on empirical methods in natural language processing: System demonstrations* (pp. 66–71). Association for Computational Linguistics. https://doi.org/10.18653/v1/D18-2012

Kuulmets, H.-A., Tattar, A., & Fishel, M. (2022). Estonian language understanding: A case study on the copa task. *Baltic Journal of Modern Computing*, *10*(3), 470–480. https://doi.org/10.22364/bjmc.2022.10.3.19

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., & Zettlemoyer, L. (2020). BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In D. Jurafsky, J. Chai, N. Schluter, & J. Tetreault (Eds.), *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 7871–7880). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.703

Lita, L. V., Ittycheriah, A., Roukos, S., & Kambhatla, N. (2003). Truecasing. *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1*, 152–159. https://doi.org/10.3115/1075096.1075116

Luong, M.-T., & Manning, C. (2015). Stanford neural machine translation systems for spoken language domains. In M. Federico, S. Stüker, & J. Niehues (Eds.), *Proceedings of the 12th international workshop on spoken language translation: Evaluation campaign* (pp. 76–79). https://aclanthology.org/2015.iwslt-evaluation.11

Lyu, S., Son, B., Yang, K., & Bae, J. (2020). Revisiting Modularized Multilingual NMT to Meet Industrial Demands. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 5905–5918. https://doi.org/10.18653/v1/2020.emnlp-main.476

Marzotto, M. (2019). Marginalized mother languages: Two ways to improve the lives of the people who speak them [Accessed: 2024-06-15]. https://translatorswithoutborders.org/blog/marginalized-mother-languages-two-ways-to-improve-the-lives-of-the-people-who-speak-them/

NLLB Team, Costa-jussà, M. R., Cross, J., Çelebi, O., Elbayad, M., Heafield, K., Heffernan, K., Kalbassi, E., Lam, J., Licht, D., Maillard, J., Sun, A., Wang, S., Wenzek, G., Youngblood, A., Akula, B., Barrault, L., Mejia-Gonzalez, G., Hansanti, P., … Wang, J. (2022). No language left behind: Scaling human-centered machine translation. https://arxiv.org/abs/2207.04672

Olev, A., & Alumäe, T. (2022). Estonian speech recognition and transcription editing service. *Baltic Journal of Modern Computing*, *10*. https://doi.org/10.22364/bjmc.2022.10.3.14

OpenAI. (2023). Gpt-4 technical report. *ArXiv*, *abs/2303.08774*. https://arxiv.org/abs/2303.08774

Ott, M., Edunov, S., Baevski, A., Fan, A., Gross, S., Ng, N., Grangier, D., & Auli, M. (2019). Fairseq: A fast, extensible toolkit for sequence modeling. *Proceedings of NAACL-HLT 2019: Demonstrations*.

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). Bleu: A method for automatic evaluation of machine translation. In P. Isabelle, E. Charniak, & D. Lin (Eds.), *Proceedings of the 40th annual meeting of the association for computational linguistics* (pp. 311–318). Association for Computational Linguistics. https://doi.org/10.3115/1073083.1073135

Popović, M. (2015). ChrF: Character n-gram F-score for automatic MT evaluation. In O. Bojar, R. Chatterjee, C. Federmann, B. Haddow, C. Hokamp, M. Huck, V. Logacheva, & P. Pecina (Eds.), *Proceedings of the tenth workshop on statistical machine translation* (pp. 392–395). Association for Computational Linguistics. https://doi.org/10.18653/v1/W15-3049

Popović, M. (2017). ChrF++: Words helping character n-grams. In O. Bojar, C. Buck, R. Chatterjee, C. Federmann, Y. Graham, B. Haddow, M. Huck,

A. J. Yepes, P. Koehn, & J. Kreutzer (Eds.), *Proceedings of the second conference on machine translation* (pp. 612–618). Association for Computational Linguistics. https://doi.org/10.18653/v1/W17-4770

Postimees. (2023). *Tartu ülikooli masintõlkemootor toetab nüüd veel 17 soome-ugri keelt* [Accessed: 12.02.2025]. https://teadus.postimees.ee/7731822/tartu-ulikooli-masintolkemootor-toetab-nuud-veel-17-soome-ugri-keelt

Purason, T., & Tättar, A. (2022). Multilingual neural machine translation with the right amount of sharing. *Proceedings of the 23rd Annual Conference of the European Association for Machine Translation*, 91–100.

Purason, T., Tättar, A., & Fishel, M. (2024). Mixing and matching: Combining independently trained translation model components. In R. Vázquez, T. Mickus, J. Tiedemann, I. Vulić, & A. Üstün (Eds.), *Proceedings of the 1st workshop on modular and open multilingual nlp (moomin 2024)* (pp. 44–56). Association for Computational Linguistics. https://aclanthology.org/2024.moomin-1.5

Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving language understanding by generative pre-training.

Rei, R., Stewart, C., Farinha, A. C., & Lavie, A. (2020). COMET: A neural framework for MT evaluation. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2685–2702. https://www.aclweb.org/anthology/2020.emnlp-main.213

Rikters, M., Tomingas, M., Tuisk, T., Ernštreits, V., & Fishel, M. (2022). Machine translation for Livonian: Catering to 20 speakers. In S. Muresan, P. Nakov, & A. Villavicencio (Eds.), *Proceedings of the 60th annual meeting of the association for computational linguistics (volume 2: Short papers)* (pp. 508–514). Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.acl-short.55

Sennrich, R., Haddow, B., & Birch, A. (2016a). Improving neural machine translation models with monolingual data. In K. Erk & N. A. Smith (Eds.), *Proceedings of the 54th annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp. 86–96). Association for Computational Linguistics. https://doi.org/10.18653/v1/P16-1009

Sennrich, R., & Haddow, B. (2016b). Linguistic input features improve neural machine translation. In O. Bojar, C. Buck, R. Chatterjee, C. Federmann, L. Guillou, B. Haddow, M. Huck, A. J. Yepes, A. Névéol, M. Neves, P. Pecina, M. Popel, P. Koehn, C. Monz, M. Negri, M. Post, L. Specia, K. Verspoor, J. Tiedemann, & M. Turchi (Eds.), *Proceedings of the first conference on machine translation: Volume 1, research papers* (pp. 83–91). Association for Computational Linguistics. https://doi.org/10.18653/v1/W16-2209

Sennrich, R., Haddow, B., & Birch, A. (2016c). Neural machine translation of rare words with subword units. In K. Erk & N. A. Smith (Eds.), *Proceedings of the 54th annual meeting of the association for computational linguistics*

(volume 1: Long papers) (pp. 1715–1725). Association for Computational Linguistics. https://doi.org/10.18653/v1/P16-1162

Sennrich, R., Birch, A., Currey, A., Germann, U., Haddow, B., Heafield, K., Miceli Barone, A. V., & Williams, P. (2017). The University of Edinburgh's neural MT systems for WMT17. In O. Bojar, C. Buck, R. Chatterjee, C. Federmann, Y. Graham, B. Haddow, M. Huck, A. J. Yepes, P. Koehn, & J. Kreutzer (Eds.), *Proceedings of the second conference on machine translation* (pp. 389–399). Association for Computational Linguistics. https://doi.org/10.18653/v1/W17-4739

Shi, T., Zhao, S., Li, X., Wang, X., Zhang, Q., Ai, D., Dang, D., Zhengshan, X., & Hao, J. (2020). OPPO's machine translation systems for WMT20. In L. Barrault, O. Bojar, F. Bougares, R. Chatterjee, M. R. Costa-jussà, C. Federmann, M. Fishel, A. Fraser, Y. Graham, P. Guzman, B. Haddow, M. Huck, A. J. Yepes, P. Koehn, A. Martins, M. Morishita, C. Monz, M. Nagata, T. Nakazawa, & M. Negri (Eds.), *Proceedings of the fifth conference on machine translation* (pp. 282–292). Association for Computational Linguistics. https://aclanthology.org/2020.wmt-1.30/

Snover, M., Dorr, B., Schwartz, R., Micciulla, L., & Makhoul, J. (2006). A study of translation edit rate with targeted human annotation. *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, 223–231. https://aclanthology.org/2006.amta-papers.25

Study in Estonia. (2023). *Tartu university's machine translation engine now translates 23 finno-ugric languages* [Accessed: 12.02.2025]. https://www.studyinestonia.ee/news/tartu-universitys-machine-translation-engine-now-translates-23-finno-ugric-languages-0

Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to sequence learning with neural networks. *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, 3104–3112.

Tars, M., Tättar, A., & Fišel, M. (2021). Extremely low-resource machine translation for closely related languages. *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, 41–52.

Tars, M., Tattar, A., & Fishel, M. (2022a). Cross-lingual transfer from large multilingual translation models to unseen under-resourced languages. *Baltic Journal of Modern Computing*, *10*(3), 435–446. https://doi.org/10.22364/bjmc.2022.10.3.16

Tars, M., Purason, T., & Tättar, A. (2022b). Teaching unseen low-resource languages to large translation models. *Proceedings of the Seventh Conference on Machine Translation (WMT)*, 375–380.

Tättar, A., & Fishel, M. (2017). Bleu2vec: The painfully familiar metric on continuous vector space steroids. *Proceedings of the Second Conference on Machine Translation, Volume 2: Shared Task Papers*, 619–622. http://www.aclweb.org/anthology/W17-4771

Tättar, A., Korotkova, E., & Fishel, M. (2019). University of Tartu's multilingual multi-domain WMT19 news translation shared task submission. *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, 382–385. https://doi.org/10.18653/v1/W19-5342

Tättar, A., Purason, T., Kuulmets, H.-A., Luhtaru, A., Ratsep, L., Tars, M., Pinnis, M., Bergmanis, T., & Fishel, M. (2022). Open and competitive multilingual neural machine translation in production. *Baltic Journal of Modern Computing*, *10*(3), 422–434. https://doi.org/10.22364/bjmc.2022.10.3.15

Tiedemann, J. (2012). Parallel data, tools and interfaces in OPUS. *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*.

Tiedemann, J. (2020). The Tatoeba Translation Challenge – Realistic data sets for low resource and multilingual MT. *Proceedings of the Fifth Conference on Machine Translation*, 1174–1182. https://www.aclweb.org/anthology/2020.wmt-1.139

University of Tartu. (2023). *University of tartu machine translation engine now supports 17 new finno-ugric languages* [Accessed: 12.02.2025]. https://ut.ee/en/content/university-tartu-machine-translation-engine-now-supports-17-new-finno-ugric-languages

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances in neural information processing systems*. Curran Associates, Inc.

Yankovskaya, E., Tättar, A., & Fishel, M. (2018). Quality estimation with force-decoded attention and cross-lingual embeddings. *Proceedings of the Third Conference on Machine Translation*, 829–834. http://www.aclweb.org/anthology/W18-64094

Yankovskaya, E., Tättar, A., & Fishel, M. (2019). Quality estimation and translation metrics via pre-trained word and sentence embeddings. *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, 103–107. http://www.aclweb.org/anthology/W19-5410

Yankovskaya, L., Tars, M., Tättar, A., & Fishel, M. (2023). Machine translation for low-resource Finno-Ugric languages. *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, 762–771.

YLE. (2023). *Tarton yliopiston käännöskone tukee nyt 17 uutta suomalais-ugrilaista kieltä* [Accessed: 12.02.2025]. https://yle.fi/a/74-20125964

Zhang, B., Williams, P., Titov, I., & Sennrich, R. (2020). Improving massively multilingual neural machine translation and zero-shot translation. In D. Jurafsky, J. Chai, N. Schluter, & J. Tetreault (Eds.), *Proceedings of the 58th annual meeting of the association for computational linguistics*. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.148

ACKNOWLEDGEMENTS

I want to say a huge thank you to my supervisor, Mark Fišel. Your help and guidance made this research possible. During the PhD process, I would say that we became friends, not just people who worked together. You have been a great teacher and supporter through this PhD journey. I have especially enjoyed our back-and-forth discussions on science, applications, and life in general.

I also want to thank my coworker Liisa Rätsep for all the professional and mental help and support. Liisa and Mark also introduced me to coffee; now, I am hooked for life.

A special thanks go to Taido and Maali who made a lot of my ideas possible by choosing me to supervise them for their Bachelor's and Master's theses, project work (mainly MTee), and research work, which was a big part of this thesis as well. I am very thankful for your hard work and dedication.

On a personal note, I couldn't have achieved this without the love and support of my wife, Annika Laumets-Tättar. You always gave me strength and motivation to keep going and reach my dreams. I would also like to say thanks to my daughter Amelia, who has been a lovely distraction in the final dissertation writing. I love you both.

I want to thank my family and friends who have supported me along the way, coming from a small town near the border of Estonia, I never imagined I would achieve a PhD degree. Knowing how proud my family is of me has been a huge motivation. I am very grateful to have you all in my life.

I should also confess that several AI assistants (ChatGPT, Claude, Gemini) and the trusty Grammarly gave me some extra editorial eyes. While they helped smooth out my rough writing, all ideas and final decisions are, for better or worse, my own.

Finally, thanks to everyone who has been a part of this journey; your support means the world to me. So, a big thanks go to all the collaborators, mentors, and institutions that supported this research.

SISUKOKKUVÕTE

Mitmekeelne masintõlge ressursivaestele keeltele

Käesolev doktoritöö uurib, kuidas integreerida ressursivaesed ja keerukad soome-ugri keeled tänapäevastesse mitmekeelsetesse masintõlkesüsteemidesse. Masintõlge (MT) on tehnoloogia, mis kasutab tarkvara ja tehisintellekti teksti või kõne automaatseks tõlkimiseks ühest keelest teise.

Soome-ugri keelkonda kuuluvad riiklikud keeled nagu ungari, soome ja eesti, kuid selles peres on kokku 42 keelt, millest paljud on väljasuremisohus, nagu näiteks Liivi keel, millel on ainult üks emakeelekõneleja, ja mitmed Saami keeled, millel on kümmekond või sadakond kõnelejat. Selle töö eesmärk on pakkuda nendele väiksematele keeltele avatud lähtekoodiga töötavaid masintõlkemudeleid.

Sissejuhatus

Masintõlge on kiiresti arenev valdkond, mis on toonud olulisi teaduslikke edusamme loomuliku keele töötlemise ja tehisintellekti vallas. Transformer (Vaswani *et al.*, 2017), mis on praegu väga populaarne tänu GPT-laadsetele ja suurtele keelemudelitele, kasvas välja teadustöö käigus masintõlke valdkonnas. Minu doktorantuuri algus langes kokku transformeerimismudeli läbimurdelise avaldamisega ning käesolev väitekiri käsitleb *Transformer*-ite kasutamist mitmekeelses närvivõrgupõhises masintõlkes madala kuni kõrge ressursitasemega stsenaariumides. Ressursivaesteks loetakse keeli, millel on vähe paralleelseid tekstiandmeid (vähem kui 100K lausepaari).

Masintõlge (MT) viitab tarkvara kasutamisele teksti või kõne tõlkimiseks ühest keelest teise. MT eesmärk on automaatselt toota kõrgekvaliteedilisi tõlkeid, mis on võrreldavad inimtõlkijate loodutega. Tabelis:1 on toodud masintõlke näited ning illustreerib keelte ja tõlkesüsteemide nüansse.

Motivatsioon

Uurimistöö keskendub mitmekeelsetele mudelitele, eelkõige motiveeritud ressursside piiratusest. Peamised põhjused on järgmised:

- Kuluefektiivsus: Masintõlkemudelite treenimine ja juurutamine, ehk teisisõnu teenuse üleval hoidmine avalikult veebis kättesaadavalt, on kallis. Enne (Johnson *et al.*, 2017) oli standardne lähenemine masintõlkele ühe suunalised mudelid. Ühe suunalised mudelid on sellised, kus tõlkesüsteem võtab sisse ainult ühekeelsed laused ja tõlgib need teise keelde. See tähendav, et täieliku mitmekeelse tõlkesüsteemi jaoks, mis toetab seitset keelt, on vaja treenida ja pilves majutada 42 erinevat mudelit (kõik kombinatsioonid, $7 \cdot 6$), üks iga tõlkesuuna jaoks. See on äärmiselt kallis, ei skaleeru hästi ja nõuab palju monitoorimist.

- Ühtne süsteem: Alternatiivina ühe-suunaliste mudelitele on meie eesmärk olnud treenida ja juurutada ühtset mitmekeelset süsteemi, mille maht oleks sama suur kui ühe suunaga mudelil. Võrdlusmomendina, üks mitmekeelne mudel on parameetrite poolest peaaegu keskmiselt kuni 2 korda rohkem parameetreid hõlmav, aga üksikuid mudeleid on 42. Iga mudeli pilve panemine on kallis, hind on ühe mitmekeelse ja ühe-suunalise mudeli puhul sama, ehk sisuliselt on mitmekeelse mudeliga 40-korda soodsam.
- Keelte sarnasus: Keeled jagavad sageli sarnaseid tunnuseid ja sõnavorme (eriti sama keelkonna keeled)– seega on mõistlik uurida mudelite võimet õppida mitmekeelseid representatsioone, et parandada tõlkemudeli üldistamise võimet.
- Madala ressursiga tõlke kvaliteet: Eesmärk on kasutada kõiki võimalikke andmeid, et parandada masintõlke toimivust keelesuundades, kus on väga vähe andmeid saadaval.

Selle väitekirja läbiv teema on, et enamik katseid on keskendunud eesti keelele, sest üldine eesmärk on olnud eesti masintõlke kvaliteedi edendamine. Eesti keel on huvitav, sest koos teiste keeltega saame mõned kõrge ressursiga keelepaarid (eesti-inglise, eesti-saksa), mõned keskmise ressursiga keelepaarid (eesti-tšehhi, eesti-itaalia) ja palju madala (eesti-araabia, eesti-võru) või paralleelsete andmeteta keelepaare (eesti-põhjasaami).

Lisaks eesti keelele esitab soome-ugri keelte grupp ainulaadse väljakutse seoses ressursipiirangutega. Keeled võivad näidata märkimisväärseid variatsioone murretes, kirjutamissüsteemides (ortograafia) ja isegi geograafiliselt lähedastes piirkondades. Näiteks võru keelel, mida räägitakse Lõuna-Eestis, on erinevad ortograafilised variatsioonid sõltuvalt asukohast. See väitekiri püüab arendada mitmekeelset masintõlkesüsteemi, mis suudab lahendada ressursinappuse probleeme soome-ugri keelte jaoks. Lisaks kasutavad paljud soome-ugri keeled, nagu komi või mari, ladina tähestiku asemel kirillitsat, mis muudab üldistamise keerulisemaks.

Kõiki motivatsioone arvesse võttes jõuame uurimistöö üldise teemani, kus me püüame leida ühte mitmekeelset masintõlkesüsteemi, mis on tõhus (andmete ja arvutusressursside mõttes), saavutades samal ajal optimaalse tõlketäpsuse.

Peamised tulemused

Käesolev doktoritöö annab põhjaliku ülevaate mitmekeelsete masintõlkemudelite kasutamisest ressursivaestele keeltele. Esiteks vaadatakse universaalseid jagatud parameetritega mudeleid, mis näitasid, et kõrgema ressursiga keelte kasutamine võib aidata parandada madala ressursiga keelte tõlke kvaliteeti. Modulaarne lähenemine, mis keskendub keelespetsiifilistele enkooder-dekooder paaridele, parandas tõlke tulemusi võrreldes universaalsete mudelitega suuremate ressurssidega keeltele. Suurte eeltreenitud mudelite peenhäälastamine näitas märkimisväärset kvaliteedi tõusu, eriti ressursivaese stsenaariumi korral. Töö esitab uusi

metoodikaid, nagu suurte eeltreenitud mudelite edasitreening seninägemata keelte jaoks ja teadmusjagamise tehnikate rakendamist lähedaste keelte vahel. Teadustöö kirjeldab, kuidas esmakordselt loodi mitmekeelne soome-ugri mudel, mis algselt hõlmas viit soome-ugri keelt, aga doktoritöö lõpus toetas meie avalikult kättesaadav platvorm juba 23 keelt, sealhulgas ohustatud keeli nagu liivi keel.

3. peatükk: Universaalsed mitmekeelsed masintõlkemudelid. Kolmandas peatükis käsitletakse universaalseid mitmekeelseid masintõlkemudeleid, kus üks mudel suudab töödelda mitut keelt ühiste parameetrite abil. Peatükk esitleb kahte lähenemist, millset mõlemad kasutavad universaalset mudeli arhitektuuri, ehk mudelit, kus kõik tehisnärvivõrgu parameetrid on jagatud ja keelespetsiifilisi mooduleid ei ole.

4. peatükk: Modulaarsed masintõlkemudelid. Neljandas peatükkis uuritakse modulaarset lähenemist, kus mitmekeelses süsteemis kasutatakse keelespetsiifilisi enkoodereid ja dekoodereid. Keelespetsiifiliste moodulite motivatsioon tuleneb sellest, et need võiksid ühe keele väga hästi ära õppida. Teine artikkel peatükis tutvustab modulaarse mudeli edasiarendust, mis ühendab ja võrdleb erinevaid mitmekeelsusele suunatud masintõlkesüsteemide arhitektuuri lähenemisi.

5. peatükk: Suured eeltreenitud (keele-)mudelid. Viiendas peatükis keskendutakse suurte eeltreenitud mudelite järeltreenimisele (ehk peenhäälestamisele) tõlkes, demonstreerides nende peenhäälastamise efektiivsust spetsiifiliste keelepaaride jaoks. Peatükis esitatakse kolm teadusartiklit, mis igaüks näitavad hüppeid tõlke kvaliteedis. Näidatakse, kuidas suurte keelemudelite peenhäälastamine võimaldab head tõlkekvaliteeti isegi selliste keelepaaride puhul, millel on väga piiratud paralleelkorpused.

Kokkuvõte

Käesolev töö annab põhjaliku ülevaate mitmekeelsete masintõlkemudelite kasutamisest ressursivaestele keeltele, näidates erinevate strateegiate tõhusust. Töö keskendub eesti ja soome-ugri keeltele ning töö tulemusel on valminud väga tugevad eesti keele tõlkemudelid ning esimesed mitmekeelsed soome-ugri tõlkemudelid. Eriti tähelepanuväärne on suurte eeltreenitud mudelite kasutamine, mis võimaldab saavutada kõrge kvaliteediga tõlkesüsteemid isegi väga piiratud andmetega keelte puhul.

Tuleviku perspektiivid ja mõju

Töö tulemused pakuvad tugeva aluse olulisteks edasisteks sammudeks masintõlkes ressursivaeste keelte jaoks globaalselt. Edasine uurimistöö võiks keskenduda teadmiste jagamise parandamisele ja suurte eel-treenitud mudelite järeltreenimise tõhustamise täiendamisele. Reaalses maailmas aitavad need leiud soome-ugri kogukondadel paremini suhelda ja säilitada oma keelelist pärandit.

PUBLICATIONS

CURRICULUM VITAE

Personal data

Name:	Andre Tättar
Birth:	13.07.1993
Nationality:	Estonian
Contact:	andre.tattar@gmail.com

Education

2018–2025	PhD in Computer Science, Institute of Computer Science, University of Tartu, Estonia
2016–2018	MSc in Computer Science, Institute of Computer Science, University of Tartu, Estonia
2012–2016	BSc in Computer Science, Institute of Computer Science, University of Tartu, Estonia
2000–2012	Valga Gymnasium, Estonia

Employment

2023–...	Head of Artificial Intelligence, Impactly OÜ
2020–2024	Junior Research Fellow in Natural Language Processing, Institute of Computer Science, University of Tartu
2022–2023	Machine Learning Researcher, Klaus
2017–2019	Research and Teaching Assistant in Natural Language Processing, Institute of Computer Science, University of Tartu

Scientific work

Main fields of interest:

- Machine Translation
- Large Language Models
- Natural Language Processing
- Artificial Intelligence

ELULOOKIRJELDUS

Isikuandmed

Nimi:	Andre Tättar
Sünniaeg:	13.07.1993
Kodakondsus:	Eesti
E-post:	andre.tattar@gmail.com

Haridus

2018–2025	Doktoriõpe informaatika erialal, Arvutiteaduse instituut, Tartu Ülikool, Eesti
2016–2018	Magistriõpe informaatika erialal, Arvutiteaduse instituut, Tartu Ülikool, Eesti
2012–2016	Bakalaureuseõpe informaatika erialal, Arvutiteaduse instituut, Tartu Ülikool, Eesti
2000-2012	Valga Gümnaasium, Eesti

Teenistuskäik

2023–...	Tehisintellekti valdkonna juht, Impactly OÜ
2020–2024	Nooremteadur loomulike keele töötluse alal, Arvutiteaduse Instituut, Tartu Ülikool, Eesti
2022–2023	Masinõppe Teadur, Klaus
2017–2019	Teadus- ja õppeassistent loomuliku keele töötluse alal, Arvutiteaduse Instituut, Tartu Ülikool, Eesti

Teadustegevus

Peamised uurimisvaldkonnad:

- Masintõlge
- Suured Keelemudelid
- Loomuliku Keele Töötlus
- Tehisintellekt

DISSERTATIONES INFORMATICAE PREVIOUSLY PUBLISHED IN DISSERTATIONES MATHEMATICAE UNIVERSITATIS TARTUENSIS

19. **Helger Lipmaa.** Secure and efficient time-stamping systems. Tartu, 1999, 56 p.
22. **Kaili Müürisep.** Eesti keele arvutigrammatika: süntaks. Tartu, 2000, 107 lk.
23. **Varmo Vene.** Categorical programming with inductive and coinductive types. Tartu, 2000, 116 p.
24. **Olga Sokratova.** Ω-rings, their flat and projective acts with some applications. Tartu, 2000, 120 p.
27. **Tiina Puolakainen.** Eesti keele arvutigrammatika: morfoloogiline ühestamine. Tartu, 2001, 138 lk.
29. **Jan Villemson.** Size-efficient interval time stamps. Tartu, 2002, 82 p.
45. **Kristo Heero.** Path planning and learning strategies for mobile robots in dynamic partially unknown environments. Tartu 2006, 123 p.
49. **Härmel Nestra.** Iteratively defined transfinite trace semantics and program slicing with respect to them. Tartu 2006, 116 p.
53. **Marina Issakova.** Solving of linear equations, linear inequalities and systems of linear equations in interactive learning environment. Tartu 2007, 170 p.
55. **Kaarel Kaljurand.** Attempto controlled English as a Semantic Web language. Tartu 2007, 162 p.
56. **Mart Anton.** Mechanical modeling of IPMC actuators at large deformations. Tartu 2008, 123 p.
59. **Reimo Palm.** Numerical Comparison of Regularization Algorithms for Solving Ill-Posed Problems. Tartu 2010, 105 p.
61. **Jüri Reimand.** Functional analysis of gene lists, networks and regulatory systems. Tartu 2010, 153 p.
62. **Ahti Peder.** Superpositional Graphs and Finding the Description of Structure by Counting Method. Tartu 2010, 87 p.
64. **Vesal Vojdani.** Static Data Race Analysis of Heap-Manipulating C Programs. Tartu 2010, 137 p.
66. **Mark Fišel.** Optimizing Statistical Machine Translation via Input Modification. Tartu 2011, 104 p.
67. **Margus Niitsoo**. Black-box Oracle Separation Techniques with Applications in Time-stamping. Tartu 2011, 174 p.
71. **Siim Karus.** Maintainability of XML Transformations. Tartu 2011, 142 p.
72. **Margus Treumuth.** A Framework for Asynchronous Dialogue Systems: Concepts, Issues and Design Aspects. Tartu 2011, 95 p.
73. **Dmitri Lepp.** Solving simplification problems in the domain of exponents, monomials and polynomials in interactive learning environment T-algebra. Tartu 2011, 202 p.

74. **Meelis Kull.** Statistical enrichment analysis in algorithms for studying gene regulation. Tartu 2011, 151 p.
77. **Bingsheng Zhang.** Efficient cryptographic protocols for secure and private remote databases. Tartu 2011, 206 p.
78. **Reina Uba.** Merging business process models. Tartu 2011, 166 p.
79. **Uuno Puus.** Structural performance as a success factor in software development projects – Estonian experience. Tartu 2012, 106 p.
81. **Georg Singer.** Web search engines and complex information needs. Tartu 2012, 218 p.
83. **Dan Bogdanov.** Sharemind: programmable secure computations with practical applications. Tartu 2013, 191 p.
84. **Jevgeni Kabanov.** Towards a more productive Java EE ecosystem. Tartu 2013, 151 p.
87. **Margus Freudenthal.** Simpl: A toolkit for Domain-Specific Language development in enterprise information systems. Tartu, 2013, 151 p.
90. **Raivo Kolde.** Methods for re-using public gene expression data. Tartu, 2014, 121 p.
91. **Vladimir Šor.** Statistical Approach for Memory Leak Detection in Java Applications. Tartu, 2014, 155 p.
92. **Naved Ahmed.** Deriving Security Requirements from Business Process Models. Tartu, 2014, 171 p.
94. **Liina Kamm.** Privacy-preserving statistical analysis using secure multi-party computation. Tartu, 2015, 201 p.
100. **Abel Armas Cervantes.** Diagnosing Behavioral Differences between Business Process Models. Tartu, 2015, 193 p.
101. **Fredrik Milani.** On Sub-Processes, Process Variation and their Interplay: An Integrated Divide-and-Conquer Method for Modeling Business Processes with Variation. Tartu, 2015, 164 p.
102. **Huber Raul Flores Macario.** Service-Oriented and Evidence-aware Mobile Cloud Computing. Tartu, 2015, 163 p.
103. **Tauno Metsalu.** Statistical analysis of multivariate data in bioinformatics. Tartu, 2016, 197 p.
104. **Riivo Talviste.** Applying Secure Multi-party Computation in Practice. Tartu, 2016, 144 p.
108. **Siim Orasmaa.** Explorations of the Problem of Broad-coverage and General Domain Event Analysis: The Estonian Experience. Tartu, 2016, 186 p.
109. **Prastudy Mungkas Fauzi.** Efficient Non-interactive Zero-knowledge Protocols in the CRS Model. Tartu, 2017, 193 p.
110. **Pelle Jakovits.** Adapting Scientific Computing Algorithms to Distributed Computing Frameworks. Tartu, 2017, 168 p.
111. **Anna Leontjeva.** Using Generative Models to Combine Static and Sequential Features for Classification. Tartu, 2017, 167 p.
112. **Mozhgan Pourmoradnasseri.** Some Problems Related to Extensions of Polytopes. Tartu, 2017, 168 p.

113. **Jaak Randmets.** Programming Languages for Secure Multi-party Computation Application Development. Tartu, 2017, 172 p.
114. **Alisa Pankova.** Efficient Multiparty Computation Secure against Covert and Active Adversaries. Tartu, 2017, 316 p.
116. **Toomas Saarsen.** On the Structure and Use of Process Models and Their Interplay. Tartu, 2017, 123 p.
121. **Kristjan Korjus.** Analyzing EEG Data and Improving Data Partitioning for Machine Learning Algorithms. Tartu, 2017, 106 p.
122. **Eno Tõnisson.** Differences between Expected Answers and the Answers Offered by Computer Algebra Systems to School Mathematics Equations. Tartu, 2017, 195 p.

DISSERTATIONES INFORMATICAE UNIVERSITATIS TARTUENSIS

1. **Abdullah Makkeh**. Applications of Optimization in Some Complex Systems. Tartu 2018, 179 p.
2. **Riivo Kikas**. Analysis of Issue and Dependency Management in Open-Source Software Projects. Tartu 2018, 115 p.
3. **Ehsan Ebrahimi**. Post-Quantum Security in the Presence of Superposition Queries. Tartu 2018, 200 p.
4. **Ilya Verenich**. Explainable Predictive Monitoring of Temporal Measures of Business Processes. Tartu 2019, 151 p.
5. **Yauhen Yakimenka**. Failure Structures of Message-Passing Algorithms in Erasure Decoding and Compressed Sensing. Tartu 2019, 134 p.
6. **Irene Teinemaa**. Predictive and Prescriptive Monitoring of Business Process Outcomes. Tartu 2019, 196 p.
7. **Mohan Liyanage.** A Framework for Mobile Web of Things. Tartu 2019, 131 p.
8. **Toomas Krips.** Improving performance of secure real-number operations. Tartu 2019, 146 p.
9. **Vijayachitra Modhukur.** Profiling of DNA methylation patterns as biomarkers of human disease. Tartu 2019, 134 p.
10. **Elena Sügis.** Integration Methods for Heterogeneous Biological Data. Tartu 2019, 250 p.
11. **Tõnis Tasa.** Bioinformatics Approaches in Personalised Pharmacotherapy. Tartu 2019, 150 p.
12. **Sulev Reisberg.** Developing Computational Solutions for Personalized Medicine. Tartu 2019, 126 p.
13. **Huishi Yin.** Using a Kano-like Model to Facilitate Open Innovation in Requirements Engineering. Tartu 2019, 129 p.
14. **Faiz Ali Shah.** Extracting Information from App Reviews to Facilitate Software Development Activities. Tartu 2020, 149 p.
15. **Adriano Augusto**. Accurate and Efficient Discovery of Process Models from Event Logs. Tartu 2020, 194 p.
16. **Karim Baghery.** Reducing Trust and Improving Security in zk-SNARKs and Commitments. Tartu 2020, 245 p.
17. **Behzad Abdolmaleki.** On Succinct Non-Interactive Zero-Knowledge Protocols Under Weaker Trust Assumptions. Tartu 2020, 209 p.
18. **Janno Siim.** Non-Interactive Shuffle Arguments. Tartu 2020, 154 p.
19. **Ilya Kuzovkin.** Understanding Information Processing in Human Brain by Interpreting Machine Learning Models. Tartu 2020, 149 p.
20. **Orlenys López Pintado.** Collaborative Business Process Execution on the Blockchain: The Caterpillar System. Tartu 2020, 170 p.
21. **Ardi Tampuu.** Neural Networks for Analyzing Biological Data. Tartu 2020, 152 p.

22. **Madis Vasser.** Testing a Computational Theory of Brain Functioning with Virtual Reality. Tartu 2020, 106 p.
23. **Ljubov Jaanuska.** Haar Wavelet Method for Vibration Analysis of Beams and Parameter Quantification. Tartu 2021, 192 p.
24. **Arnis Parsovs.** Estonian Electronic Identity Card and its Security Challenges. Tartu 2021, 214 p.
25. **Kaido Lepik.** Inferring causality between transcriptome and complex traits. Tartu 2021, 224 p.
26. **Tauno Palts.** A Model for Assessing Computational Thinking Skills. Tartu 2021, 134 p.
27. **Liis Kolberg.** Developing and applying bioinformatics tools for gene expression data interpretation. Tartu 2021, 195 p.
28. **Dmytro Fishman.** Developing a data analysis pipeline for automated protein profiling in immunology. Tartu 2021, 155 p.
29. **Ivo Kubjas.** Algebraic Approaches to Problems Arising in Decentralized Systems. Tartu 2021, 120 p.
30. **Hina Anwar.** Towards Greener Software Engineering Using Software Analytics. Tartu 2021, 186 p.
31. **Veronika Plotnikova.** FIN-DM: A Data Mining Process for the Financial Services. Tartu 2021, 197 p.
32. **Manuel Camargo.** Automated Discovery of Business Process Simulation Models From Event Logs: A Hybrid Process Mining and Deep Learning Approach. Tartu 2021, 130 p.
33. **Volodymyr Leno.** Robotic Process Mining: Accelerating the Adoption of Robotic Process Automation. Tartu 2021, 119 p.
34. **Kristjan Krips.** Privacy and Coercion-Resistance in Voting. Tartu 2022, 173 p.
35. **Elizaveta Yankovskaya.** Quality Estimation through Attention. Tartu 2022, 115 p.
36. **Mubashar Iqbal.** Reference Framework for Managing Security Risks Using Blockchain. Tartu 2022, 203 p.
37. **Jakob Mass.** Process Management for Internet of Mobile Things. Tartu 2022, 151 p.
38. **Gamal Elkoumy.** Privacy-Enhancing Technologies for Business Process Mining. Tartu 2022, 135 p.
39. **Lidia Feklistova.** Learners of an Introductory Programming MOOC: Background Variables, Engagement Patterns and Performance. Tartu 2022, 151 p.
40. **Mohamed Ragab.** Bench-Ranking: A Prescriptive Analysis Approach for Large Knowledge Graphs Query Workloads. Tartu 2022, 158 p.
41. **Mohammad Anagreh.** Privacy-Preserving Parallel Computations for Graph Problems. Tartu 2023, 181 p.
42. **Rahul Goel.** Mining Social Well-being Using Mobile Data. Tartu 2023, 104 p.
43. **Anti Ingel.** Algorithms using information theory: classification in brain-computer interfaces and characterising reinforcement-learning agents. Tartu 2023, 142 p.

44. **Shakshi Sharma.** Fighting Misinformation in the Digital Age: A Comprehensive Strategy for Characterizing, Identifying, and Mitigating Misinformation on Online Social Media Platforms. Tartu 2023, 158 p.
45. **Kristiina Rahkema.** Quality Analysis of iOS Applications with Focus on Maintainability and Security Aspects. Tartu 2023, 182 p.
46. **Ivan Slobozhan.** Studying Online Social Media Engagement in CIS Countries during Protests, Mass Demonstrations and War. Tartu 2023, 81 p.
47. **Nurlan Kerimov.** Building a catalogue of molecular quantitative trait loci to interpret complex trait associations. Tartu 2023, 248 p.
48. **Pavlo Tertychnyi.** Machine Learning Methods for Anti-Money Laundering Monitoring. Tartu 2023, 117 p.
49. **Abasi-amefon Obot Affia.** A Framework and Teaching Approach for IoT Security Risk Management. Tartu 2023, 180 p.
50. **Raimond-Hendrik Tunnel.** Video Game Design and Development Bachelor's Curriculum for Estonia. Tartu 2024, 137 p.
51. **Ahto Salumets.** Bioinformatics analysis of various aspects in immunology. Tartu 2024, 198 p.
52. **Mohammed Abdulhameed Shaif Ali.** Deep Learning Methods for Cell Microscopy Image Analysis. Tartu 2024, 143 p.
53. **Pille Pullonen-Raudvere**. Foundations of Efficient and Secure Algorithm Development for Secure Multiparty Computation. Tartu 2024, 265 p.
54. **Marili Rõõm.** Multiple approaches to learners' success and factors affecting it in computer programming MOOCs. Tartu 2024, 170 p.
55. **Shivananda Rangappa Poojara.** Design and Orchestration of Scalable, Event-Driven Serverless Data Pipelines for Internet of Things (IoT) Applications. Tartu 2024, 172 p.
56. **Hassan Abdulgaleel Hassan Salim Eldeeb.** Empowering Machine Learning Pipelines with Automated Feature Engineering. Tartu 2024, 121 p.
57. **Muhammad Uzair.** Soft decision making for agri-food 4.0. Tartu 2024, 158 p.
58. **Kirill Milintsevich.** Estimation of Depression Level from Text: Symptom-Based Approach, External Knowledge, Dataset Validity. Tartu 2024, 130 p.
59. **Maksym Del.** Multilingual and Multi-Domain Representational Patterns Across Trpansformer-Based Models. Tartu 2024, 131 p.
60. **Kristo Raun.** Adaptive Out-of-order Handling in Streaming Conformance Checking. Tartu 2024, 118 p.
61. **Toivo Vajakas.** Towards integration of mobile network data into analyzing human mobility. Tartu 2024, 103 p.
62. **Katsiaryna Lashkevich.** Data-Driven Analysis and Optimization of Waiting Times in Business Processes. Tartu 2024, 169 p.
63. **Alejandra Duque-Torres.** Classifying, Constraining and Ranking Metamorphic Relations. Tartu 2025, 159 p.
64. **Mariia Bakhtina.** A Method for Information Security and Privacy Management in Smart Solutions. Tartu 2025, 199 p.