universität wien

MASTERARBEIT | MASTER'S THESIS

Titel | Title

Transfer Learning to Improve Estonian Machine Translation: The Role of Genealogical and Areal Relationships

verfasst von | submitted by

Greete Hollas

angestrebter akademischer Grad | in partial fulfilment of the requirements for the degree of

Master of Science (MSc)

Wien | Vienna, 2025

Studienkennzahl lt. Studienblatt | Degree programme code as it appears on the student record sheet:	UA 066 587
Studienrichtung lt. Studienblatt | Degree programme as it appears on the student record sheet:	Joint-Masterstudium Multilingual Technologies
Betreut von | Supervisor:	Assoz. Prof. Mag. Dr. Dagmar Gromann BSc

Abstract

Machine translation has advanced substantially with the rise of neural machine translation. However, these models rely heavily on large parallel corpora, which disadvantages low-resource languages. Estonian, a morphologically complex Uralic language, exemplifies this challenge: despite a stable online presence, it lacks sufficient annotated and parallel data for effective neural machine translation. While transfer learning has been applied to low-resource machine translation, little attention has been paid to how genealogical and areal linguistic relationships influence transfer effectiveness.

This thesis investigates how linguistic similarity can improve Estonian machine translation through transfer learning. Specifically, it examines the impact of genealogical proximity and areal relationships on translation performance. The experiments use the pre-trained NLLB-200 model and are conducted in four stages: zero-shot baseline evaluation, English–Estonian fine-tuning, cross-lingual fine-tuning with related and contact languages, and comparative analysis.

The results show that genealogically closer languages, particularly Finnish, yield the most effective transfer, confirming the role of shared grammatical and structural features. Areal languages also provide measurable benefits, especially when lexical overlap exists due to historical contact. Combining these perspectives highlights how both genealogical and areal factors shape transfer learning outcomes in low-resource settings.

The contributions of this thesis are threefold: (1) it provides a systematic analysis of genealogical and areal influences in transfer learning for Estonian machine translation; (2) it demonstrates that linguistic relationships can complement data-driven methods in improving translation quality; and (3) it offers insights applicable to other Uralic and contact languages with limited resources, supporting broader efforts toward linguistic diversity in machine translation.

Zusammenfassung

Die maschinelle Übersetzung hat mit dem Aufstieg der neuronalen maschinellen Übersetzung erhebliche Fortschritte erzielt. Diese Modelle sind jedoch stark auf große parallele Korpora angewiesen, sodass Sprachen mit geringen Ressourcen benachteiligt werden. Estnisch, eine morphologisch komplexe uralische Sprache, veranschaulicht diese Herausforderung: Trotz einer stabilen Online-Präsenz fehlen ausreichende annotierte und parallele Daten für eine effektive neuronale maschinelle Übersetzung. Während Transferlernen bereits auf maschinelle Übersetzungen mit geringen Ressourcen angewendet wurde, wurde bisher wenig Aufmerksamkeit darauf gerichtet, wie genealogische und areale sprachliche Beziehungen die Transferwirksamkeit beeinflussen.

Diese Arbeit untersucht, wie sprachliche Ähnlichkeit die estnische maschinelle Übersetzung durch Transferlernen verbessern kann. Konkret wird der Einfluss genealogischer Nähe und arealer Beziehungen auf die Übersetzungsleistung untersucht. Die Experimente verwenden das vortrainierte NLLB-200-Modell und werden in vier Stufen durchgeführt: einer Zero-Shot-Baseline-Bewertung, einer Feinabstimmung für Englisch–Estnisch, einer sprachübergreifenden Feinabstimmung mit verwandten und Kontaktsprachen sowie einer vergleichenden Analyse.

Die Ergebnisse zeigen, dass genealogisch näher verwandte Sprachen, insbesondere Finnisch, den effektivsten Transfer erzielen und dadurch die Rolle gemeinsamer grammatikalischer und struktureller Merkmale bestätigt wird. Auch areale Sprachen bieten messbare Vorteile, insbesondere wenn aufgrund historischer Sprachkontakte lexikalische Überschneidungen bestehen. Die Kombination dieser Perspektiven verdeutlicht, wie sowohl genealogische als auch areale Faktoren die Ergebnisse des Transferlernens unter ressourcenschwachen Bedingungen beeinflussen.

Diese Arbeit leistet drei wesentliche Beiträge: (1) Sie liefert eine systematische Analyse genealogischer und arealer Einflüsse auf das Transferlernen für die estnische maschinelle Übersetzung; (2) sie zeigt, dass sprachliche Beziehungen datengestützte Methoden bei der Verbesserung der Übersetzungsqualität ergänzen können; und (3) sie bietet Erkenntnisse, die auf andere uralische und Kontaktsprachen mit begrenzten Ressourcen anwendbar sind und unterstützt damit umfassendere Bemühungen um sprachliche Vielfalt in der maschinellen Übersetzung.

Contents

List of Tables

1 Introduction

The field of Machine Translation (MT) has achieved advancements in recent years. At present, Neural Machine Translation (NMT) is the leading technique within the community (Ranathunga et al., 2023). However, the quality of the translation correlates with the amount of parallel data, which means the benefits of NMT can be seen for language pairs without language data concerns (Koehn and Knowles, 2017). Therefore, the majority of MT research has focused on high-resource languages, with English often as the focal point (Joshi et al., 2020). While low-resource translation is argued to be a popular research topic (Zhang and Zong, 2020; Ranathunga et al., 2023), the emphasis on high-resourced languages has left many low-resource languages relatively underexplored.

In addition, researchers have mentioned that the community has not agreed on threshold for low-resource settings (Ranathunga et al., 2023) and this aspect could potentially be seen "[...] as a spectrum of resource availability" (Hedderich et al., 2021, 2547). Estonian is one such language. According to Joshi et al. (2020), who categorized number of languages into six groups based on availability of raw as well as annotated datasets, the Estonian language falls under the categorization *The Rising Stars*, indicating a stable presence of the language and an existing cultural community on the web but a lack of work on collecting labeled data. This understanding is supported by Muischnek (2023), stating that Estonian lacks annotated data. In MT, however, the perception of whether the number of resources is small or large is often determined on the basis of the available parallel data for a language pair. Hence, the size of the parallel corpora is important.

The scarcity of bilingual corpora and linguistic resources presents several challenges for developing effective translation systems for these languages. For instance, low-resource languages tend to have not only limited quantity but also lower quality of available corpora (Kreutzer et al., 2022). Estonian, as a Uralic language with a complex morphological structure and limited bilingual corpora, exemplifies the difficulties faced in Low-Resource Neural Machine Translation (LRNMT). To address these challenges, researchers have explored various approaches in LRNMT, including synthetic data generation (Korotkova and Fishel, 2024) and transfer learning (Nguyen and Chiang, 2017). These techniques have demonstrated improvements in effectiveness, existing research has rarely considered systematically how linguistic relationships, both genealogical and areal, affect the transfer process.

This thesis addresses this gap by focusing on Estonian and investigating the role of

genealogical similarity and areal influence in transfer learning for MT. Genealogical relationships capture structural and grammatical commonalities inherited through shared ancestry, while areal relationships reflect contact-induced similarities that arise from geographic proximity and historical interaction. Understanding how these two dimensions shape transfer learning effectiveness can inform better strategies not only for Estonian but also for other low-resource languages.

The central objective of this thesis is therefore to improve MT quality for Estonian while contributing to the broader goal of supporting linguistic diversity. Specifically, the research seeks to answer the following questions:

1. How does genealogical similarity impact the performance of transfer learning in Estonian machine translation?

2. How do areal relationships and shared lexical influences contribute to transfer learning outcomes in Estonian machine translation?

To address these questions, this thesis evaluates transfer learning strategies using the pre-trained NLLB-200 model (Costa-jussà et al., 2022). Experiments are conducted across four stages: (1) zero-shot baseline translation for English–Estonian, (2) fine-tuning with English–Estonian parallel data, (3) cross-lingual fine-tuning with related and contact languages, and (4) comparative analysis through both quantitative (sacreBLEU, chrF) and qualitative evaluation.

It is assumed that genealogically related languages will yield greater improvements in Estonian MT performance due to shared grammatical, morphological, and lexical characteristics. The extent of this improvement is expected to correlate with the degree of linguistic proximity between the source and target languages. Furthermore, it is hypothesised that areal relationships and language contact may also have a positive effect on transfer learning outcomes, but their impact is expected to be smaller than that of genealogical similarity.

This research is motivated by the challenges faced in LRNMT, where limited parallel data constrains performance. Improving MT for Estonian not only addresses practical needs for speakers of minority languages, but also contributes to broader efforts in supporting linguistic diversity and advancing Natural Language Processing (NLP) techniques for low-resource settings. By investigating the role of genealogical and areal language relationships, this thesis aims to provide insights relevant to researchers of MT systems working with similar languages.

The contributions of this thesis are threefold. First, it provides the first systematic evaluation of genealogical and areal relationships in transfer learning for Estonian MT. Second, it demonstrates how linguistic similarity can complement data-driven approaches in low-resource settings. Third, it offers insights relevant not only for Estonian but also for other Uralic and contact languages with limited resources, such as Livonian or Võro.

The following sections outline the structure of this thesis. Chapter 2 presents the theoretical background, introducing key concepts in machine learning, neural machine translation, low-resource MT, and linguistic classification. Chapter 3 reviews related work, situating this research within existing studies on transfer learning and low-resource translation. Chapter 4 describes the experimental design, including datasets, model selection, fine-tuning procedures, and evaluation methods. Chapter 5 reports the results of the experiments, combining both quantitative and qualitative analyses. Finally, Chapter 6 discusses the findings in relation to the research questions and emphasizes implications, limitations, and future directions, and Chapter 7 summarizes the key contributions of the thesis and suggests directions for further research.

2 Theoretical Background

This chapter outlines the theoretical foundations relevant to the study of MT for low-resource languages. It begins by introducing core concepts in machine learning, followed by a summary of machine translation systems, with particular attention to neural approaches. The chapter then examines strategies specifically suited to low-resource scenarios, including data augmentation, multilingual training, and transfer learning. To contextualize the challenges of low-resource translation, the chapter also explores definitions of data scarcity, key factors influencing resource availability, and the specific case of Estonian as a low-resource language. Finally, it reviews linguistic classification systems and typological relationships, which influence the design and evaluation of NMT models.

2.1 Machine Learning Fundamentals

One of the foundational definitions of learning in the context of Machine Learning (ML) is provided by Mitchell (1997, 2), who defines it as: "A computer program is said to learn from experience E with respect to some class of tasks T, and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E." Fundamentally, ML enables systems to improve their performance on specific tasks by detecting patterns in data through algorithmic and statistical methods, without the need for explicit programming. Instead of writing explicit rules, as in classical programming, ML allows a system to learn the rules from examples. The definition by Mitchell (1997) captures three essential components of the learning concept. First, the task (T) which includes the concepts of what the model is supposed to do. Second, the experience (E) that covers the data that the model learns. Third, to evaluate how well the model performs, a performance measure is included (P). The importance of ML has grown significantly as a result of access to large datasets, increased computing resources, and progress in learning algorithms. This ability to learn and generalize makes ML a fundamental technology in various domains and it is widely applied across multiple fields, such as healthcare, finance, self-driving technologies, and language technologies (Sarker, 2021).

ML operates through a systematic process that begins with the definition of a problem and the identification of relevant input and output variables (Mitchell,

1997). Once a task is clearly formulated, appropriate data is collected, cleaned, and preprocessed to ensure its suitability for learning, as the quality and structure of the input data have an important role in the performance of the learning process (Murphy, 2022). A ML algorithm attempts to discover patterns within this data by fitting a model, which is a mathematical function mapping inputs to outputs (Sarker, 2021). During training, the algorithm updates its internal parameters to reduce a predefined loss function that reflects the gap between predicted and true values. This optimization is typically performed through iterative methods such as gradient descent (Jurafsky and Martin, 2025). The ultimate goal is not only to achieve strong performance on the training data but also to generalize, such as to make accurate predictions on previously unseen data (Mitchell, 1997). Achieving good generalization requires careful consideration of model complexity, the use of validation data for hyperparameter tuning, and regularization techniques that prevent overfitting, which arises when a model captures noise or highly specific characteristics from the training data and thus fails to generalize on unseen inputs (Bishop, 2006). Once trained and evaluated, the model is deployed for inference, where it can provide predictions or decisions on new input data (Sarker, 2021). Throughout this process, components such as data, features, the model, the loss function, and the optimization algorithm interact in a tightly integrated manner to form the basis of ML systems.

ML algorithms are typically grouped into four main types: supervised learning, unsupervised learning, self-supervised learning, and reinforcement learning (Mohammed et al., 2016). These paradigms determine how models learn patterns from data.

Supervised learning is the most prevalent form of ML, where the model learns a mapping f from inputs $x \in X$ to outputs $y \in Y$ based on labeled or annotated data (Sarker, 2021). In this context, the inputs x typically represent a fixed-dimensional vector of numbers. The outputs, y, are known as labels, targets, or responses, and they correspond to the correct answer for a given input. The experience E in supervised learning is provided in the form of a training set consisting of N input-output pairs $D = \{(x_n, y_n)\}_{n=1}^{N}$, with N representing the total sample size in the dataset (Goodfellow et al., 2016). The task is to learn a function that assigns the appropriate output to each input based on the patterns present in the data. The performance of a supervised model is measured using a performance metric, which varies based on whether the output is categorical or continuous. Supervised learning is employed in many different areas, from classification problems such as spam detection and sentiment analysis, as well as regression tasks like stock price prediction (Sarker, 2021). To train the model, the goal is to minimize an error metric, often referred to as a loss function, which measures the difference between predicted and true labels (Jurafsky and Martin, 2025). Widely used

supervised learning techniques are decision trees, linear regression, and support vector machines.

Unlike supervised learning, unsupervised learning does not rely on labeled data. Instead, the goal involves discovering hidden patterns, structures, or representations within the input data X, without access to corresponding target outputs Y (Sarker, 2021). The experience E in this case consists solely of a dataset without annotations or labels, where the learning algorithm seeks to uncover the hidden structure and relationships present in the data (Goodfellow et al., 2016). Since no explicit labels are available, there is no direct supervision guiding the learning process, making the definition of a performance measure P more complex and often task-dependent. One of the main tasks in unsupervised learning is clustering, where the objective is to group similar instances together based on some similarity measure (Sarker, 2021). Other key tasks, mentioned by Sarker (2021), are dimensionality reduction and anomaly detection. The first aims to compress high-dimensional data to a lower-dimensional form while preserving its core structure. This is commonly used in data visualization or preprocessing for supervised tasks. Anomaly detection focuses on identifying outliers or rare events within a dataset. Frequently applied algorithms in unsupervised learning are k-means clustering, autoencoders, hierarchical clustering, Principal Component Analysis (PCA), and Gaussian Mixture Models (GMMs) (Sarker, 2021).

In recent years, self-supervised learning, a subclass of unsupervised learning, has gained importance as a scalable alternative to traditional supervised learning, as it generates supervisory signals from the data itself through pretext tasks (Gui et al., 2024). These tasks allow models to learn useful representations by predicting parts of the data from other parts (Liu et al., 2023). For instance, in masked language modeling, as used in BERT (Devlin et al., 2019), the model learns to infer missing tokens within a sentence from their surrounding context. These objectives enable models to acquire general-purpose representations from raw inputs, without requiring manually annotated data (Liu et al., 2023).

Semi-supervised learning integrates aspects of supervised and unsupervised approaches by making use of a small labeled dataset together with a much larger unlabeled dataset (Mohammed et al., 2016). This approach proves especially useful in situations where obtaining labeled data requires significant resources or time, whereas unlabeled data can be accessed easily (Sarker, 2021). The model leverages the labeled examples to learn initial patterns, which are then generalized across the unlabeled data. According to Sarker (2021), applications of semi-supervised learning are common in areas like MT and labeling data.

Reinforcement Learning (RL), in contrast, is a learning paradigm in which an agent engages with an environment and improves decision-making over time through feedback given as rewards or penalties (Mitchell, 1997). The objective involves

learning a policy that maximizes cumulative reward or minimize the risk over time (Mohammed et al., 2016). Unlike supervised learning, which requires labeled input-output pairs, RL focuses on learning from delayed and often sparse feedback (Mitchell, 1997). RL is widely applied across domains like robotics and autonomous systems, in which the learning process involves exploration, exploitation, and long-term planning under uncertainty (Sarker, 2021). Additionally, the integration of neural networks with RL, known as Deep Reinforcement Learning, enables agents to handle complex, high-dimensional input spaces by approximating value functions or policies with deep neural networks (Mnih et al., 2015).

Neural networks constitute a class of models that are inspired by how biological neural systems operate and are structured (Mitchell, 1997). At a high level, they consist of interconnected layers composed of artificial neurons that process inputs by computing weighted sums and applying non-linear activation operations (Jurafsky and Martin, 2025). These non-linear activation functions, such as the sigmoid, hyperbolic tangent (tanh), or Rectified Linear Unit (ReLU), add non-linear characteristics to the model, allowing it to learn and represent complex, non-linear relationships in data. Without these non-linearities, even deep networks composed of multiple layers would be mathematically equivalent to a single-layer linear transformation, limiting the expressive capacity of the model (Jurafsky and Martin, 2025). The architecture typically includes an input layer that receives raw data, one or more hidden layers that progressively convert the input into higher-level features, followed by an output layer that generates predictions (Mitchell, 1997). Formally, a neural network represents a compositional function $f(x; \theta)$, where θ denotes the set of weights and biases that are learned from data (Goodfellow et al., 2016).

Training a neural network involves optimizing these parameters to reduce a task-specific loss function $\mathcal{L}(y, \hat{y})$, where y is the ground truth and $\hat{y}$ is the model's prediction (Goodfellow et al., 2016). This process is commonly carried out using Stochastic Gradient Descent (SGD) or its variants, which rely on the backpropagation algorithm to efficiently compute gradients (Jurafsky and Martin, 2025). The expressiveness of neural networks allows them to model non-linear relationships in data, which makes them suitable for high-dimensional and unstructured inputs, such as audio, text, and images (Goodfellow et al., 2016).

Deep learning, a subfield of ML centered on deep neural networks with many layers, has revolutionized various domains by enabling end-to-end learning systems that automatically extract features from raw data (Jurafsky and Martin, 2025). Unlike traditional ML pipelines, which rely heavily on manual feature engineering, models in deep learning, such as Convolutional Neural Networks (CNNs) (Lecun et al., 1998), Recurrent Neural Networks (RNNs) (Elman, 1990), and Transformers (Vaswani et al., 2017), automatically extract multi-level features directly from data

(Sarker, 2021).

CNNs are a specialized type of feedforward neural network designed to automatically extract spatial features from structured data such as images (Lecun et al., 1998). They benefit from three key properties: local connectivity, weight sharing, and downsampling through pooling layers (Li et al., 2022). Local connectivity restricts neuron connections to small input regions, reducing parameters and improving training efficiency, while weight sharing applies the same convolutional kernels across spatial locations, further decreasing model complexity. Pooling layers reduce spatial dimensionality by summarizing feature maps, thereby minimizing redundant information and overfitting risk. A typical CNN model consists of convolutional layers that extract hierarchical features via kernels, with padding to preserve spatial dimensions and stride controlling sampling density, while modern CNN architectures often include fully connected layers and softmax outputs for classification tasks (Li et al., 2022). Due to their ability to efficiently model spatial hierarchies, CNNs have become foundational in computer vision and have also been applied in certain NLP tasks such as sentence and text classification (Kim, 2014).

RNNs are a class of neural network developed to handle sequential data, which enables their use in applications such as language modeling, MT, and speech recognition (Elman, 1990). Unlike feedforward networks, RNNs employ loops in their architecture that allow information to persist across time steps. At each step in a sequence, the network takes both the current input and the hidden state from the previous step as input, allowing it to learn temporal dependencies (Cho et al., 2014b). However, standard RNNs have difficulty maintaining information over long sequences because of issues like vanishing and exploding gradients (Bengio et al., 1994). To address this, advanced variants such as Long Short-Term Memory (LSTM) networks (Hochreiter and Schmidhuber, 1997a) and Gated Recurrent Units (GRUs) (Cho et al., 2014a) use gating components to manage how information is updated and maintained, allowing the network to preserve relevant context across extended sequences. These enhancements have improved RNN performance in many NLP tasks, including MT, sentiment analysis, and named entity recognition (Sutskever et al., 2014; Jozefowicz et al., 2016; Wang et al., 2016). Despite their strengths, RNNs process data sequentially, limiting parallelization and efficiency, which motivated the shift toward architectures like Transformers (Vaswani et al., 2017).

Transformers are a neural network architecture specifically designed to model sequence data without relying on recurrence or convolutions (Vaswani et al., 2017). The model introduced by Vaswani et al. (2017) is based entirely on self-attention mechanisms, which enables it to determine the relevance of each token with respect to all other tokens in the sequence, regardless of their position. The main components of the Transformer are multi-head self-attention layers, position-wise

feedforward layers, and layer normalization, all organized into stacks of encoder and decoder blocks (Liu et al., 2020a). Each encoder layer transforms the input sequence by attending to all positions in parallel, while decoder layers generate the output sequence by considering both the tokens generated so far and the representations produced by the encoder. To account for word order, Transformers incorporate positional encodings, which integrate positional information into the input embeddings (Vaswani et al., 2017). This architecture enables parallel computation, making Transformers significantly faster to train and more scalable than RNN-based models (Tay et al., 2022).

The Transformer architecture marked a breakthrough in NLP, outperforming RNNs and CNNs in a range of tasks including MT, summarization, and language modeling (Vaswani et al., 2017). One of its key advantages is the ability to model long-range dependencies without the limitations of vanishing gradients or sequential processing. By enabling direct connections between all tokens in a sequence, Transformers are better equipped to capture global context. This design laid the foundation for large pretrained language models such as BERT (Devlin et al., 2019), Generative Pre-trained Transformer (GPT) (Radford et al., 2018), and Text-to-Text Transfer Transformer (T5) (Raffel et al., 2023), which use variants of the Transformer encoder or decoder to learn deep contextual representations from massive text corpora. These models are trained using self-supervised objectives and then fine-tuned for downstream tasks, setting new performance benchmarks across NLP. Due to their scalability, flexibility, and strong generalization capabilities, Transformers have become the dominant architecture in modern NLP research and applications.

2.2 Machine Translation

Kenny (2022, 23) defines translation as "[...] the production of a text in one language, the target language, on the basis of a text in another language, the source language". However, in translation, there is always a degree of approximation, since human translators can make different choices. Based on the definition mentioned above, MT is a branch of computational linguistics dealing with the automatic conversion of text or speech from a source language into a target language (Kenny, 2022). The overall goal of the translation process is to produce the target text in accordance with the principles of adequacy and fluency (Koehn, 2020). The added value and motivation behind research in MT stemmed from the increasing demand for fast, scalable, and cost-effective translation, particularly in contexts where human translation alone could not meet the volume or speed requirements (Kenny, 2022). Today, MT helps with overcoming language barriers and enabling global communication. It is widely used in domains such as international business, diplomacy, academia, and

humanitarian work, where quick and accessible translation is essential for effective interaction across linguistic boundaries.

The evolution of MT has progressed through several paradigms, beginning with Rule-Based Machine Translation (RBMT). This approach depends on manually crafted linguistic rules and bilingual dictionaries to carry out translations, typically involving a syntactic analysis of the source text, a transfer stage, and the generation of the target text. In RBMT, the system is provided with lexicons for both the source and target languages, as well as rules that define how words and grammatical structures correspond across languages (Kenny, 2022). This method requires time and linguistic expertise to develop, as the rules must be handcrafted for each language pair. Moreover, RBMT suffered from the so-called knowledge bottleneck, which stems from the difficulty of encoding all the linguistic and real-world knowledge required for high-quality translation (Sadler, 1992). Despite these challenges, RBMT dominated the field until the early 2000s and formed the basis for the first free online MT systems in the late 1990s (Joscelyne, 1998).

Statistical Machine Translation (SMT) emerged in the late 1980s, when statistical methods were first applied to translation by modeling it as a probabilistic optimization problem (Koehn, 2020). The 2000s marked a turning point, driven by increased computational power, the rise of the internet, and the growing availability of large-scale parallel corpora (Koehn, 2020). Unlike its rule-based predecessors, SMT relies on statistical models derived from bilingual datasets composed of aligned texts from both the source and target languages. The translation model pairs words or sequences from both languages based on co-occurrence frequencies, assigning each pair a probability score. During translation, known as decoding, the SMT system generates candidates for hypothetical translations and selects the most probable one based on the learned models and assigned weights (Kenny, 2022). Because SMT often relies on relatively short n-grams, sequences of n contiguous items used to model local word dependencies, such as bigrams or trigrams, it treats phrases as largely independent units (Jurafsky and Martin, 2025). This leads to issues such as word drop, inconsistent translations, difficulty in handling morphologically rich or agglutinative languages, and an inability to model long-distance dependencies (Kalchbrenner and Blunsom, 2013).

While neural network methods in MT date back to the late 1980s (Allen, 1987), they remained largely theoretical due to insufficient computational power and limited data availability. As computing resources and data volumes increased, neural methods gradually re-entered the field, initially as enhancements to SMT components (Koehn, 2020). A breakthrough came with the introduction of end-to-end neural models such as sequence-to-sequence frameworks (Cho et al., 2014b; Sutskever et al., 2014) and attention mechanisms (Bahdanau et al., 2015), which enabled better handling of long-distance dependencies, a known limitation of SMT

(Koehn, 2020). NMT uses deep learning methods, primarily neural networks, to represent and perform translation tasks. Instead of relying on predefined phrase tables or handcrafted rules, NMT systems learn to directly map input sentences to output translations in an end-to-end manner (Kalchbrenner and Blunsom, 2013). During training, the system processes source sentences through an encoder that transforms input subword tokens into dense vector representations, capturing contextual and semantic information. A decoder then generates the target sentence token by token, assigning probability scores based on both the encoded source and previously generated target tokens (Kenny, 2022). The introduction of NMT significantly improved fluency and translation quality, especially for managing long-range dependencies and complex sentence structures (Bahdanau et al., 2015), and has now become the dominant approach in MT (Bojar et al., 2017). The following section provides a more detailed overview of this method.

2.2.1 Neural Machine Translation

At the core, NMT addresses the sequence-to-sequence task, where the system is trained to transform a sequence in the source language into a corresponding sequence in the target language by learning from large parallel corpora (Kenny, 2022). Unlike rule-based or statistical approaches, NMT systems learn to perform this task end-to-end using neural networks trained on large parallel corpora. A common architecture used in NMT is the encoder-decoder framework (Koehn, 2020), in which the encoder processes the input sequence into a series of continuous vector representations, while the decoder produces the output sequence using these representations. The specific way in which encoding and decoding are performed varies depending on the underlying neural architecture, with early models relying on recurrent networks and more recent systems using attention-based mechanisms such as the Transformer (Vaswani et al., 2017).

However, large-scale parallel corpora are not strictly necessary, thanks to self-supervised learning methods (Ruiter et al., 2019). Self-Supervised Neural Machine Translation (SSNMT) approaches enable models to learn from comparable rather than strictly parallel corpora by jointly selecting training data and learning translation. These models exploit semantic representations learned during training to identify useful sentence pairs from monolingual or loosely aligned bilingual corpora (Ruiter et al., 2020). Translation mappings are learned by aligning latent representations across languages, often using techniques such as back-translation and denoising objectives, and have shown promising results, also in low-resource settings.

Beyond how training data is sourced, another component in NMT is the choice of neural architecture, which determines how translation is learned and generated. Initially, Recurrent Neural Networks (RNNs) were employed in this framework

because of their ability to process sequential data by maintaining hidden states across time steps (Bahdanau et al., 2015). In these models, the encoder processes the input sentence token by token, and the final hidden state, or a sequence of hidden states, is passed to the decoder, which generates the target sentence one word at a time in an autoregressive manner, each predicted word depending on the previous ones. However, the limitations of RNNs, especially their difficulty in modeling long-range dependencies due to the vanishing gradient problem, led to the development of Long Short-Term Memory (LSTM) networks (Hochreiter and Schmidhuber, 1997b) and Gated Recurrent Units (GRUs) (Cho et al., 2014b). These architectures introduced gating mechanisms that allowed them to retain and propagate important information over longer sequences, improving translation quality (Koehn, 2020).

Despite improvements in RNN-based architectures such as LSTMs and GRUs, they suffer from inherent limitations, particularly in modeling long-range dependencies due to their sequential nature and issues like vanishing gradients (Bengio et al., 1994). The introduction of the Transformer model by Vaswani et al. (2017) brought a shift in NMT by eliminating recurrence and convolution altogether. Instead, the Transformer architecture relies entirely on self-attention mechanisms, allowing the model to compute representations of the input sequence in parallel, improving training efficiency and scalability (Ott et al., 2018). The main advancement is the the scaled dot-product attention mechanism, allowing the model to assess the relevance of different tokens in a sequence relative to each other (Clark et al., 2019). This mechanism is employed in both the encoder and decoder, enabling the model to recognize complex dependencies, no matter how far apart they are in the sequence. In contrast to encoder-decoder architectures that compress the source sequence into a single context vector, the Transformer enables the decoder to attend to all encoder outputs at each decoding step, thereby improving alignment and translation quality, particularly for long or syntactically complex sentences (Vaswani et al., 2017). Furthermore, the Transformer's use of positional encodings compensates for the lack of recurrence by injecting information about token order directly into the input embeddings. This design enables the model to maintain sensitivity to word order while leveraging fully parallel computation. These architectural advantages have led to significant performance gains over RNN-based systems on standard MT benchmarks, and the Transformer remains the basis of most state-of-the-art MT models today (Zhang and Zong, 2020).

NMT is widely regarded as the most effective type of MT developed to date (Kenny, 2022). Unlike rule-based systems, which operate based on hand-crafted linguistic rules, or SMT, which relies on surface-level n-gram probabilities, NMT processes entire sentences at once and generates translations using deep, context-aware representations. This allows improved capture of complex linguistic phenomena,

like discontinuous dependencies and syntactic agreement. Despite its strengths, NMT also has limitations. Traditional NMT systems focus on sentence-level processing and struggle with cross-sentence coherence, such as correctly resolving pronouns or discourse relations (Stojanovski and Fraser, 2018). Additionally, NMT models require massive parallel datasets and significant computational resources, making them less feasible for low-resource languages. The following section discusses specialized strategies for improving machine translation, particularly in the low-resource settings.

In recent years, several large pretrained NMT models have advanced the field by enabling transfer learning, multilingual translation, and improved performance in low-resource settings. One such model is T5 (Text-to-Text Transfer Transformer) (Raffel et al., 2023), which reframes all NLP tasks, including translation, as a text-to-text problem. It uses a standard encoder-decoder Transformer architecture and is pretrained on a large-scale corpus using a self-supervised learning objective. By treating tasks like translation as sequence transformation task, T5 enables transfer learning across different tasks with minimal architectural adjustments. Once pretrained, the model can be fine-tuned on translation-specific data, achieving strong results due to its generalization capabilities. Multilingual BART (mBART) (Liu et al., 2020b) extends BART's denoising autoencoder architecture to the multilingual setting. It is pretrained on large-scale monolingual corpora in multiple languages using a combination of token masking and sentence permutation as noise functions. Unlike many previous pretraining approaches, mBART jointly trains the full encoder-decoder Transformer in an autoregressive manner. The model is language-agnostic and task-agnostic, enabling fine-tuning for supervised or unsupervised translation tasks across any language pair. No Language Left Behind (NLLB) (Costa-jussà et al., 2022) represents a milestone in massively multilingual translation. It uses a dense encoder-decoder Transformer architecture trained with a mixture of supervised and self-supervised learning techniques, scaling to over 200 languages. NLLB incorporates language-specific adapters, improved tokenization, and large-scale training infrastructure, allowing it to produce more balanced and accurate translations across both high- and low-resource languages. It demonstrates a shift toward universal translation models designed to minimize linguistic inequality in machine translation systems.

2.2.2 Evaluation of Machine Translation

Evaluating the quality of MT is an important component of MT research, as it provides the means to compare systems, guide development, and measure progress over time. In this thesis, the focus is on quality assessment, meaning methods that measure the quality of translation outputs, rather than quality estimation, where systems attempt to predict quality without reference translations.

A wide range of automatic metrics has been developed for MT evaluation. The most widely used is Bilingual Evaluation Understudy (BLEU) (Papineni et al., 2002), which measures the degree of n-gram overlap between system output and one or more reference translations, while incorporating a brevity penalty to prevent the system from producing translations that are too short. The underlying assumption is that a higher degree of surface-level overlap corresponds to better translation quality. Although BLEU has become a commonly used standard, it has been criticized for its sensitivity to surface-level matching and poor correlation with human judgments at the sentence level (Callison-Burch et al., 2006). To address implementation inconsistencies, Post (2018) introduced sacreBLEU, a standardized version of BLEU that ensures comparability across studies by fixing tokenization and preprocessing choices.

Other overlap-based metrics attempt to better capture linguistic similarity. The character F-score (chrF) (Popović, 2015) computes precision and recall over character n-grams, making it more robust to morphological variation and beneficial for morphologically rich languages such as Estonian. By operating at the character level, chrF implicitly captures subword patterns, making it sensitive to inflectional variation that word-based metrics may miss. METEOR (Denkowski and Lavie, 2014) extends word-level matching by including synonymy and stemming, offering improved correlation with human judgments compared to BLEU, although at the cost of higher computational complexity. ROUGE (Lin, 2004), initially developed for summarization, is also applied in MT, with variants such as ROUGE-n and ROUGE-L capturing overlap in n-grams and longest common subsequences. It assumes that high-quality translations will preserve many of the same sequences of words as the reference.

More recently, learned metrics based on pretrained language models have been proposed, aiming to overcome the limitations of string-based methods. One example is COMET (Rei et al., 2020), which leverages multilingual contextual embeddings from large pretrained models to evaluate translations in terms of adequacy and fluency. Such metrics have shown stronger correlations with human evaluation than traditional n-gram based scores and are increasingly adopted in MT shared tasks (Freitag et al., 2021).

Despite the advantages of automatic metrics, human evaluation remains the gold standard for assessing MT quality. Human assessments can take several forms, including direct assessment of adequacy and fluency, pairwise ranking of outputs from different systems, or error annotation frameworks that classify translation errors by type and severity (Lommel et al., 2013). These evaluations provide a more nuanced picture of translation quality but are costly and time-consuming to conduct. For this reason, human evaluation is often used to validate automatic metrics rather than as the sole evaluation method.

2.2.3 Machine Translation Approaches for Low-Resource Languages

NMT has substantially improved the quality of automated translation, yet it continues to face challenges in low-resource settings. The limited availability of parallel corpora remains a central issue in training effective models, as noted by both early and contemporary studies (Resnik and Smith, 2003; Sennrich and Zhang, 2019). This data sparsity issue is compounded by domain mismatch, where available training data fails to reflect specialized fields, like medical, legal, or technical content (Koehn and Knowles, 2017). As noted by Sennrich and Zhang (2019), although NMT has surpassed previous paradigms such as phrase-based SMT in many tasks, it continues to underperform under low-data conditions unless specific adaptations are employed. To overcome these challenges, several strategies have been developed (Haddow et al., 2022).

Data scarcity makes low-resource languages, such as Estonian, particularly reliant on synthetic data (Korotkova et al., 2024; Korotkova and Fishel, 2024). A widely adopted technique to address this is back-translation, which involves creating artificial parallel data by converting monolingual sentences from the target language back into the source language using an existing translation model (Sennrich et al., 2016). This approach enables models to benefit from fluent, natural target-side sentences and has been shown to improve translation quality, even in low-resource settings (Edunov et al., 2018). Variants such as iterative back-translation and tagged data further enhance effectiveness by increasing the quality and diversity of synthetic data (Hoang et al., 2018; Caswell et al., 2019). However, back-translation can introduce noise when the reverse model, which translates from the target language back into the source language, is weak, and gains often diminish after a few iterations (Haddow et al., 2022). Despite these limitations, it remains a foundational technique in LRNMT pipelines.

To further expand datasets and enhance model performance, data augmentation techniques such as paraphrasing, synonym replacement, and sentence generation are used (Fadaee et al., 2017). These methods aim to increase data diversity, thereby improving model robustness and generalization. Arthaud et al. (2021) applied a similar approach to adapt a model for low-resource translation by using a pre-trained Bidirectional Encoder Representations from Transformers (BERT) model (Devlin et al., 2019) to identify contextually appropriate substitutions for new vocabulary. Their results highlighted the trade-off between vocabulary specialization and overall translation quality, showing that data augmentation could enhance the translation of new words while maintaining high performance.

Multilingual training offers a complementary strategy to data-centric methods by rethinking how models are trained rather than focusing solely on data enrichment. These models are trained jointly on multiple languages, enabling them to

share representations and generalize across linguistic boundaries. This approach facilitates the development of a universal representation space and improves translation performance, especially for languages with shared grammatical structures or vocabulary (Johnson et al., 2017). The trade-off between parameter sharing and language-specific modeling has been studied, with findings showing that multilingual models can outperform bilingual ones, especially for low-resource language pairs (Lakew et al., 2018). However, adding more languages increases model complexity, and the balance between transfer and interference must be carefully managed (Aharoni et al., 2019). Despite these challenges, multilingual training has been shown to improve performance on low-resource languages, particularly when using a shared encoder-decoder architecture (Johnson et al., 2017).

One major benefit of multilingual models is their capability to handle zero-shot and few-shot translation. In the zero-shot setting, these models translate between language pairs without having encountered direct parallel data during training, relying instead on cross-lingual transfer through shared representations (Johnson et al., 2017). This is particularly effective in setups where bridge languages link otherwise disconnected pairs. In contrast, few-shot learning involves fine-tuning with minimal supervised data for a new language pair. Studies have shown that even a few hundred or thousand parallel sentences can yield strong performance, especially when the new language is typologically similar to those seen during training (Neubig and Hu, 2018). These methods reduce the dependency on large parallel corpora and are crucial for supporting languages with limited data.

Beyond multilingual training, transfer learning more broadly encompasses a range of techniques aimed at reusing learned knowledge across languages, domains, or tasks (Torrey and Shavlik, 2009). In LRNMT, it provides a mechanism for improving translation quality by leveraging knowledge from high-resource settings. A common approach is pretraining a model on a high-resource language pair and fine-tuning it on a low-resource one, allowing the model to transfer general linguistic and translation knowledge (Zoph et al., 2016). This can involve reusing encoder-decoder weights, shared embeddings, or even full multilingual models (Dabre et al., 2020). According to Dabre et al. (2017), fine-tuning benefits particularly from language similarity, where typological proximity between high- and low-resource languages improves the effectiveness of parameter reuse. Overall, transfer learning serves as a foundational technique in LRNMT, offering a flexible and scalable way to boost translation quality, particularly when combined with multilingual pretraining and data augmentation strategies.

In addition to data and training strategies, the efficient deployment of NMT models for low-resource languages relies on model compression techniques. Large-scale architectures, such as Transformers, demand significant resources, which may be unavailable in low-resource settings. To address this, Gurgurov et al. (2025) propose

a compression pipeline that combines knowledge distillation, structured pruning, hidden size truncation, and vocabulary trimming. Their approach compresses multilingual encoder models by up to 92% with only a modest performance drop of 2–10% across different downstream tasks. Such compression methods are crucial for enabling low-resource language translation systems to operate effectively under hardware constraints, supporting more scalable and environmentally sustainable NMT deployment.

2.3 Understanding Data Scarcity

Advancements in the field of NLP, including NMT, are largely attributed to the availability of large, high-quality datasets. For example, the quality of the output translation produced by an NMT system is closely linked to the volume of available parallel data, demonstrating that NMT performs best with language pairs where sufficient data exists (Koehn and Knowles, 2017). However, challenges arise when working with languages that lack sufficient data to train robust models (Joshi et al., 2020). Data scarcity refers to the lack of adequate and diverse training data, which is essential for building effective machine learning models, particularly for multilingual and cross-lingual tasks, such as MT. High-quality results depend not only on the volume of data, but also on its diversity, since a heterogeneous dataset improves the model's ability to generalize to unseen scenarios (Alzubaidi et al., 2023).

In this chapter, the concept of Low-Resource Language (LRL) will be explored, focusing on the challenges they present. The Estonian language, as the focal point of this thesis, will be examined in the context of its classification as an LRL. In addition, insights will be provided into its linguistic features, making it a relevant subject for low-resource MT research.

2.3.1 Definitions of Low-Resource Languages

The term **low-resource settings** extends beyond languages and also covers domains and tasks that lack dedicated training data (Hedderich et al., 2021). Thus, it applies to a variety of different scenarios characterized by differing levels of resource availability. This master's thesis, however, focuses specifically on under-resourced languages rather than domain- or task-specific low-data settings.

There is no general agreement on the unified definition of LRLs (Ranathunga et al., 2023). An indication of this is the variation of terminology, with a variety of synonyms to describe the concept of low-resource languages, such as low-density, low-data, resource-poor, and under-resourced (Besacier et al., 2014). In the context of NLP, Cieri et al. (2016, 4545) describe LRLs as languages that "[...] have fewer

technologies and especially datasets relative to some measure of their international importance". Within NMT, Zhang and Zong (2020) define LRLs as languages that lack bilingual training resources. Therefore, the definition of LRL depends on the specific languages, language pairs, and the NLP tasks being considered.

A common criterion for classifying languages as LRL or not is the availability of raw and annotated datasets, as well as parallel corpora. In the context of MT, the availability of parallel corpora is a key determinant of resource richness. However, there is no consensus on a threshold that defines high- versus low-resource languages by comparing the amount of available resources (Ranathunga et al., 2023). For instance, Yang et al. (2019) consider working with 350.000 labeled training instances in a text generation task as low-resource. Consequently, different thresholds are applied across tasks, and resource availability can be understood as "[...] a spectrum of resource availability" (Hedderich et al., 2021, 2547).

The scarcity of bilingual corpora and linguistic resources presents several challenges that impede the development of reliable translation systems for these languages. Even when a language achieves a substantial level of available data, the diversity and comprehensiveness of the data may be insufficient to capture all linguistic phenomena. For example, LRLs often suffer from not only limited quantity but also lower quality of available corpora (Kreutzer et al., 2022).

To better understand the digital status and resource availability of languages, Joshi et al. (2020) categorized around 2.000 languages into six groups based on the availability of language resources and their representation in NLP. These groups range from The Left-Behinds, such as Bora, which have no digital presence, to The Winners, which dominate the digital landscape and benefit from strong industrial and governmental investments for future development. Joshi et al. (2020) emphasize the importance of enhancing language inclusion, not only to support endangered languages and provide broader access to language technologies, but also for practical applications, such as developing emergency response solutions in underrepresented languages.

2.3.2 Factors Influencing Resource Availability

Nigatu et al. (2024) reported that their findings on different definitions for LRLs identified four interrelated factors: socio-political aspects, resources (both human and digital), linguistic artifacts, and agency of the community. These factors collectively influence data collection, accessibility, and availability for NLP applications.

Financial, economic, and political forces significantly impact resource availability for languages. As noted by Blasi et al. (2022), the economic influence of a language's speakers is a more decisive factor in the development of language technologies than their population size. While large speaker populations can help attract attention, it is often institutional and market power, not just demographics, that drives

investment. Nonetheless, widely spoken languages, especially those with official status in multiple countries, tend to benefit from governmental and educational support (Joshi et al., 2020), reinforcing their dominance in NLP research and resource creation. In contrast, marginalized languages frequently lack representation in public media, education systems, and official policies, limiting opportunities for linguistic preservation and digital inclusion. Policies that support multilingualism and open data initiatives can facilitate resource development, whereas the absence of such support contributes to digital exclusion and endangerment (Bird, 2020).

As several studies have shown, the development of NLP resources depends not only on language demographics but also on the availability of human expertise and technological infrastructure (Joshi et al., 2020; Nekoto et al., 2020). Human resources include first-language speakers, linguistic experts, and NLP researchers, all of whom are essential for data curation, annotation, and validation. Yet some languages remain low-resource despite a large speaker base due to limited digital presence or insufficient research engagement (Joshi et al., 2020). Access to digital devices and platforms determines how much linguistic data is available, as many LRLs lack substantial online content, making it difficult to collect monolingual or parallel corpora through conventional web crawling methods (Nigatu et al., 2024). Digital inequality further exacerbates these issues, as communities with limited access to technology face additional challenges in contributing to data collection efforts.

The third factor mentioned by Nigatu et al. (2024) is linguistic artifacts, including structured linguistic knowledge, data, and computational tools, which are critical for NLP applications. The availability of high-quality parallel corpora is particularly important for MT, however, curating such datasets is labor-intensive and expensive, limiting access for many LRLs (Joshi et al., 2020). Additionally, linguistic complexity, such as morphological richness and non-standardized orthographies, can increase data sparsity, making conventional NLP approaches less effective (Ponti et al., 2019). Beyond data, technological exclusion also plays a role as many LRLs lack foundational NLP tools, pre-trained language models, or even basic text-processing utilities (Simons et al., 2022). Without computational resources and tailored linguistic models, developing language technologies for these languages remains a significant challenge.

Community involvement is an important yet often overlooked factor in language technology development. Even when linguistic data exists, financial and infrastructural barriers may prevent communities from actively contributing to NLP research. Moreover, the misalignment between academic or commercial NLP goals and community needs can limit the relevance and usability of the resulting technologies (Bird, 2020; Hershcovich et al., 2022). In contrast, when communities are directly involved in the research and development process, the resulting tools tend to better

reflect their linguistic, cultural, and communicative priorities (Nekoto et al., 2020). Collaborative efforts that incorporate community-driven initiatives can help address data gaps and ensure that language technologies serve their intended users more effectively (Bella et al., 2022).

2.3.3 Estonian as Low-resource Language

The Estonian language serves as an example of a LRL in the context of MT research. In the categorization conducted by Joshi et al. (2020), discussed in Section 2.3.1, Estonian is classified among The Rising Stars alongside languages such as Ukrainian and Hebrew. This classification reflects its stable online presence and an existing active cultural community on the web, but it also highlights the limited efforts to collect labeled data. Although Estonian is the official national language of Estonia and is supported by state institutions (Eberhard et al., 2024), its relatively small speaker base of approximately one million limits the interest for private-sector investment in language technologies. While the general digital presence of Estonian is acceptable, the availability of linguistic resources is extremely limited compared to high-resource languages (Muischnek, 2023). Muischnek (2023) mentions particularly the lack of parallel corpora for Estonian and non-English languages. However, the categorization of Estonian as low-, medium-, or high-resource language can vary depending on the task or language pair in question (Tars et al., 2022a,b; Kocmi and Bojar, 2018). For instance, Estonian has significantly more resources available than some closely related languages, such as Livonian (Rikters et al., 2022), and in certain contexts, this relative abundance allows it to be considered a high-resource language.

Moreover, several linguistic characteristics of the Estonian language present particular challenges for MT. While most European languages belong to the Indo-European family, Estonian is classified under the Finno-Ugric group within the Uralic family. For this reason, its basic vocabulary differs from most of the languages spoken in that region. Furthermore, Estonian uses additional characters beyond the standard ASCII set, which are $\ddot{A}$, $\ddot{O}$, $\ddot{U}$, $\tilde{O}$, $\check{S}$, and $\check{Z}$ (Muischnek, 2023). Although Estonian shares word formation processes, such as derivation and compounding with Germanic languages, it retains distinct Uralic features like extensive use of inflection (Erelt, 2007). Estonian has an extensive system of inflection, particularly for nouns and verbs. Nouns decline in 14 grammatical cases, such as nominative (*sõber*, friend), genitive (*sõbra*, of the friend), elative (*sõbrast*, from a friend), and absessive (*sõbrata*, without a friend) (Chuang et al., 2020). The variation in stem forms due to consonant gradation and vowel alternation further increases morphological complexity. These characteristics contribute to the challenges of MT, particularly in tasks such as word alignment, lemmatization, and morphological generation. Additional influences and characteristics shaped by Estonian's genealogical origins

and regional context will be discussed in Section 2.4.

Estonian is typologically a transitional language. It exhibits characteristics of both agglutinating and fusional languages (Muischnek, 2023). Historically, Estonian developed as an agglutinating language, where words are formed by combining discrete morphemes that each represent a single grammatical function. However, over time it has developed fusional features, such as irregular forms that extend beyond simple suffix addition (Erelt, 2007). Estonian also has a relatively free word order influenced by information structure, although there is a tendency to position the verb in the second position in a sentence (Erelt, 2007). This syntactic flexibility makes parsing and translating Estonian more challenging for NMT systems. Other notable characteristics relevant to NMT include the absence of grammatical gender and articles, as well as a rich morphological system. There are 14 grammatical cases in both singular and plural forms, and verbs inflect for person, number, tense, mood, and voice (Muischnek, 2023).

The availability of datasets for Estonian is limited, particularly in the context of MT, which requires large volumes of parallel data. Therefore, the first step for many researchers is to compile data from different sources, including web crawling and back-translation of the collected texts (Purason et al., 2024; Tättar et al., 2022; Rikters et al., 2022). Initiatives such as the European Language Resource Coordination (ELRC) and the European Language Grid (ELG) have contributed to the development of multilingual resources for Estonian. As Estonian is one of the official languages of the European Union (EU), it is included in parallel datasets made available by EU institutions, such as (Muischnek, 2023):

- Europarl: A collection of European Parliament speeches, providing approximately 620,000 sentence pairs for most EU language pairs (Koehn, 2005).
- EMEA: A corpus containing up to 1.1 million aligned sentences from medical domain texts (Tiedemann, 2012).
- JRC-Acquis: A legal domain corpus that includes up to 1.7 million sentences between Estonian and other European languages (Steinberger et al., 2006).
- ELRC-SHARE: A repository offering a variety of parallel datasets between Estonian and other European languages across domains, including legal, news, medical, culture, and scientific research (Lösch et al., 2018).

Additional Estonian–English datasets are available, such as ParaCrawl (Bañón et al., 2020). However, the volume of existing parallel corpora between Estonian and non-European languages is considerably smaller. Available resources include:

- OpenSubtitles: A collection of movie subtitles containing up to 12 million aligned sentences, including translations into non-European languages such as Arabic, Hebrew, and Chinese (Tiedemann, 2012).

- NLLB: A significant dataset providing parallel corpora for Estonian with languages such as Catalan (1.2 million sentence pairs), Hindi (1.4 million), and Korean (1.7 million) (Schwenk et al., 2021; Fan et al., 2021)

A common approach in Estonian-related NMT research involves the creation of custom datasets, which are later published for broader use (Tars et al., 2021; Rikters et al., 2022). One of the most notable outcomes of this approach is the development of SynEst dataset, which contains over 1 billion sentence pairs translating between Estonian and 11 other languages (Korotkova et al., 2024). This corpus was later extended by Korotkova and Fishel (2024), who incorporated additional data sources and applied a back-translation approach. These improvements resulted in and expanded dataset containing over 2 billion filtered sentence pairs, covering 12 new translation directions.

By leveraging these parallel datasets and employing data augmentation techniques such as back-translation, researchers can improve the quality of Estonian-language NMT systems. However, further efforts are required to increase the availability of high-quality parallel corpora, particularly for Estonian and non-European language pairs (Muischnek, 2023).

2.4 Language Classification and Linguistic Relationships

Language classification can be performed in a variety of ways, depending on the purpose and focus of the analysis, with four major approaches commonly recognized: genetic, areal, lexicostatistical, and typological (Brown and Ogilvie, 2009). A widely used method is genetic classification, which groups languages into families based on descent from a presumed common ancestor. Areal classification offers a different perspective by grouping languages based on either structural similarities due to contact or their location within the same geographic region. Lexicostatistical classification relies on word comparison to assess relationships between languages. Finally, typological classification groups languages according to structural characteristics, such as word and sentence patterns, regardless of historical origin.

While these linguistic classifications are central to cross-linguistic studies, it is important to note that languages can also be grouped according to non-structural criteria, such as the number of speakers or social factors like formality and speaker demographics (Velupillai, 2012). Nevertheless, for the purpose of linguistic typology and transfer learning, genealogical affiliation and geographic distribution remain especially critical. The following sections explore typological, genealogical, and

areal classification in more detail, highlighting their relevance to transfer learning and Estonian machine translation.

2.4.1 Language Typology

Language typology classifies languages based on structural features rather than historical origins. As Comrie (1989) notes, linguistic typology examines the variation across the world's languages by systematically comparing their structures. In contrast to genealogical classification, which groups languages by shared ancestry, typological classification is based on formal similarities in aspects such as phonology, grammar, or vocabulary (Crystal, 2010). As Velupillai (2012, 15) defines it, linguistic typology "[...] is the study and interpretation of types of linguistic systems", which generally involves cross-linguistic comparison, although it can also include comparisons within a single language. This approach helps to identify linguistic universals and variations, offering insights into how languages function and how they can be processed computationally.

In NLP, understanding linguistic typology is crucial because language structure significantly impacts model performance on certain tasks (O'Horan et al., 2016). More specifically in NMT, typological knowledge can guide the selection of an appropriate source language for transfer learning, as it helps identify languages with structural similarities that facilitate more accurate translations, even in the absence of large parallel corpora (Ponti et al., 2019).

Most NLP systems are trained on high-resource languages like English, which has relatively simple morphology and a fixed word order. However, many languages, particularly those with rich morphological systems or flexible word orders, present challenges for MT (Koehn and Knowles, 2017). For instance, morphologically rich languages, such as Russian or Turkish, require translation models to handle complex word structures, where a single source word may correspond to multiple words or an entire phrase in the target language (Passban et al., 2018).

One key typological factor influencing translation is morphology (Nzeyimana, 2024). Following traditional morphological classifications, languages can be broadly classified based on their morphological characteristics into several types (Velupillai, 2012). Isolating languages, e.g., Chinese, tend to use single-morpheme words and depend primarily on word order and auxiliary words to express grammatical relationships. Agglutinative languages, e.g., Finnish, form words by combining distinct morphemes, each representing a specific grammatical function. Fusional languages, e.g., Russian, also use affixes, but individual morphemes often encode multiple grammatical meanings at once. Some languages are polysynthetic, e.g., Central Siberian Yupik, combining many morphemes into a single complex word that can represent a full sentence. Others, like Arabic, are classified as introflexive, where grammatical distinctions are marked through internal modifications of the

word stem. It is important to note, however, that few languages fit perfectly into a single type, and many languages employ a mix of morphological strategies (Bickel and Nichols, 2007).

Word order plays a central role in MT, especially in transfer learning setups where parallel data is scarce (Murthy et al., 2019). NMT systems often rely on training data where the assisting language has a different word order from the source language, which can lead to reordering challenges. The ordering of words varies across languages. English, for example, follows a fixed Subject-Verb-Object (SVO) structure, where syntactic roles are largely determined by position (Dryer, 2013). In contrast, languages such as Turkish exhibit Subject-Object-Verb (SOV) order, and many others show different or more flexible patterns, such as Verb-Subject-Object (VSO) or free word order (Brown and Ogilvie, 2009). Moreover, some languages such as Hungarian allow relatively free constituent order due to rich case marking, enabling a greater variety of surface arrangements without changing grammatical relations (Brown and Ogilvie, 2009). This flexibility increases the complexity for MT systems, particularly when the source and target languages differ in their default or dominant word orders.

Determining the dominant word order of a language is not always straightforward, as it often requires large corpora of simple, pragmatically neutral declarative sentences (Velupillai, 2012). As Dryer (2013) notes, some languages exhibit flexible constituent order but still have a dominant word order, while others do not show a clearly dominant structure at all. Furthermore, word order typology is closely linked with genealogical and areal influences, given that shared ancestry or prolonged contact can result in structural convergence across otherwise unrelated languages (Heine and Kuteva, 2005). Case marking further complicates translation, particularly in languages with flexible word order, since the two features are often correlated. Languages such as Russian or Tamil, which exhibit flexible word order, also employ case marking to disambiguate constituent roles, adding another layer of complexity for MT systems (Bisazza et al., 2021). For MT, recognizing these interactions is essential for building models that can handle the complexities of low-resource languages like Estonian, where typological knowledge can be leveraged to improve translation performance through transfer learning.

2.4.2 Role of Genealogical and Areal Relationships

Languages are classified into families based on their shared ancestry, with related languages typically exhibiting similar phonetic, grammatical, and lexical features. Genealogical classification is, as Crystal (2010, 303) explain, "[...] based on the assumption that languages have diverged from a common ancestor. It uses early written remains as evidence, and when this is lacking, deductions are made using the comparative method to enable the form of the parent language to be reconstructed."

Understanding these genealogical relationships is crucial in NLP and MT, as they provide insights into structural similarities that can be leveraged for transfer learning and multilingual modeling. However, as Bickel (2007) points out, typological classifications are primarily grounded in empirical structural data, and while they often align with genealogical or areal patterns, such causal interpretations are secondary to the observed linguistic evidence. This distinction is important when applying linguistic typology in computational contexts, where structural similarity may or may not align neatly with language family trees.

Languages that share a common ancestor often exhibit overlapping vocabulary, morphological patterns, and syntactic structures. As Velupillai (2012) explains, genealogical classification groups languages into families based on historical descent using methods of historical and comparative linguistics, such as the identification of cognates and the reconstruction of proto-forms through the comparative method. This process enables linguists to build family trees where a single parent language gives rise to multiple daughter languages. For instance, the Indo-European language family has been reconstructed by aligning forms like Sanskrit *pitár*, Greek *patér*, and Latin *pater* to propose a common root with the consonants *p*, *t*, and *r* (Blake, 2009). Systematic sound changes, such as the shift from *p* to *f* in English *father*, reveal how modern descendants have evolved, while still preserving evidence of shared origin. These shared phonological, morphological, and syntactic features are especially evident in subfamilies such as Romance, e.g., Spanish, French, Italian, and Germanic, e.g., English, Dutch, German (Besters-Dilger et al., 2014). In computational terms, this inherited congruence allows for more effective transfer learning and cross-lingual adaptation in NLP models, as structural features learned from one language can generalize more effectively to related languages.

The Uralic languages comprise a geographically and structurally diverse family spoken across Northern Eurasia, from Hungary to Western Siberia (Blake, 2009). As Marcantonio (2006) outlines, this family includes languages such as Hungarian, Finnish, Estonian, the Saami languages, and a number of minority languages like Komi, Mari, and Khanty. Estonian is classified within the Finnic branch of the Uralic family, is closely linked to Finnish and is more distantly to languages like Karelian and Veps, which together form the Northern Finnic group (Laakso, 2001). Among these, Finnish is its closest major relative, with which it shares many phonological and syntactic features, though mutual intelligibility is limited. While conventional classifications place Hungarian and the Ob-Ugric languages within a Ugric subgroup, these languages differ in vocabulary, syntax, and phonology, which has made it difficult to reconstruct a common proto-language from direct linguistic evidence (Marcantonio, 2006). Uralic languages often share typological features such as vowel harmony, consonant gradation, agglutinative morphology, and rich case systems. These features are also found in neighboring Altaic languages, complicating

genealogical delineation (Johanson, 2006). Despite these shared features, there is substantial internal variation. For instance, Finnish has a relatively small consonant inventory, while North Saami dialects can have over thirty consonants (Marcantonio, 2006). These typological properties and regional patterns reflect both inherited traits and language contact, underscoring the complexity of establishing clear-cut genealogical relationships.

While genealogical relationships shape languages through common descent, languages also evolve due to areal influence, which results from prolonged contact between speakers of different linguistic backgrounds. These influences can lead to contact-induced linguistic changes, including lexical borrowings, phonetic shifts, and grammatical restructuring (Velupillai, 2012). Unlike genealogical relationships, which stem from shared ancestry, areal influence occurs through social, political, and economic interactions, often creating linguistic similarities between unrelated languages. As languages in close proximity share features due to these sustained interactions, they may develop linguistic areas (Heine and Kuteva, 2005).

Estonian has been influenced by several neighboring languages due to prolonged contact over the centuries. Its geographic proximity to Baltic languages has resulted in shared features, such as grammaticalization of the verb *tulema* (to come) functioning as a modal auxiliary to convey deontic necessity (must, have to), with the agent marked in an oblique case (Heine and Kuteva, 2005). This construction closely parallels a similar development in Latvian, where the reflexive form of the verb *nák* (to come) also expresses obligation, with the agent likewise marked obliquely (Stolz, 1991). According to Heine and Kuteva (2005), given the genetic distance between Estonian and Latvian, and the rarity of such constructions cross-linguistically, this parallel is best explained as a result of language contact rather than shared inheritance.

Additionally, Estonian has been significantly shaped by Germanic languages, particularly through centuries of German dominance in the region, which has left a mark on vocabulary, word order, and certain aspects of its syntax (Harms, 2006). Scandinavian influence, particularly Swedish, is also notable due to periods of Swedish rule in Estonia, contributing to lexical borrowings and some phonetic changes, however, the influence was more remarkable in the costal areas (Erelt, 2007). Furthermore, Russian influence, particularly in the modern period, has affected Estonian's lexicon and phonetics due to Estonia's historical connection with the Russian Empire and later the Soviet Union (Verschik, 2005). In recent decades, English has also become a significant source of lexical influence on Estonian, especially in domains such as technology, business, online communication, and popular culture, reflecting Estonia's post-independence globalization, the widespread exposure to English in media and education, and its role as a global lingua franca (Kask, 2016). These diverse influences have contributed to the development

of Estonian within a multilingual and multicultural context, where contact-induced linguistic features are widespread.

Recent research in computational linguistics has increasingly examined how genealogical and areal relationships influence language modeling and transfer in NLP. As Bakker (2010) explains, structural similarities among languages may arise from not only language-internal features, but shared ancestry or contact. Therefore, both play critical roles in multilingual representation learning. For example, Eger et al. (2016), studying bilingual vector spaces from Europarl, found that geographic proximity sometimes explains cross-lingual similarity better than genetic relatedness, while Bjerva et al. (2019) showed structural features, for instance from dependency trees, to be even more predictive than genealogy in clustering languages. Other studies, such as Rama and Borin (2015), leverage representational distances to infer phylogenetic relationships, often aligning with traditional family trees. At the same time, visualization of language vectors, such as in Libovický et al. (2019), reveals that learned embeddings cluster languages in ways influenced by both genetic and areal factors. Typological features, particularly those with high diachronic stability such as word order, tend to align closely with genealogical information (Ponti et al., 2019). As a result, integrating discrete typological data, such as that provided by the World Atlas of Language Structures (WALS) (Dryer and Haspelmath, 2013), with continuous language representations offers a promising direction for capturing cross-linguistic variation in NLP systems. These considerations are particularly relevant for low-resource languages like Estonian, which exhibit both genealogically distinctive and areally influenced features. As such, approaches that explicitly model inherited and contact-induced similarities offer a promising direction for improving multilingual representation and transfer in NLP.

3 Related Work

Research in LRNMT has explored various strategies, including architectural innovations, data augmentation techniques, and the influence of linguistic relationships on transfer learning. Several studies have demonstrated that pretraining or fine-tuning models on high-resource language pairs can substantially improve performance on low-resource targets. This chapter reviews contributions in this space, focusing on methods that evaluate transfer learning between related and unrelated languages, multilingual NMT for morphologically rich languages such as Estonian, and experiments involving Uralic and other closely related languages.

Kocmi and Bojar (2018) present a simple approach to transfer learning for NMT in low-resource scenarios. Their central contribution lies in demonstrating that effective knowledge transfer does not require language relatedness, shared vocabulary, or specialized training modifications. Their method involves pretraining a high-resource parent NMT model, and then continuing training on a low-resource child language pair using the same model architecture and training regime, a process termed by the authors as trivial transfer learning. All experiments use the standard Transformer model (Vaswani et al., 2017), with shared subword vocabularies across encoder and decoder. This architecture choice plays a key role in enabling parameter reuse and effective subword-level transfer.

Their work diverges from earlier transfer learning studies that typically relied on shared target languages or closely related language pairs (Zoph et al., 2016; Nguyen and Chiang, 2017). Instead, Kocmi and Bojar (2018) eliminate the need for shared language components altogether, showing that substantial translation quality improvements are possible even when the parent and child language pairs are entirely unrelated, for example Arabic-Russian as a parent for Estonian–English. Their results include significant improvements of up to +3.38 BLEU for English–Estonian when transferring from a Czech-English parent, which is notably higher than gains from using more closely related Finnish–English data. They also simulate very low-resource settings and show that their method maintains its efficacy, highlighting the practicality of trivial transfer learning in highly data-scarce settings. While Kocmi and Bojar (2018) adopt a language-agnostic approach, this study aims to investigate whether genealogical and areal linguistic similarities can enhance transfer effects. Their findings serve as a useful baseline and contrast, since if relatedness is found to be beneficial, it would refine their broader claim by demonstrating that while not required, linguistic relatedness can still yield further gains in cohesive language

groups.

In contrast to the language-agnostic approach proposed by Kocmi and Bojar (2018), Dabre et al. (2017) previously explicitly investigate the role of language relatedness in transfer learning for LRNMT. Their study empirically evaluates how different parent models, each trained on a high-resource source language to English pair, affect translation quality when used to initialize models for low-resource source languages.

Using a 4-layer LSTM encoder-decoder architecture with attention and word-piece segmentation, based on the model design from Zoph et al. (2016), they assess transfer performance across ten typologically diverse source languages. A central finding is that transfer learning yields greater gains when the source language of the parent model is linguistically similar to the child source language. For example, Hindi as a parent language improved performance for Indo-Aryan languages like Punjabi, with gains up to +2.41 BLEU, compared to unrelated but higher-resource parents such as French and German.

Their results suggest that genealogical similarity outweighs factors like dataset size or shared script, with related source languages consistently outperforming unrelated ones across both exhaustive and opportunistic experiments. This supports the hypothesis that shared linguistic features facilitate more effective parameter transfer. Their findings challenge the claim by Kocmi and Bojar (2018) that transfer success is largely independent of language similarity, and instead highlight relatedness as a factor, especially in low-resource scenarios. This raises important questions for the present master's thesis about whether these patterns hold across other language families and to what extent structural or lexical features drive effective transfer.

Rikters et al. (2018) add to this line by investigating how multilingual and multi-way NMT models handle related and morphologically rich languages. Unlike Kocmi and Bojar (2018), who emphasize language-agnostic sequential transfer, and more in line with Dabre et al. (2017), Rikters et al. (2018) focus on relatedness within a multilingual joint training setup. Building on the method of prepending language tokens to source sentences (Johnson et al., 2017), they train and compare multiple architectures: 6-layer Transformer models with convolutional embeddings (Transformer-M), 4-layer GRU-based encoder-decoder models (GRU-DM), and 15-layer convolutional models (FConv-M). The experiments are conducted across both one-way and multilingual multi-way configurations.

The findings show that multilingual training can offer substantial BLEU gains (up to +5.28) for low-resource directions, especially when involving typologically similar languages. However, they also observe performance trade-offs, such as degradation on high-resource pairs, likely due to conflicting gradients or parameter competition. Their results highlight the effectiveness of parameter sharing among

related languages, reinforcing the view that linguistic proximity can enhance model performance. While Rikters et al. (2018) adopt a joint multilingual strategy, their findings provide a point of comparison for the present study, which examines relatedness effects in a more controlled, sequential transfer setting, contributing to broader discussions on designing effective LRNMT strategies.

Another relevant contribution is the work by Tars et al. (2021), which focuses on extremely low-resource languages within the Uralic language family, specifically Võro, North Saami, and South Saami. Their study demonstrates how a combination of multilingual training, synthetic data generation through back-translation, and transfer learning can substantially improve translation quality even when parallel corpora are nearly absent. A key strength of their approach is the use of high-resource related languages to support the training of NMT models for closely related but under-resourced languages.

Their experimental setup involves first training a baseline Estonian–Finnish model, and then using its parameters for initializing a new model that is fine-tuned on the much smaller Estonian–Võro dataset. This approach leads to significant gain, up to +13 BLEU over the multilingual baseline, demonstrating the efficacy of leveraging linguistic similarity and pre-trained weights for targeted low-resource scenarios. This is especially relevant to the present work, which also centers on Estonian and explores its relationships with closely related Uralic languages. A notable trade-off highlighted in their findings is that while this strategy excels for specific target pairs, it sacrifices the generality of multilingual systems capable of translating in multiple directions simultaneously.

As with Tars et al. (2021), the work by Korotkova and Fishel (2024) also relies on synthetic data. They present a comprehensive approach to Estonian-centric machine translation by developing MT models for 18 translation directions involving Estonian, both from and into the language. Their key methodological innovation lies in leveraging the encoder of the multilingual NLLB model, which they keep frozen, while training modular, language-specific Transformer decoders for each target language. This modular design enables independent training and deployment of decoders, improving efficiency and flexibility. Their models consistently surpass the original NLLB system in translation quality, demonstrating significant gains across multiple language pairs. This modeling work is supported by an extension of the Synthetic Corpus of Parallel Estonian (SynEst), which they augment with over 2 billion back-translated sentence pairs. Their results indicate that such modular architectures, combined with targeted synthetic data, can yield state-of-the-art open-source MT systems for low-resource languages like Estonian. In contrast, rather than relying on large-scale synthetic data or pretraining on high-resource pairs, this study investigates naturalistic transfer between genealogically or areally related languages to better understand the extent to which linguistic proximity

alone can support successful low-resource NMT. This allows for a clearer assessment of cross-linguistic generalization in multilingual models, particularly in the absence of artificial data augmentation.

A recent contribution in this field is the study of Language verY Rare for All (LYRA) by Merad et al. (2025), which addresses the problem of MT for extremely low-resource languages, focusing on French-Monégasque. The work is particularly relevant for its methodological emphasis on transfer learning from typologically related high-resource languages and its focus on realistic low-resource conditions. The authors demonstrate that initializing training with a French–Italian model, given the close relationship between Italian and Monégasque, substantially improves performance compared to training from scratch. Their approach is further enhanced by careful data curation, where standardization of orthographic conventions, such as punctuation and capitalization, leads to performance gains, underscoring the impact of preprocessing decisions in data-scarce settings. In addtion, they explore Retrieval-Augmented Generation (RAG) techniques with decoder-only LLMs, retrieving similar training examples at inference time to guide translation output. The LYRA results suggest that even in the absence of extensive synthetic corpora or massive parallel data, targeted transfer from structurally similar languages remains an effective strategy for supporting endangered and under-resourced languages.

4 Method

This chapter outlines the methodological framework used to investigate the impact of genealogical and areal relationships on LRNMT, with a focus on Estonian as the target language. It begins by describing the overall experimental setup, followed by sections detailing the dataset selection, model choice, fine-tuning procedure, and evaluation strategy employed throughout the study.

4.1 Overview of Experiments

The goal of this study is to examine how fine-tuning a multilingual NMT model with different source languages affects the generation quality of Estonian as a target language. The underlying assumption is that both genealogically related and areally proximate languages may provide useful inductive transfer when fine-tuning low-resource NMT systems. The experiments were designed to systematically explore this by measuring the impact of fine-tuning on translation performance across multiple source languages, compared to both zero-shot and in-language baselines. The multi-step setup allows to isolate and analyze the effect of different source languages on Estonian generation, while keeping the evaluation procedure consistent across runs.

The experimental workflow consists of the following four stages:

1. Zero-shot baseline
2. English fine-tuning
3. Cross-lingual fine-tuning
4. Comparison and analysis

First, the pre-trained NLLB-200 model (600M parameter version) (Costa-jussà et al., 2022) is directly evaluated on English–Estonian translation without any task-specific fine-tuning. This serves as a baseline to understand how well the model performs in a zero-shot setting and provides a reference point for later improvements.

Second, the model is fine-tuned on in-domain English–Estonian parallel data. Performance is then evaluated on a held-out English–Estonian test set. This step

establishes a performance ceiling for fine-tuning with the actual source language used during testing.

Third, for each selected source language X (Finnish, Hungarian, Latvian, German, Russian, Swedish), the model is fine-tuned on X–Estonian data points. However, evaluation is still conducted on the same English–Estonian test set. The objective is to assess whether fine-tuning on a non-English language still improves Estonian generation quality, and whether linguistic proximity, based on genealogy or geography, plays a role in the extent of transfer.

Final test scores are compared across all experiments, including both baselines and each cross-lingually fine-tuned model. The applied evaluation scores are sacreBLEU (Post, 2018) and chrF (Popović, 2015). In addition to quantitative evaluation, qualitative inspection of translated outputs is conducted to identify patterns of improvement or degradation, particularly with respect to lexical and syntactic fluency.

4.2 Dataset

This study focuses on MT into Estonian, using different source languages selected for their genealogical or areal relation to Estonian. The source languages used in the experiments are as follows:

- **Genealogically related**: Finnish, Hungarian
- **Areally related**: Latvian, German, Russian, Swedish

The source languages were chosen to represent a range of linguistic and typological relationships to Estonian. Finnish and Hungarian are both members of the Uralic language family along with Estonian. Finnish is the closest major linguistic relative of Estonian (Karlsson, 2009). However, Hungarian has been separated from other Finno-Ugric languages and therefore provides a further relative with still strong connections, such as complex morphology (Kiss, 2009). Latvian, German, Russian, and Swedish were included due to their areal proximity and historical contact with Estonian, offering an opportunity to assess whether shared linguistic features arising from geographic or cultural contact also facilitate transfer during fine-tuning.

Parallel corpora were sourced from the European Language Resource Coordination (ELRC) Presscorner_COVID collection (Piperidis et al., 2018), ensuring domain consistency through the use of EU press releases related to the COVID-19 pandemic. The dataset was distributed via the OPUS platform (Tiedemann, 2012). The only exception is the Russian–Estonian pair, for which parallel data from the Wikimedia corpus (Tiedemann, 2012) was used due to the absence of a Russian–Estonian dataset in the ELRC collection.

In order to minimize overlap with the NLLB model's pretraining data, the selected corpora were chosen to be highly unlikely to have been part of its original training pipeline. While full exclusion cannot be guaranteed due to the lack of transparency around the NLLB training pipeline (Costa-jussà et al., 2022), widely used general-domain sources were explicitly avoided. The selected corpora are domain-specific and relatively narrow in topical scope, which reduces the risk of direct overlap with large-scale pretraining datasets. Additionally, prior to experimentation, simple sanity checks were conducted to evaluate whether the pretrained NLLB model without any fine-tuning was able to translate example sentences from the selected corpora unexpectedly high accuracy, which would indicate that they may have been seen during pretraining. As translation quality was relatively low, this suggested that the corpora had not been part of the model's training data and were therefore suitable for use in the experiments.

To maintain comparability across experiments, all language pairs were aligned to the same data size: 1,000 sentence pairs for training, and 200 pairs each for validation and test sets. This standardization helps ensure that observed differences in performance can more confidently be attributed to linguistic or structural factors rather than discrepancies in data volume. Initially, smaller datasets were tested due to computational constraints, but the final configuration was chosen as a balance between feasibility and the need for meaningful performance evaluation.

Preprocessing steps were applied uniformly across all language pairs, involving strict filtering to ensure clean and well-formed sentence pairs. Empty or non-string entries were removed from both source and target sides. Sentence pairs were further filtered to exclude those exceeding 250 tokens in either language. Tokenization and encoding were consistently applied using the NLLB-200 tokenizer across all language pairs.

The overall dataset design aimed to balance domain consistency and size across language pairs while including a typologically and geographically diverse set of source languages. This combination supports the broader research goal of assessing whether linguistic relatedness, either genealogical or areal, has a measurable impact on the effectiveness of cross-lingual fine-tuning in low-resource MT scenarios.

4.3 Model Selection

For all experiments in this study, the pre-trained multilingual MT model NLLB-200-distilled-600M (Costa-jussà et al., 2022) was used. This model, released by Meta AI as part of the No Language Left Behind (NLLB) initiative, supports 200 languages and offers a balance between translation quality, language coverage, and computational feasibility. The 600M-parameter distilled version was selected primarily due to its more efficient memory and runtime characteristics, which make

it suitable for fine-tuning in a resource-constrained environment such as Google Colab. As noted in Section 4.2, model selection was also informed by the desire to minimize overlap between the model's original pretraining data and the fine-tuning corpora used in this study.

Estonian is among the supported languages in NLLB-200 and classified as high-resource language by Costa-jussà et al. (2022), although it remains relatively low-resource in the broader multilingual context. Choosing a model that already includes Estonian in its training data ensures that the model has foundational knowledge of the language's structure and vocabulary. This is important for facilitating effective fine-tuning rather than learning the language from scratch. At the same time, Estonian's low-resource status means that meaningful performance improvements from additional fine-tuning are still possible and measurable.

4.4 Fine-Tuning

The fine-tuning process was carried out using the Hugging Face Transformers and Datasets libraries, alongside PyTorch as the underlying deep learning framework. All training scripts were run on Google Colab Pro environment, using the T4 GPU. This setup allowed for reproducible experimentation across language pairs and parameter settings.

Each experiment involved fine-tuning the pre-trained nllb-200-distilled-600M model on 1,000 parallel training examples, with 200 held-out samples each for validation and testing. The goal was to adapt the model to better generate Estonian given a specific source language, under consistent low-resource conditions.

Fine-tuning hyperparameters were adjusted through multiple experimental runs. The initial configuration used a learning rate of $2e^{-5}$ and 3 training epochs. However, this setting often led to underfitting or unstable convergence. After further testing, a lower learning rate of $1e^{-5}$, and for some source languages $5e^{-6}$, combined with 5–8 training epochs produced significantly better validation loss and final translation quality. A per-device batch size of 4 was used throughout all runs due to hardware constraints. All experiments used the Adam optimizer with a weight decay of 0.01 and mixed precision training enabled via `fp16=True`, which improved memory efficiency. Evaluation and checkpoint saving were performed at the end of every epoch (`eval_strategy="epoch"`, `save_strategy="epoch"`), ensuring that validation loss was consistently monitored throughout training. The model checkpoint with the lowest validation loss was retained using the argument `load_best_model_at_end=True`.

Training times per run varied between 10 and 20 minutes, depending on the number of epochs. Each fine-tuning job was conducted independently per language pair. Table 1 summarizes the best hyperparameter settings and training outcomes

for each source language used in fine-tuning.

Table 1: Best fine-tuning parameters and selected checkpoints for each source language

Source Language	Learning Rate	Epochs	Best Epoch	Validation Loss (Best)
English (EN)	$1e^{-5}$	8	5	0.2291
Latvian (LV)	$5e^{-6}$	8	5	0.2504
Swedish (SV)	$5e^{-6}$	5	5	0.2279
Finnish (FI)	$1e^{-5}$	8	5	0.2231
German (DE)	$1e^{-5}$	8	5	0.2789
Russian (RU)	$1e^{-5}$	8	5	0.2865
Hungarian (HU)	$1e^{-5}$	8	5	0.2202

This fine-tuning process ensured that the final models used for evaluation were not only trained under consistent low-resource conditions, but were also individually optimized for performance and convergence quality. The results of these models on the shared English–Estonian test set are discussed in the following section.

4.5 Evaluation

To monitor training progress and select the most effective checkpoints, validation loss was used as the primary indication of performance during fine-tuning. Evaluation and model saving were performed at the end of each epoch, and the checkpoint with the lowest validation loss was retained. Final evaluation was carried out on a held-out English–Estonian test set, consistent across all experiments, to ensure comparability between models fine-tuned on different source languages.

Two automatic evaluation metrics were used. sacreBLEU, a standardized implementation of BLEU (Post, 2018), measures n-gram overlap between model outputs and reference translations. It captures surface-level fluency and adequacy, particularly for high-resource or structurally similar language pairs. Additionally, chrF (Popović, 2015), a character n-gram F-score metric, was used due to its sensitivity to morphological variation and orthographic closeness. Given Estonian's rich inflectional morphology, chrF offers a valuable complementary perspective to BLEU by better capturing partial correctness and linguistic similarity at the subword level.

In addition to quantitative evaluation, a qualitative error analysis was performed on a representative sample of model outputs. The investigated categories were named entity translation, handling of morphological variations, word order, hallucinations and omissions, and orthographic consistency.

5 Results

This chapter presents the empirical findings of the study, focusing on how fine-tuning with different source languages affects Estonian translation quality in a multilingual NMT setting. The experiments were designed to test the hypothesis that genealogically related and areally proximate languages provide varying degrees of inductive transfer when used to fine-tune a pre-trained model for Estonian translation. The central objective is to quantify and analyze how source language choice influences the model's ability to generate Estonian, using consistent evaluation on a shared English–Estonian test set. This chapter begins with a quantitative analysis based on sacreBLEU and chrF scores, followed by a qualitative assessment of linguistic patterns in selected model outputs in the subsequent section.

5.1 Quantitative Analysis

The translation quality of each fine-tuned model is evaluated on the held-out English–Estonian test set using two automatic metrics: sacreBLEU and chrF. sacreBLEU measures surface-level n-gram overlap, reflecting lexical and syntactic accuracy, while chrF evaluates character n-gram similarity, which is especially sensitive to morphological correctness. Together, these metrics provide complementary perspectives on the quality of Estonian translation outputs. Table 2 summarizes the sacreBLEU and chrF scores for all experiments, including the zero-shot baseline, English fine-tuning, and cross-lingual fine-tuning with different source languages.

The zero-shot baseline, evaluated using the original NLLB-200-distilled-600M model without any fine-tuning, achieves a sacreBLEU score of 21.45 and chrF score of 58.99. This serves as the initial reference point for assessing the impact of task-specific adaption. Fine-tuning on English–Estonian data leads to a substantial performance increase, raising the scores to 25.60 sacreBLEU and 61.41 chrF, and setting an upper bound for expected performance when the test-time source language is included during fine-tuning.

Among the cross-lingual fine-tuning experiments, Hungarian yields the highest scores overall, achieving 26.36 sacreBLEU and 62.01 chrF. It is the only source language that outperforms English fine-tuning, suggesting particularly strong transfer effectiveness. Comparable, though slightly lower, results are observed with Swedish, Finnish, and German are used as source languages. German achieves the

Table 2: Test set results on EN–ET using models fine-tuned on different source-to-Estonian datasets

Source Language	sacreBLEU	chrF
Baseline: zeroshot EN–ET	21.45	58.99
English (EN)	25.60	61.41
Genealogical proximity		
Hungarian (HU)	**26.36**	**62.01**
Finnish (FI)	24.38	60.70
Areal proximity		
German (DE)	24.52	60.95
Swedish (SV)	24.08	61.06
Latvian (LV)	22.82	58.78
Russian (RU)	20.01	56.56

highest sacreBLEU score among these three with 24.52, followed by Finnish with 24.48 and Swedish with 24.08. In terms of chrF, Swedish performs best with 61.06, while German and Finnish fall just below the 61 mark, scoring 60.95 and 60.70, respectively. The lowest scores were obtained when Russian and Latvian were used as source languages. Latvian yields a slightly higher sacreBLEU score than the baseline, with 22.82 compared to 21.45. However, its chrF score falls just below the baseline, reaching 58.78. Russian is the only source language that performs significantly below the baseline on both metrics, achieving a sacreBLEU score of 20.01 and a chrF score of 56.56.

While the results suggest some variation in translation quality depending on the source language used for fine-tuning, the overall differences remain relatively modest across most conditions, ranging by approximately 5 score points in sacreBLEU and 5.5 for chrF. This may in part be attributed to the limited size of the training and evaluation data. Nevertheless, this setup reflects a realistic low-resource scenario and provides valuable insight into the behavior of multilingual models under constrained conditions. The use of a consistent evaluation set and standardized training setup across all language conditions allows for controlled and directly comparable results.

The strongest cross-lingual transfer effect is observed when fine-tuning on Hungarian, which outperforms even English fine-tuning. This is a notable result, as although both Hungarian and Estonian belong to the Uralic language family, they are not closely related within it. For instance, Finnish is genealogically much closer to Estonian than Hungarian is. If genealogical proximity were the dominant factor, one would expect Finnish to yield the highest performance. Instead, Hungarian achieves the top scores across both metrics. This suggests that genealogical related-

ness may contribute to transfer effectiveness but does not fully explain it. One plausible factor is the shared typological profile of Hungarian and Estonian, as both are agglutinative languages with rich inflectional morphology, which may support more effective parameter adaptation in a multilingual model under low-resource conditions. In addition, it is possible that the Hungarian–Estonian subset used for fine-tuning contained features such as higher lexical diversity more consistent alignments that facilitated stronger learning. This highlights the possibility that structural properties, combined with experimental factors such as data characteristics or model sensitivity, can sometimes outweigh genealogical distance in determining transfer success.

Finnish, German, and Swedish yield relatively strong results, performing just below the English fine-tuned model. Finnish is the closest genealogical relative to Estonian among the tested languages, and its performance supports the idea that linguistic relatedness can enhance transfer. Like Estonian, Finnish is an agglutinative language with rich inflectional morphology, which may support better parameter alignment during fine-tuning. However, its slightly lower scores compared to Hungarian suggest that genealogical closeness alone does not guarantee optimal transfer.

German and Swedish, while not genealogically related to Estonian, still achieve competitive results. This may be attributed to several factors, such as their status as high-resource languages in the model's pretraining corpus, which likely led to more robust internal representations. Additionally, both languages share surface-level typological features with Estonian, such as SVO word order and certain vocabulary overlaps, particularly due to historical borrowing. The strong performance of these typologically and genealogically diverse languages underscores that multiple factors, including data availability, model familiarity, and structural similarity, can interact to shape cross-lingual transfer success.

Latvian and Russian achieved the weakest results among the cross-lingual fine-tuning experiments. While Latvian attains a slightly higher sacreBLEU score than the zero-shot baseline, its chrF score remains lower, suggesting limited overall benefit. Russian, on the other hand, is the only source language to perform significantly below the baseline on both metrics, indicating negative transfer. These outcomes may reflect several factors. Although Latvian is geographically close to Estonian and shares regional contact in the Baltics, it lacks the morphological and structural similarity found in Uralic languages such as Finnish or Hungarian. This may limit the effectiveness of transfer at the subword and morphosyntactic level. For Russian, the relatively poor performance may be due to a combination of structural divergence, such as Slavic morphology and free word order, and orthographic distance, as Russian uses the Cyrillic script, while Estonian uses the Latin alphabet. These differences could cause greater representational mismatch in

the model's encoder-decoder alignment. Additionally, unlike the other languages, the Russian–Estonian training data was sourced from a different corpus, which may have impacted domain alignment and consistency during fine-tuning. This discrepancy in training material could have contributed to the weaker fine-tuning outcomes observed for Russian.

The quantitative results demonstrate that genealogical closeness appears to be a stronger predictor of cross-lingual transfer effectiveness than geographical proximity. Hungarian, despite being more distantly related to Estonian than Finnish, yields the best transfer improvements. Finnish, the closest genealogical relative among the tested languages, and Swedish, which is geographically close but genealogically distant, both show relatively strong transfer performance, highlighting that areal proximity can also have a moderate positive effect. German, though neither genealogically nor geographically close, also performs competitively, possibly due to its high-resource status in the model's pretraining and its historical influence on Estonian, which has resulted in lexical overlap. Most source languages tested outperform the zero-shot baseline, confirming that fine-tuning with related languages generally benefits translation quality. Latvian's moderate gains suggest that regional contact alone may not suffice to enable strong transfer without closer genealogical or typological similarity. Russian notably underperforms, reflecting both structural divergence and differences in training data, underscoring that linguistic distance and experimental factors jointly influence transfer success. These findings highlight the interplay between genealogical, typological, areal, and data-related factors in shaping cross-lingual adaptation in LRNMT.

5.2 Qualitative Analysis

A qualitative error analysis was carried out on the model outputs to gain insight into recurring patterns that affect translation quality and the types of errors produced. A subset of each model's output was examined individually. Errors in each sentence were identified, categorized, and then compared across models to highlight differences in performance. The analysis focused on several dimensions, including named entity recognition, morphological accuracy, word order, lexical choices, and orthographic correctness. These categories were selected to capture the main sources of error in Estonian translations and to highlight differences between transfer settings.

5.2.1 Named Entity Translation

Named entity translation posed challenges for all fine-tuned models. While entities such as personal names were generally preserved, institutional or policy names were often mistranslated, inconsistently rendered, or distorted through morphological or orthographic errors.

Table 3: Comparison of named entity translation across models (example sentence number 1)

Source (EN)	**The Fit for Future Platform** succeeds the **REFIT platform**, building on its experiences.
Reference (ET)	**Platvorm „Fit for Future"** asendab **REFITi platvormi** ja tugineb selle kogemustele.
Zeroshot	**Fit for Future platvorm** on **REFIT platvormi** järeltulija, võttes aluseks selle kogemusi.
EN–ET	**Suurepärane tulevikuplatvorm** järgneb **REFIT platvormi**, tuginedes selle kogemustele.
HU–ET	**Plattform Suuneb tulevikuks"** järgneb **REFIT platvormile** ja põhineb selle kogemustel.
FI–ET	**Plattvorm sobivad tulevikule** järgib **REFIT platvormi** ja toetub selle kogemustele.
DE–ET	**Suurepärane tulevikuplatvorm** järgib **REFIT platvormi**, võttes selle kogemuste põhjal kasutusele.
SV–ET	**Suurepärane tulevikuplatvorm** on selle kogemuste põhjal **REFIT platvormi** järeltulija.
LV–ET	Suunistes välja töötatud **platvormi Fit for Future" (Kohas tulevikuks)** on saanud järeltulija **REFIT platvormi**, võttes selle kogemustest välja.
RU–ET	**Suuneb tulevikuks sobiv platvorm REFIT platvormi** järel ja põhineb selle kogemustel.

A first example is shown in Table 3, where the phrase *Fit for Future Platform* was incorrectly handled in all cases. Only the zero-shot baseline attempted a literal transfer, while fine-tuned models tended to paraphrase or replace the entity with semantically odd expressions such as *Suurepärane tulevikuplatvorm* ("excellent future platform") or *Plattform Suuneb tulevikuks*, the latter being both incorrect and largely unintelligible. The Latvian–Estonian finetuned model attempted to preserve the original entity but added mistranslated elements. These outputs illustrate a tendency for models to normalize unfamiliar named entities into more natural-sounding, yet incorrect, constructions. The second named entity in this example, *REFIT platform*, was generally correctly preserved, but all models failed to correctly conjugate the name, producing form such as *REFIT platvormi* instead

of the correct *REFITi platvormi*, resulting in a mismatch with Estonian case inflection.

Table 4: Comparison of named entity translation across models (example sentence number 2)

Source (EN)	**Janez Lenarčič**, **Commissioner for Crisis Management**, said:"Sanctions should not impede the delivery of humanitarian assistance, including medical assistance, in line with International Humanitarian Law.
Reference (ET)	**Kriisiohjevolinik Janez Lenarčič** ütles: „Rahvusvahelise humanitaarõiguse kohaselt ei tohiks sanktsioonid takistada humanitaarabi, sealhulgas arstiabi andmist.
Zeroshot	**Kriisijuhtimise volinik Janez Lenarčič** ütles: "Sanktsioonid ei tohiks takistada humanitaarabi, sealhulgas meditsiinilist abi, osutamist kooskõlas rahvusvahelise humanitaarabiga.
EN–ET	**Kriisiohje volinik Janez Lenarčič** ütles: Sanktsioonid ei tohiks takistada humanitaarabi, sealhulgas meditsiinise abi, osutamist kooskõlas rahvusvahelise humanitaarabega.
HU–ET	**Kriisiohje volinik Janez Lenarčič** lisas: "Sanktsioonid ei tohiks takistada humanitaarabi, sealhulgas meditsiiniabi, osutamist kooskõlas rahvusvahelise humanitaarabega.
FI–ET	**Kriisiohje volinik Janez Lenarčič** ütles: Sanktsioonid ei tohiks takistada humanitaarabi, sealhulgas meditsiinitoetuse, osutamist kooskõlas rahvusvahelise humanitaaravaga.
DE–ET	**Kriisiohjevolinik Janez Lenarčič** ütles: "Kohustusrakkumised ei tohiks takistada humanitaarabi, sealhulgas meditsiinitarbi, osutamist kooskõlas rahvusvahelise humanitaarabi õigusega.
SV–ET	**Kriisiohjamise volinik Janez Lenarčič** ütles: Sanktsioonid ei tohiks takistada humanitaarabi, sealhulgas meditsiiniabi, osutamist kooskõlas rahvusvahelise humanitaarabega.
LV–ET	**Kriisiohjamise volinik Janez Lenarčič** lisas: Sanktsioonid ei tohiks takistada humanitaarabi, sealhulgas meditsiinitarbi, osutamist kooskõlas rahvusvahelise humanitaarabega.
RU–ET	**Kriisijuhtimise volinik Janez Lenarčič** ütles: "Sanktsioonid ei tohiks takistada rahvusvahelise humanitaarõigusega kooskõlas humanitaarbi, sealhulgas meditsiinilise abi, andmist.

A second example, shown in Table 4, illustrates how the models handle personal names and formal titles. All systems correctly preserve the name *Janez Lenarčič*, indicating that personal names are generally easier for the models to transfer than institutional or policy names. However, variation is observed in the translation of the title *Commissioner for Crisis Management*, with outputs such as *kriisiohjevolinik*,

kriisijuhtimise volinik, or *Kriisiohjamise volinik*. These differences reflect subtle morphological and orthographic inconsistencies.

A third example, in Table 5, demonstrates the translation of *the Farm to Fork Strategy*. Here, none of the models succeeded in producing the correct Estonian term, *"Talust taldrikule"*. Instead, all outputs contain highly unnatural compounds such as *Põllumajandusettevõtte-kelle strateegia* or *Põllumaast torki strateegia*. These forms combine fragments of plausible agricultural terminology with nonsensical constructions, indicating both lexical hallucination and morphological errors.

Table 5: Comparison of named entity translation across models (example sentence number 3)

Source (EN)	**The Farm to Fork Strategy** aims to promote a global transition to sustainable food systems, in close cooperation with its international partners.
Reference (ET)	**Strateegia „Talust taldrikule"** eesmärk on edendada tihedas koostöös rahvusvaheliste partneritega ülemaailmset üleminekut kestlikele toidusüsteemidele.
Zeroshot	**Põllumajandusettevõtte-kork-strateegia** eesmärk on edendada ülemaailmset üleminekut jätkusuutlike toidussüsteemide poole, tehes seda tihedal koostöös oma rahvusvaheliste partneritega.
EN–ET	**Põllumajandusettevõtte-kelle strateegia** eesmärk on edendada ülemaailmset ülemineku kestlike toidussüsteemide suundule tihedas koostöös rahvusvaheliste partneritega.
HU–ET	**Põllumajandusettevõtte-kest-kiiruse strateegia** eesmärk on edendada tihedas koostöös oma rahvusvaheliste partneritega ülemaailmset ülemineku kestlikele toidussüsteemidele.
FI–ET	**Strateegia põllumajandusettevõte-kõrguni** eesmärk on edendada ülemaailmset ülemineku kestlike toidussüsteemide suures koostöös oma rahvusvaheliste partneritega.
DE–ET	**Põllumajandustootmise-korkistrateegia** eesmärk on edendada ülemaailmset ülemineku kestlike toidussüsteemide suures koostöös oma rahvusvaheliste partneritega.
SV–ET	**Strateegia põllumajandusettevõte-kiiruse strateegia** eesmärk on edendada ülemaailmset üleminekut kestlike toidussüsteemide suunas koostöös oma rahvusvaheliste partneritega.
LV–ET	**Strateegia põllumajandusettevõtted torki** jaoks on ette nähtud selleks, et edendada ülemaailmset ülemineku kestlike toidussüsteemide suunas koostöös rahvusvaheliste partneritega.
RU–ET	**Põllumaast torki strateegia** eesmärk on edendada ülemaailmset üleminekut jätkusuutlikesse toidussüsteemidele tihedas koostöös oma rahvusvaheliste partneritega.

5.2.2 Morphological Comparison

A prominent type of morphological error observed across models concerns case agreement between adjectives and nouns. In the example shown in Table 6, nearly all systems produce the phrase *järgmiste viie aasta jooksul* ("of the following five years"), whereas the correct form in Estonian would be *järgmise viie aasta jooksul* ("for the next five years"). The mistake arises because the models treat *järgmine* ("next") as if it required the plural genitive ending *-te*, aligning it with *aasta(d)*, rather than the singular genitive *-se*. This results in ungrammatical output, even though the intended meaning remains recoverable.

Interestingly, this error is systematic across languages, suggesting that fine-tuning did not help models internalize the relatively rigid agreement patterns of Estonian genitives. Instead, the models tend to overgeneralize morphological agreement, producing surface forms that appear plausible but are grammatically incorrect. The example illustrates how morphosyntactic mismatches, particularly in case and number agreement, remain a persistent challenge for MT into morphologically rich languages like Estonian.

In addition to adjective–noun agreement issues, case government with nouns in complex infinitive constructions also revealed systematic errors across models. A second recurrent issue in this sentence concerns case government with the noun *ohver* ("victim"). The reference construction is *võimestada ohvreid kuriteost teada andma*, where *ohvreid* appears in the partitive plural. Some models, such as FI–ET and DE–ET, correctly produce *ohvritel*/*ohvreid* in the partitive or otherwise appropriate form, showing successful morphological adaptation. In contrast, other systems, including LV–ET and RU–ET, assign an incorrect case that is neither nominative nor partitive, producing forms that are ungrammatical in context. Many outputs, like EN–ET or HU–ET, simply use nominative plural (*ohvrid teatama*), reflecting a failure to capture the case requirements imposed by the infinitive phrase. This illustrates that while some models can handle complex case government, others struggle, particularly under low-resource cross-lingual transfer conditions.

Additionally, errors also appear in other case constructions, such as the relative clause *mis keskenduvad kahele eesmärgile*. In models the zero-shot baseline and DE–ET model, the case of *kahele eesmärgile* does not properly align with the governing verb, resulting in ungrammatical forms. This further highlights that morphological agreement challenges are widespread, affecting multiple syntactic environments beyond infinitive phrases.

Table 6: Comparison of morphological issues across models

Source (EN)	The strategy sets a number of actions **for the next five years**, **focusing on two objectives**: first, **to empower victims to report crime**, **claim** compensation and ultimately **recover** from consequences of crime; second, to work together with all relevant actors for victims' rights.
Reference (ET)	Strateegiaga on **järgmiseks viieks aastaks** ette nähtud mitu meedet, mis **keskenduvad kahele eesmärgile**: esiteks, **võimestada ohvreid kuriteost teada andma**, hüvitist **nõudma** ja lõpuks kuriteo tagajärgedest **taastuma**; teiseks, teha ohvrite õiguste valdkonnas koostööd kõigi asjakohaste osalejatega.
Zeroshot	Strateegia sätestab **järgmise viie aasta jooksul** mitmeid meetmeid, mille **eesmärk on kahte eesmärki**: esmalt **võimaldada ohvrid teatama kuritegudest**, **nõuda** hüvitist ja lõpuks **taastada** kuritegude tagajärgedest; teiseks töötada koos kõigi asjaomaste osapooltega ohvrite õiguste eest.
EN–ET	Strateegia näitab ette mitmeid meetmeid **järgmise viie aasta jooksul**, mille **keskmes on kaks eesmärki**: esiteks **võimaldada ohvrid teatama kuritegudest**, **nõuda** hüvitist ja lõpuks **taastada** kuritegude tagajärgede eest; teiseks, töötada koos kõigi asjaomaste osalistega ohvrite õiguste kaitseks.
HU–ET	Strateegias on ette nähtud **järgmiste viie aasta** tegevuskavade arv, mille **keskmes on kaks eesmärki**: esiteks **võimaldada ohvrid teatama kuritegudest**, **taotlema** hüvitist ja lõpuks **taastuma** kuritegude tagajärgedest; teiseks on töötada koos kõigi asjaomaste osapooltega, et kaitsta ohvrite õigusi.
FI–ET	Strateegias on sätestatud **järgmiste viie aasta jooksul** mitmeid meetmeid, mille **keskmes on kaks eesmärki**: esiteks **võimaldada ohvritel teatada kuritegudest**, **nõuda** hüvitist ja lõpuks **taastada** kuritegude tagajärgedest; teiseks töötada koos kõigi asjaomaste osalistega ohvrite õiguste nimel.
DE–ET	Strateegias on ette nähtud **järgmiste viie aasta jooksul** mitmeid meetmeid, mille eesmärk on **keskenduda kahte eesmärki**: esmalt **võimaldada ohvritel teatada kuritegevusest**, **nõuda** hüvitist ja lõpuks **taastada** kuritegevuse tagajärgedest; teiseks, töötada koos kõigi asjaomaste osalistega võitlemiseks ohvritele õiguste nimel.
SV–ET	Strateegias on sätestatud **järgmiste viie aasta jooksul** mitmeid meetmeid, mille **keskmes on kaks eesmärki**: esiteks **võimaldada ohvreid teatama kuritegudest**, **taotlema** hüvitist ja lõpuks **taastuma** kuriteo tagajärgedest; teiseks töötada kõigi asjaomaste osalistega koos ohvrite õiguste kaitsmiseks.
LV–ET	Strateegias sätestatakse **järgmise viie aasta jooksul** mitmeid meetmeid, mille **keskmes on kaks eesmärki**: esiteks, **võimaldada ohvritele teatada kuritegudest**, **taotleda** hüvitist ja lõpuks **taastada** kuritegude tagajärgedest; teiseks, töötada koos kõigi asjaomaste teguritega ohvrite õiguste nimel.
RU–ET	Strategia sätestab **järgnevate viie aasta jooksul** mitmeid meetmeid, mille **keskendus on kahes eesmärgis**: esmalt **ohvrite võimaldada teatada kuritegudest**, **nõuda** hüvitist ja lõpuks **taastada** kuritegevuse tagajärgedest; teiseks, töötada koos kõigi asjakohaste teguritega ohvrite õiguste eest.

A third recurrent morphological challenge involves verb forms in infinitive constructions, particularly those ending in -ma or -da. In the reference sentence, verbs such as *võimestada*, *nõudma*, and *taastuma* appear in specific infinitive forms dictated by the syntactic context. Several models, including EN–ET and HU–ET, fail to produce the correct forms, either using inappropriate infinitives or mismatching endings, leading to ungrammatical sequences. In contrast, FI–ET and DE–ET generally produce the correct forms, demonstrating better adaptation to the morphological requirements of Estonian infinitives. LV–ET and RU–ET frequently generate incorrect forms, further illustrating that low-resource cross-lingual transfer can struggle with complex verbal morphology.

5.2.3 Word Order

Word order errors in the model outputs are generally stylistic rather than strictly grammatical, but they can affect readability and comprehension, particularly in sentences with multiple clauses, temporal expressions, or numerical data. Estonian allows relatively free constituent order compared to English, but it also has clear preferences for neutral and natural sequencing, especially in adverbial placement, subordinate clauses, and reporting structures. In this section, we focus on deviations that reduce clarity or distort emphasis to the extent that they interfere with understanding, rather than on purely stylistic variation.

The example in Table 7 illustrates these issues. In the reference sentence, temporal expressions, numerical values, and subordinate clauses are arranged in a natural sequence to convey the forecasted economic growth. Some models, such as EN–ET, FI–ET, and DE–ET, maintain an understandable order, with only minor deviations from natural Estonian phrasing. In contrast, outputs from the zero-shot baseline, RU–ET, and LV–ET show more pronounced disruptions, including misplaced verbs, shifted temporal modifiers, and awkward numerical placements, which reduce clarity and readability. These patterns suggest that while fine-tuning improves syntactic acceptability, capturing Estonian's preferred clause and modifier order remains challenging under low-resource cross-lingual transfer conditions.

The most severe word order problems, where comprehension is directly impaired, occur primarily in LV–ET and RU–ET. This observation aligns with their comparatively low automatic evaluation scores, suggesting that word order disruptions are a key contributor to poor overall performance. By contrast, higher-scoring models such as FI–ET and DE–ET produce sentences with fewer or only minor stylistic deviations, indicating a correlation between metric-based quality and the extent of word order issues. Lower-scoring systems not only make more lexical and morphological errors, but also struggle more with organizing clauses and modifiers into intelligible sequences, highlighting the interplay between syntactic control and overall translation quality.

Table 7: Comparison of word order choices across models

Source (EN)	For the EU as a whole, growth is forecast to ease marginally to 1.4% in 2020 and 2021, down from 1.5% in 2019.
Reference (ET)	ELis tervikuna peaks majanduskasv 2020. ja 2021. aastal pisut vähenema ning püsima 1,4% juures, mis on väiksem näitaja kui 2019. aasta 1,5%.
Zeroshot	Euroopa Liidu puhul prognoosib majanduskasv 2020. ja 2021. aastal veidi kergemaks, vähenes 1.5%-st 2019. aastal.
EN–ET	Üldiselt prognoositakse, et ELi majanduskasv leeveneb 2020. ja 2021. aastal veidi 1,4 protsendi võrra, võrreldes 2019. aasta 1,5 protsendiga.
HU–ET	ELi kui terviku puhul eeldatakse, et majanduskasv kergendub 2020. ja 2021. aastal vähenevalt, jõudes 1,4%-le, võrreldes 2019. aastal 1,5%-ga.
FI–ET	ELi puhul prognoositatakse, et majanduskasv kergendub 2020. ja 2021. aastal veidi 1,4%-le, võrreldes 2019. aasta 1,5%-ga.
DE–ET	ELi puhul prognoositatakse, et majanduskasv kergendub 2020. ja 2021. aastal veidi 1,4%-ni, võrreldes 2019. aasta 1,5%-ga.
SV–ET	Prognooside kohaselt leeveneb majanduskasv 2020. ja 2021. aastal mõningalt ja jõuab 1,4%-ni, võrreldes 2019. aastal 1,5%-ga.
LV–ET	ELi puhul prognoosib majanduskasv 2020. ja 2021. aastal vähenevalt 1,4%-ni, võrreldes 1,5%-ga 2019. aastal
RU–ET	Euroopa Liidu majanduskasvu prognoosib vähene vähendamist 2020. ja 2021. aastal 1,4%-ni, võrreldes 1,5%-ga 2019. aastal.

5.2.4 Lexical Choice

Lexical accuracy represents another major challenge for the models. This category covers errors involving incorrect or awkward word choices, mistranslations of key terms, and distortions of meaning. It also includes more severe problems such as hallucinations, where the model generates content absent from the source, as well as omissions, where essential information is dropped. In some cases, models also produced repetitive or looping sequences, further undermining readability. These errors are particularly detrimental because they directly affect the semantic fidelity of the translation, making the output misleading or incoherent even when morphology or word order are otherwise acceptable.

In enumerative context, such as lists of protective equipment, all models struggled to maintain lexical accuracy and terminological consistency, as shown in Table 8. The reference sentence lists concrete, domain-specific terms such as *maskid*, *kaitseprillid ja -sirmid*, *näokaitsmed*, *suu- ja ninakaitsevahendid*, and *kaitseriietus*. By contrast, system outputs frequently contain hallucinated forms that do not exist in Estonian, including *kaitsekõidud*, *kaitsekõsed*, and *kaitsekõbed*. These pseudo-words mimic compound formation but are nonsensical, reflecting failed attempts to approximate technical terminology. Other outputs substitute incorrect existing

words, e.g. *viirused* ("viruses") instead of *visiirid* ("visors") and *kaitselinnad* (translateable as "defence city") instead on *kaitseprillid* ("protective spectacles"), or omit entire items from the list. Across models, this combination of fabricated, mistranslated, or omitted terms severely distort the meaning of the source. While the overall sentence structure often remains intact, the accumulation of lexical errors makes the translations difficult to interpret and undermines trust in the output for technical or regulatory texts.

Table 8: Comparison of lexical choices across models

Source (EN)	Personal Protective Equipment covers equipment such as masks, protective spectacles and visors, face shields, mouth-nose protection and protective garments.
Reference (ET)	Isikukaitsevahendite hulka kuuluvad sellised vahendid nagu maskid, kaitseprillid ja -sirmid, näokaitsmed, suu- ja ninakaitsevahendid ja kaitseriietus.
Zeroshot	Isikliku kaitsevahend hõlmab maske, kaitseklaasid ja -näitsid, näo- ja nina kaitset ja kaitsekõidud.
EN–ET	Isikliku kaitsevahendite hulka kuuluvad sellised seadmed nagu maskid, kaitsbrillid ja viirid, nägu kaitsed, suu-naise kaitse ja kaitsekõsed.
HU–ET	Isiklike kaitsevahendite hulka kuuluvad sellised seadmed nagu maskid, kaitsebrillid ja -näitsid, näoga kaitse, suu-naus kaitse ja kaitsevahendid.
FI–ET	Isikukaitsevahend hõlmab selliseid varustusi nagu maskid, kaitseprillid ja -näitsid, näokindlad, suu-naase kaitsevahend ja kaitsevahendid.
DE–ET	Isikukaitsevahend hõlmab selliseid vahendeid nagu maskid, kaitsbrilli ja viirused, näoga kaitseskeed, suu-neesi kaitseskeed ja kaitsekõbed.
SV–ET	Isikukaitsevahendid hõlmavad selliseid vahendeid nagu maskid, kaitselinnad ja viirused, näokindlad, suu-naisi kaitsevahendid ja kaitseluba.
LV–ET	Isiku kaitsevahend hõlmab selliseid vahendeid nagu maskid, kaitsebrillid ja visiirid, näokindlad, suu-nausekaitsevahend ja kaitsepikkad.
RU–ET	Isiklikku kaitseküsimust hõlmab selliseid seadmeid nagu maski, kaitselinnasid ja -kaitsed, näokindlad, suu-naus-kaitseküsimus ja kaitsekõidud.

Another type of lexical error observed in multiple models was looping or uncontrolled repetition of tokens. This phenomenon occurred in three different systems: twice in LT–ET, once in FI–ET, and in EN–ET. In these cases, the model repeatedly generated the same word or phrase (*erakorraline*, "emergency") dozens of times, or cycled through fragments of regulation references, e.g. *1280/2013, 1280/2013*. Such outputs are unusable, as the repetition entirely blocks interpretation of the sentence. While rare in the dataset, these cases highlight a fragility of sequence generation in low-resource or domain-mismatched conditions. This issue is qualitatively different from ordinary mistranslations or lexical substitutions, since it reflects a failure of

generation control rather than lexical choice per se, but its inclusion here underlines the range of lexical disruptions observed.

Beyond hallucinations and repetition, a subtler but recurring error type was the substitution of semantically related but stylistically inappropriate synonyms. In Table 7, the reference sentence uses the verb *vähenema* ("to decrease"), which is the conventional choice in economic forecasting contexts. Several model outputs instead introduce alternatives such as *leevendub* ("alleviates") or *kergendub* ("lightens"), which are technically grammatical but sound unnatural and misleading in this domain. These deviations highlight that lexical substitution errors are often subtler than outright hallucinations. The outputs remain interpretable but fail to capture the idiomatic register of the target language. While nearly all models occasionally select awkward wording, the most problematic patterns were observed in the LV–ET and RU–ET outputs, where lexical errors frequently co-occur with morphological mistakes and disrupted word order, compounding their negative impact on overall quality. This reinforces the broader finding that these two systems consistently underperformed across evaluation dimensions.

5.2.5 Orthographic Issues

Orthographic issues in the model outputs are generally less frequent than lexical or word order problems, but they still affect the readability and perceived reliability of translations. Most of these errors involve misspellings or malformed compounds that result in non-existent words, often resembling plausible Estonian forms but deviating from correct orthography. In some cases, models also mishandle punctuation, for example by omitting necessary commas, misplacing colons or semicolons in lists, or inconsistently using decimal separators. While these issues rarely block comprehension, they give the output a machine-like appearance and undermine its suitability for formal or technical contexts where orthographic precision is expected.

A representative example of orthographic errors is found in the translation of the verb expressing "to recruit." In Estonian, the correct infinitive form for this context is *värvata*. Only the FI–ET model produced this correct form. All other models, including the zero-shot baseline, EN–ET, HU–ET, DE–ET, and SV–ET, produced *värbada*, an incorrect approximation that violates standard infinitive spelling rules. LV–ET and RU–ET outputs went further, generating *värbida*, a completely non-existent form in Estonian. Notably, the reference sentence avoids this issue by using the idiomatic construction *tööle võtma*, which is also correct and semantically equivalent. This example illustrates that models often fail to produce accurate Estonian infinitives, leading to orthographically invalid forms. While comprehension may remain possible, these errors reduce the perceived reliability and formal correctness of the translation.

Table 9: Comparison of orthographic issues across models (example sentence)

Source (EN)	The money will, for example, enable EPPO **to recruit** more quickly qualified staff and to buy IT equipment to start processing the first cases.
Reference (ET)	Seda summat kasutatakse näiteks selleks, et Euroopa Prokuratuur saaks **võtta** kiiremini **tööle** kvalifitseeritud töötajaid ja osta IT-vahendeid esimeste juhtumite uurimise alustamiseks.
Zeroshot	Näiteks võimaldab EPPO rahaliselt kiiremini **värbada** kvalifitseeritud töötajat ja osta IT-seadmeid, et alustada esimese juhtumi menetlemist.
EN–ET	Näiteks võimaldab vahendid Euroopa Prokuratuurile kiiremini **värbada** kvalifitseeritud töötajat ja osta IT-vahendeid, et alustada esimese kohtuasja menetlemist.
HU–ET	Näiteks võimaldab see Euroopa Prokuratuurel kiiresti **värbada** kvalifitseeritud töötajaid ja osta IT-varustust, et alustada esimese juhtumi menetlemist.
FI–ET	Näiteks võimaldab see raha Euroopa Üldprokurüarsusele kiiremini **värvata** kvalifitseeritud töötajad ja osta IT-varustust, et alustada esimese juhtumi menetlemist.
DE–ET	Näiteks võimaldab rahaline vahend Euroopa Prokuratuurel kiiresti **värbada** kvalifitseeritud töötajat ja osta esimese juhtumi menetlemiseks IT-seadmeid.
SV–ET	See raha võimaldab näiteks Euroopa Prokuratuuril kiiremini **värbada** kvalifitseeritud töötajaid ning osta IT-vahendeid, et alustada esimese juhtumi menetlemist.
LV–ET	Raha võimaldab näiteks Euroopa Prokuratuurile kiiremini **värbida** kvalifitseeritud töötajaid ja osta esimese juhtumi menetlemise alustamiseks IT-seadmeid.
RU–ET	Raha võimaldab näiteks EPPO-l kiiremini **värbida** kvalifitseeritud töötajaid ja osta IT-seadmeid, et alustada esimese juhtumi menetlemist.

Another common type of orthographic error involves simple misspellings of standard terms. For instance, in Table 8, the reference sentence uses *kaitseprillid* ("protective spectacles"), a standard compound in Estonian. All models except FI–ET incorrectly render this word as *kaitsebrillid* or similar variants, demonstrating a consistent orthographic deviation. Similarly, in another sentence, the term *toidusüsteemid* ("food systems") is misspelled by most models as *toidussüsteemid*, with only HU–ET producing the correct form. These errors highlight that even common or semantically transparent words can be mishandled, affecting the readability and perceived reliability of the output.

In addition, proper nouns, domain-specific terminology, and complex compounds also pose challenges. For example, translations of "coronavirus (pandemic)" vary across systems, with outputs including *koronaviruse*, *koronaviruse pandemia*,

koronaviruspandemie, and *koronavirusuga*, reflecting inconsistent attempts to render the term in correct Estonian, including matching the intended grammatical case. Zero-shot models and RU–ET struggle particularly with these items. Overall, RU–ET is notable for generating the most frequent and diverse orthographic mistakes, including both minor spelling deviations and malformed compounds, underscoring the model's relative fragility in orthographic accuracy compared to other systems.

6 Discussion

This master's thesis set out to investigate whether genealogical and areal relationships influence the effectiveness of transfer learning in LRNMT, using Estonian as a test case. The results indicate that genealogical proximity plays a measurable role, with Hungarian and Finnish providing the most consistent improvements within this setup. Areal proximity, while somewhat beneficial, proved more uneven. German and Swedish produced moderate improvements, but Latvian and Russian frequently underperformed, both in terms of automatic metrics and qualitative evaluation. This can be partly explained by the nature of the areal connections. Latvian and Estonian are neighboring languages, but share relatively little vocabulary due to limited historical influence, which reduces the effectiveness of transfer. Russian has contributed more loanwords to Estonian, yet differences in grammatical structure and script limit transferability. Swedish and German, although geographically more distant, have exerted stronger historical influence on Estonian and share more vocabulary, in addition to using the same script, which may account for their comparatively better performance. It is also possible that the model's pretraining on German and Swedish was more extensive or domain-representative than for Russian or Latvian, which could have further contributed to the observed differences, though this effect is difficult to quantify in the current setup.

These findings align with prior research that highlights the utility of related languages in transfer learning. Dabre et al. (2017) and Rikters et al. (2018) similarly observed that shared linguistic structures improve cross-lingual transfer, particularly in low-resource conditions. By contrast, Kocmi and Bojar (2018) demonstrated that transfer can succeed even without relatedness, but the present results add to this claim by showing that while relatedness is not strictly necessary, it often enhances performance and reduces lexical and orthographic instability. The role of Hungarian is especially noteworthy. Despite being typologically related but geographically distant, it provided the strongest performance, suggesting that genealogical factors outweigh areal contact when the two are compared directly.

The qualitative analysis offers additional insights into the types of errors characteristic of different transfer settings. Genealogically related models, Hungarian–Estonian and Finnish–Estonian, not only scored higher but also produced translations that were more natural in terms of word order and morphology, with fewer critical lexical errors. Areal transfer, by contrast, was more prone to word-order disruptions, problematic synonym choices, and orthographic mistakes,

particularly in the case of Russian and Latvian. This suggests that while contact languages may share some vocabulary with Estonian, their divergent morpho-syntactic structures limit their usefulness for transfer learning. A finding was the relatively poor performance of Russian, despite its status as a major contact language. This is plausibly connected to the domain mismatch in the training data, as well as to structural differences that complicate parameter transfer.

These results underscore the importance of carefully selecting transfer languages in low-resource scenarios. Genealogically related sources, even when not areally proximate, offer a stronger foundation than contact languages with high lexical overlap but divergent grammatical systems. More broadly, the study supports the view that linguistic knowledge, particularly genealogical relationships, can guide the design of transfer learning pipelines for low-resource languages. This complements data-driven methods such as large-scale synthetic corpora (Korotkova and Fishel, 2024) and multilingual joint training (Rikters et al., 2018), offering a more resource-efficient alternative in cases where such data is unavailable.

At the same time, several limitations temper the generalizability of these findings. First, the experiments were conducted with relatively small datasets: 1,000 sentence pairs for training and 200 each for validation and test. This setup was deliberately chosen to simulate a low-resource scenario and to remain computationally feasible. However, the limited data size constrains the robustness of the conclusions, as larger datasets would likely produce more stable estimates of performance and could shift the relative effectiveness of different source languages. Additionally, the use of validation loss alone to select the optimal epoch may have limited the reliability of the model selection, and incorporating an evaluation metric on the validation set would have provided a more robust criterion.

Second, not all language pairs were drawn from the same domain. All data except for Russian–Estonian came from the ELRC Presscorner_COVID corpus, which consists of EU press releases on COVID-19. By contrast, the Russian–Estonian data had to be sourced from Wikimedia due to the lack of available ELRC material. This mismatch introduces potential domain effects, making Russian less directly comparable to the other languages and likely contributing to its weaker performance.

Third, as part of preprocessing, sentence pairs exceeding 250 tokens in either language were removed. While this step ensured data cleanliness and computational manageability, it disproportionately excluded long or complex sentences. Consequently, the evaluation may not fully capture model behavior on structurally demanding inputs, which are often where translation systems struggle most.

Taken together, these limitations mean that the results should be interpreted as indicative rather than definitive. Nevertheless, the consistency of the observed trends across both quantitative metrics and qualitative error analysis suggests that the main conclusions remain reasonably well supported.

In sum, this master's thesis contributes to ongoing debates about the role of linguistic similarity in transfer learning. It shows that genealogical relatedness, more than areal contact, improves translation quality for Estonian, particularly in terms of word order, morphology, and lexical accuracy. These findings extend previous work by demonstrating that linguistic relatedness is not only theoretically significant but also practically beneficial for building translation systems for low-resource Uralic and contact languages.

7 Conclusion

This thesis investigated the role of linguistic similarity in transfer learning for LRNMT, with Estonian as the target language. Using the NLLB-200 model, a series of fine-tuning experiments were conducted with source languages chosen for either genealogical relatedness (Hungarian, Finnish) or areal proximity (German, Swedish, Latvian, Russian), alongside English as a direct baseline.

The results show that genealogical similarity offers the most consistent benefits, with Hungarian and Finnish yielding stronger improvements in translation quality than areal contact languages. While some areal sources such as German and Swedish provided moderate gains, Latvian and Russian were less effective, suggesting that shared vocabulary alone does not compensate for structural divergence. Qualitative analysis further indicated that genealogically related models produced translations with more accurate word order, morphology, and lexical choices, whereas areal models were more prone to orthographic and word-order errors.

These findings demonstrate that genealogical relatedness is a valuable criterion for selecting transfer languages in low-resource settings, especially when large-scale synthetic data or multilingual joint training are not available. At the same time, the study highlights the limits of areal similarity for transfer learning, showing that lexical overlap without structural compatibility yields limited advantages.

Future research could expand on these findings by testing larger datasets, exploring additional domains, and examining how genealogical transfer interacts with complementary approaches such as data augmentation. By clarifying the role of genealogical and areal similarity, this work provides a basis for more principled transfer strategies in the development of translation systems for Estonian and other low-resource languages.

Bibliography

Roee Aharoni, Melvin Johnson, and Orhan Firat (2019). Massively multilingual neural machine translation. In *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)*, pages 3874–3884, Minneapolis, Minnesota. Association for Computational Linguistics.

Robert B. Allen (1987). Several studies on natural language and back-propagation: Natural language translation. In *Proceedings of the First International Conference on Neural Networks*, pages 338–339. IEEE.

Laith Alzubaidi, Jiaming Bai, Ali Al-Sabaawi, et al. (2023). A survey on deep learning tools dealing with data scarcity: definitions, challenges, solutions, tips, and applications. *Journal of Big Data*, 10(46).

Farid Arthaud, Rachel Bawden, and Alexandra Birch (2021). Few-shot learning through contextual data augmentation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1049–1062, Online. Association for Computational Linguistics.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio (2015). Neural machine translation by jointly learning to align and translate. In *Proceedings of the International Conference on Learning Representations (ICLR)*, San Diego, CA.

Dik Bakker (2010). Language sampling. In Jae Jung Song, editor, *The Oxford Handbook of Linguistic Typology*, pages 100–127. Oxford University Press.

Marta Bañón, Pinzhen Chen, Barry Haddow, Kenneth Heafield, Hieu Hoang, Miquel Esplà-Gomis, Mikel L. Forcada, Amir Kamran, Faheem Kirefu, Philipp Koehn, Sergio Ortiz Rojas, Leopoldo Pla Sempere, Gema Ramírez-Sánchez, Elsa Sarrías, Marek Strelec, Brian Thompson, William Waites, Dion Wiggins, and Jaume Zaragoza (2020). ParaCrawl: Web-scale acquisition of parallel corpora. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4555–4567, Online. Association for Computational Linguistics.

Gábor Bella, Erdenebileg Byambadorj, Yamini Chandrashekar, Khuyagbaatar Batsuren, Danish Cheema, and Fausto Giunchiglia (2022). Language diversity:

Visible to humans, exploitable by machines. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 156–165, Dublin, Ireland. Association for Computational Linguistics.

Y. Bengio, P. Simard, and P. Frasconi (1994). Learning long-term dependencies with gradient descent is difficult. *IEEE Transactions on Neural Networks*, 5(2):157–166.

Laurent Besacier, Etienne Barnard, Alexey Karpov, and Tanja Schultz (2014). Automatic speech recognition for under-resourced languages: A survey. *Speech Communication*, 56:85–100.

Juliane Besters-Dilger, Cynthia Dermarkar, Stefan Pfänder, and Achim Rabus, editors (2014). *Congruence in Contact-Induced Language Change: Language Families, Typological Resemblance, and Perceived Similarity*, volume 27 of *Language Contact and Bilingualism*. De Gruyter, Berlin/Boston.

Balthasar Bickel (2007). Typology in the 21st century: Major current developments. *Linguistic Typology*, 11:239–251.

Balthasar Bickel and Johanna Nichols (2007). Inflectional morphology. In Timothy Shopen, editor, *Language Typology and Syntactic Description, Vol. III: Grammatical Categories and the Lexicon*, pages 169–240. Cambridge University Press, Cambridge.

Steven Bird (2020). Decolonising speech and language technology. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3504–3519, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Arianna Bisazza, Ahmet Üstün, and Stephan Sportel (2021). On the difficulty of translating free-order case-marking languages. *Transactions of the Association for Computational Linguistics*, 9:1233–1248.

Christopher M. Bishop (2006). *Pattern Recognition and Machine Learning*. Springer, New York.

Johannes Bjerva, Robert Östling, Maria Han Veiga, Jörg Tiedemann, and Isabelle Augenstein (2019). What do language representations really represent? *Computational Linguistics*, 45(2):381–389.

Barry J. Blake (2009). Classification of languages. In Keith Brown and Sarah Ogilvie, editors, *Concise Encyclopedia of Languages of the World*, pages 246–257. Elsevier, Oxford, UK.

Damian Blasi, Antonios Anastasopoulos, and Graham Neubig (2022). Systematic inequalities in language technology performance across the world's languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5486–5505, Dublin, Ireland. Association for Computational Linguistics.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Shujian Huang, Matthias Huck, Philipp Koehn, Qun Liu, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Raphael Rubino, Lucia Specia, and Marco Turchi (2017). Findings of the 2017 conference on machine translation (WMT17). In *Proceedings of the Second Conference on Machine Translation*, pages 169–214, Copenhagen, Denmark. Association for Computational Linguistics.

Keith Brown and Sarah Ogilvie, editors (2009). *Concise Encyclopedia of Languages of the World*. Elsevier, Oxford, UK.

Chris Callison-Burch, Miles Osborne, and Philipp Koehn (2006). Re-evaluating the role of Bleu in machine translation research. In *11th Conference of the European Chapter of the Association for Computational Linguistics*, pages 249–256, Trento, Italy. Association for Computational Linguistics.

Isaac Caswell, Ciprian Chelba, and David Grangier (2019). Tagged back-translation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, pages 53–63, Florence, Italy. Association for Computational Linguistics.

Kyunghyun Cho, Bart van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio (2014a). On the properties of neural machine translation: Encoder–decoder approaches. In *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*, pages 103–111, Doha, Qatar. Association for Computational Linguistics.

Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio (2014b). Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, Doha, Qatar. Association for Computational Linguistics.

Yu-Ying Chuang, Kaidi Lõo, James P. Blevins, and R. Harald Baayen (2020). *Estonian Case Inflection Made Simple: A Case Study in Word and Paradigm*

Morphology with Linear Discriminative Learning, page 119–141. Cambridge University Press.

Christopher Cieri, Mike Maxwell, Stephanie Strassel, and Jennifer Tracey (2016). Selection criteria for low resource language programs. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4543–4549, Portorož, Slovenia. European Language Resources Association (ELRA).

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning (2019). What does BERT look at? an analysis of BERT's attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 276–286, Florence, Italy. Association for Computational Linguistics.

Bernard Comrie (1989). *Language Universals and Linguistic Typology: Syntax and Morphology*. University of Chicago Press, Chicago.

Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang (2022). No language left behind: Scaling human-centered machine translation.

David Crystal (2010). *The Cambridge Encyclopedia of Language*, 3rd edition. Cambridge University Press, Cambridge.

Raj Dabre, Chenhui Chu, and Anoop Kunchukuttan (2020). A survey of multilingual neural machine translation. *ACM Comput. Surv.*, 53(5).

Raj Dabre, Tetsuji Nakagawa, and Hideto Kazawa (2017). An empirical study of language relatedness for transfer learning in neural machine translation. In *Proceedings of the 31st Pacific Asia Conference on Language, Information and Computation*, pages 282–286. The National University (Phillippines).

Michael Denkowski and Alon Lavie (2014). Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 376–380, Baltimore, Maryland, USA. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *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)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Matthew S. Dryer (2013). Order of subject, object and verb (v2020.4). In Matthew S. Dryer and Martin Haspelmath, editors, *The World Atlas of Language Structures Online*. Zenodo.

Matthew S. Dryer and Martin Haspelmath, editors (2013). *The World Atlas of Language Structures Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

David M. Eberhard, Gary F. Simons, and Charles D. Fennig, editors (2024). *Ethnologue: Languages of the World*, twenty-seventh edition. SIL International, Dallas, Texas. Online version: `http://www.ethnologue.com`.

Sergey Edunov, Myle Ott, Michael Auli, and David Grangier (2018). Understanding back-translation at scale. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 489–500, Brussels, Belgium. Association for Computational Linguistics.

Steffen Eger, Armin Hoenen, and Alexander Mehler (2016). Language classification from bilingual word embedding graphs. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3507–3518, Osaka, Japan. The COLING 2016 Organizing Committee.

Jeffrey L. Elman (1990). Finding structure in time. *Cognitive Science*, 14(2):179–211.

Mati Erelt, editor (2007). *Estonian Language*, 2nd edition, volume 1 of *Linguistica Uralica Supplementary Series*. Estonian Academy Publishers, Tallinn.

Marzieh Fadaee, Arianna Bisazza, and Christof Monz (2017). Data augmentation for low-resource neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 567–573, Vancouver, Canada. Association for Computational Linguistics.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov,

Edouard Grave, Michael Auli, and Armand Joulin (2021). Beyond english-centric multilingual machine translation. *J. Mach. Learn. Res.*, 22(1).

Markus Freitag, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey (2021). Experts, errors, and context: A large-scale study of human evaluation for machine translation. *Transactions of the Association for Computational Linguistics*, 9:1460–1474.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville (2016). *Deep Learning*. MIT Press. `http://www.deeplearningbook.org`.

Jie Gui, Tuo Chen, Jing Zhang, Qiong Cao, Zhenan Sun, Hao Luo, and Dacheng Tao (2024). A survey on self-supervised learning: Algorithms, applications, and future trends. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12):9052–9071.

Daniil Gurgurov, Michal Gregor, Josef van Genabith, and Simon Ostermann (2025). On multilingual encoder language model compression for low-resource languages.

Barry Haddow, Rachel Bawden, Antonio Valerio Miceli Barone, Jindřich Helcl, and Alexandra Birch (2022). Survey of low-resource machine translation. *Computational Linguistics*, 48(3):673–732.

Robert Thomas Harms (2006). Estonian. In Keith Brown and Sarah Ogilvie, editors, *Concise Encyclopedia of Languages of the World*, pages 377–378. Elsevier, Oxford.

Michael A. Hedderich, Lukas Lange, Heike Adel, Jannik Strötgen, and Dietrich Klakow (2021). A survey on recent approaches for natural language processing in low-resource scenarios. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2545–2568, Online. Association for Computational Linguistics.

Bernd Heine and Tania Kuteva (2005). Language contact and grammatical change. *Language Contact and Grammatical Change*, pages 1–308.

Daniel Hershcovich, Stella Frank, Heather Lent, Miryam de Lhoneux, Mostafa Abdou, Stephanie Brandl, Emanuele Bugliarello, Laura Cabello Piqueras, Ilias Chalkidis, Ruixiang Cui, Constanza Fierro, Katerina Margatina, Phillip Rust, and Anders Søgaard (2022). Challenges and strategies in cross-cultural NLP. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6997–7013, Dublin, Ireland. Association for Computational Linguistics.

Vu Cong Duy Hoang, Philipp Koehn, Gholamreza Haffari, and Trevor Cohn (2018). Iterative back-translation for neural machine translation. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pages 18–24, Melbourne, Australia. Association for Computational Linguistics.

Sepp Hochreiter and Jürgen Schmidhuber (1997a). Long short-term memory. *Neural Computation*, 9(8):1735–1780.

Sepp Hochreiter and Jürgen Schmidhuber (1997b). Long short-term memory. *Neural Computation*, 9(8):1735–1780.

Lars Johanson (2006). Altaic languages. In Keith Brown and Sarah Ogilvie, editors, *Concise Encyclopedia of Languages of the World*, pages 30–32. Elsevier, Oxford, UK.

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean (2017). Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

Andrew Joscelyne (1998). Altavista translates in real time. *Language International*, 10(1):6–7.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury (2020). The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Rafal Jozefowicz, Oriol Vinyals, Mike Schuster, Noam Shazeer, and Yonghui Wu (2016). Exploring the limits of language modeling.

Daniel Jurafsky and James H. Martin (2025). Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition with language models. Online manuscript released January 12, 2025.

Nal Kalchbrenner and Phil Blunsom (2013). Recurrent continuous translation models. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1700–1709, Seattle, Washington, USA. Association for Computational Linguistics.

Finnish F. Karlsson (2009). Finnish. In Keith Brown and Sarah Ogilvie, editors, *Concise Encyclopedia of Languages of the World*, pages 413–415. Elsevier, Oxford, UK.

Helin Kask (2016). English-estonian code-copying in estonian blogs. *Philologia Estonica Tallinnensis*, 1:80–101.

Dorothy Kenny, editor (2022). *Machine translation for everyone*. Number 18 in Translation and Multilingual Natural Language Processing. Language Science Press, Berlin.

Yoon Kim (2014). Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751, Doha, Qatar. Association for Computational Linguistics.

Katalin É. Kiss (2009). Hungarian. In Keith Brown and Sarah Ogilvie, editors, *Concise Encyclopedia of Languages of the World*, pages 514–516. Elsevier, Oxford, UK.

Tom Kocmi and Ondřej Bojar (2018). Trivial transfer learning for low-resource neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 244–252, Brussels, Belgium. Association for Computational Linguistics.

Philipp Koehn (2005). Europarl: A parallel corpus for statistical machine translation. In *Proceedings of Machine Translation Summit X: Papers*, pages 79–86, Phuket, Thailand.

Philipp Koehn (2020). *Neural Machine Translation*. Cambridge University Press.

Philipp Koehn and Rebecca Knowles (2017). Six challenges for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 28–39, Vancouver. Association for Computational Linguistics.

Elizaveta Korotkova and Mark Fishel (2024). Estonian-centric machine translation: Data, models, and challenges. In *Proceedings of the 25th Annual Conference of the European Association for Machine Translation (Volume 1)*, pages 647–660, Sheffield, UK. European Association for Machine Translation (EAMT).

Elizaveta Korotkova, Taido Purason, Agnes Luhtaru, and Mark Fishel (2024). Multilinguality or back-translation? a case study with Estonian. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 11838–11848, Torino, Italia. ELRA and ICCL.

Julia Kreutzer, Isaac Caswell, Lisa Wang, Ahsan Wahab, Daan van Esch, Nasanbayar Ulzii-Orshikh, Allahsera Tapo, Nishant Subramani, Artem Sokolov, Claytone

Sikasote, Monang Setyawan, Supheakmungkol Sarin, Sokhar Samb, Benoît Sagot, Clara Rivera, Annette Rios, Isabel Papadimitriou, Salomey Osei, Pedro Ortiz Suarez, Iroro Orife, Kelechi Ogueji, Andre Niyongabo Rubungo, Toan Q. Nguyen, Mathias Müller, André Müller, Shamsuddeen Hassan Muhammad, Nanda Muhammad, Ayanda Mnyakeni, Jamshidbek Mirzakhalov, Tapiwanashe Matangira, Colin Leong, Nze Lawson, Sneha Kudugunta, Yacine Jernite, Mathias Jenny, Orhan Firat, Bonaventure F. P. Dossou, Sakhile Dlamini, Nisansa de Silva, Sakine Çabuk Ballı, Stella Biderman, Alessia Battisti, Ahmed Baruwa, Ankur Bapna, Pallavi Baljekar, Israel Abebe Azime, Ayodele Awokoya, Duygu Ataman, Orevaoghene Ahia, Oghenefego Ahia, Sweta Agrawal, and Mofetoluwa Adeyemi (2022). Quality at a glance: An audit of web-crawled multilingual datasets. *Transactions of the Association for Computational Linguistics*, 10:50–72.

Johanna Laakso (2001). The finnic languages. In Östen Dahl and Maria Koptjevskaja-Tamm, editors, *Circum-Baltic Languages. Volume 1: Past and Present*, volume 54, pages clxxix–ccxii. John Benjamins Publishing Company, Amsterdam.

Surafel Melaku Lakew, Aliia Erofeeva, and Marcello Federico (2018). Neural machine translation into language varieties. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 156–164, Brussels, Belgium. Association for Computational Linguistics.

Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324.

Zewen Li, Fan Liu, Wenjie Yang, Shouheng Peng, and Jun Zhou (2022). A survey of convolutional neural networks: Analysis, applications, and prospects. *IEEE Transactions on Neural Networks and Learning Systems*, 33(12):6999–7019.

Jindřich Libovický, Rudolf Rosa, and Alexander Fraser (2019). How language-neutral is multilingual bert?

Chin-Yew Lin (2004). ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Liyuan Liu, Xiaodong Liu, Jianfeng Gao, Weizhu Chen, and Jiawei Han (2020a). Understanding the difficulty of training transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5747–5763, Online. Association for Computational Linguistics.

Xiao Liu, Fanjin Zhang, Zhenyu Hou, Li Mian, Zhaoyu Wang, Jing Zhang, and Jie Tang (2023). Self-supervised learning: Generative or contrastive. *IEEE Transactions on Knowledge and Data Engineering*, 35(1):857–876.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer (2020b). Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Arle Richard Lommel, Aljoscha Burchardt, and Hans Uszkoreit (2013). Multidimensional quality metrics: a flexible system for assessing translation quality. In *Proceedings of Translating and the Computer 35*, London, UK. Aslib.

Andrea Lösch, Valérie Mapelli, Stelios Piperidis, Andrejs Vasiļjevs, Lilli Smal, Thierry Declerck, Eileen Schnur, Khalid Choukri, and Josef van Genabith (2018). European language resource coordination: Collecting language resources for public sector multilingual information management. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Angela Marcantonio (2006). Uralic languages. In Keith Brown and Sarah Ogilvie, editors, *Concise Encyclopedia of Languages of the World*, pages 1129–1133. Elsevier, Oxford, UK.

Ibrahim Merad, Amos Wolf, Ziad Mazzawi, and Yannick Léo (2025). Language verY rare for all. In *Proceedings of the First Workshop on Language Models for Low-Resource Languages*, pages 166–174, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Tom M Mitchell (1997). *Machine learning*, volume 1. McGraw-hill New York.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves, Martin Riedmiller, Andreas K. Fidjeland, Georg Ostrovski, Stig Petersen, Charles Beattie, Amir Sadik, Ioannis Antonoglou, Helen King, Dharshan Kumaran, Daan Wierstra, Shane Legg, and Demis Hassabis (2015). Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533.

Mehmed Mohammed, Mohammad B. Khan, and Eihab B. M. Bashier (2016). *Machine Learning: Algorithms and Applications*, 1 edition. CRC Press.

Kadri Muischnek (2023). *Language Report Estonian*, pages 131–134. Springer International Publishing.

Kevin P. Murphy (2022). *Probabilistic Machine Learning: An introduction*. MIT Press.

Rudra Murthy, Anoop Kunchukuttan, and Pushpak Bhattacharyya (2019). Addressing word-order divergence in multilingual neural machine translation for extremely low resource languages. In *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)*, pages 3868–3873, Minneapolis, Minnesota. Association for Computational Linguistics.

Wilhelmina Nekoto, Vukosi Marivate, Tshinondiwa Matsila, Timi Fasubaa, Taiwo Fagbohungbe, Solomon Oluwole Akinola, Shamsuddeen Muhammad, Salomon Kabongo Kabenamualu, Salomey Osei, Freshia Sackey, Rubungo Andre Niyongabo, Ricky Macharm, Perez Ogayo, Orevaoghene Ahia, Musie Meressa Berhe, Mofetoluwa Adeyemi, Masabata Mokgesi-Selinga, Lawrence Okegbemi, Laura Martinus, Kolawole Tajudeen, Kevin Degila, Kelechi Ogueji, Kathleen Siminyu, Julia Kreutzer, Jason Webster, Jamiil Toure Ali, Jade Abbott, Iroro Orife, Ignatius Ezeani, Idris Abdulkadir Dangana, Herman Kamper, Hady Elsahar, Goodness Duru, Ghollah Kioko, Murhabazi Espoir, Elan van Biljon, Daniel Whitenack, Christopher Onyefuluchi, Chris Chinenye Emezue, Bonaventure F. P. Dossou, Blessing Sibanda, Blessing Bassey, Ayodele Olabiyi, Arshath Ramkilowan, Alp Öktem, Adewale Akinfaderin, and Abdallah Bashir (2020). Participatory research for low-resourced machine translation: A case study in African languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2144–2160, Online. Association for Computational Linguistics.

Graham Neubig and Junjie Hu (2018). Rapid adaptation of neural machine translation to new languages. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 875–880, Brussels, Belgium. Association for Computational Linguistics.

Toan Q. Nguyen and David Chiang (2017). Transfer learning across low-resource, related languages for neural machine translation. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 296–301, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Hellina Hailu Nigatu, Atnafu Lambebo Tonja, Benjamin Rosman, Thamar Solorio, and Monojit Choudhury (2024). The zeno's paradox of 'low-resource' languages. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17753–17774, Miami, Florida, USA. Association for Computational Linguistics.

Antoine Nzeyimana (2024). Low-resource neural machine translation with morphological modeling. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 182–195, Mexico City, Mexico. Association for Computational Linguistics.

Helen O'Horan, Yevgeni Berzak, Ivan Vulić, Roi Reichart, and Anna Korhonen (2016). Survey on the use of typological information in natural language processing. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1297–1308, Osaka, Japan. The COLING 2016 Organizing Committee.

Myle Ott, Sergey Edunov, David Grangier, and Michael Auli (2018). Scaling neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 1–9, Brussels, Belgium. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu (2002). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Peyman Passban, Andy Way, and Qun Liu (2018). Tailoring neural architectures for translating from morphologically rich languages. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3134–3145, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Stelios Piperidis, Penny Labropoulou, Miltos Deligiannis, and Maria Giagkou (2018). Managing public sector data for multilingual applications development. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Edoardo Maria Ponti, Helen O'Horan, Yevgeni Berzak, Ivan Vulić, Roi Reichart, Thierry Poibeau, Ekaterina Shutova, and Anna Korhonen (2019). Modeling language variation and universals: A survey on typological linguistics for natural language processing. *Computational Linguistics*, 45(3):559–601.

Maja Popović (2015). chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Matt Post (2018). A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Taido Purason, Aleksei Ivanov, Lisa Yankovskaya, and Mark Fishel (2024). SMUGRI-MT - machine translation system for low-resource Finno-Ugric languages. In *Proceedings of the 25th Annual Conference of the European Association for Machine Translation (Volume 2)*, pages 31–32, Sheffield, UK. European Association for Machine Translation (EAMT).

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever (2018). Improving language understanding by generative pre-training. `https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf`.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu (2023). Exploring the limits of transfer learning with a unified text-to-text transformer.

Taraka Rama and Lars Borin (2015). Comparative evaluation of string similarity measures for automatic language classification. In Jan Mačutek and George K. Mikros, editors, *Sequences in Language and Text*, pages 203–231. De Gruyter Mouton.

Surangika Ranathunga, En-Shiun Annie Lee, Marjana Prifti Skenduli, Ravi Shekhar, Mehreen Alam, and Rishemjit Kaur (2023). Neural machine translation for low-resource languages: A survey. *ACM Comput. Surv.*, 55(11).

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie (2020). COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Philip Resnik and Noah A. Smith (2003). The web as a parallel corpus. *American Journal of Computational Linguistics*, 29(3):349–380.

Matīss Rikters, Mārcis Pinnis, and Rihards Krišlauks (2018). Training and adapting multilingual NMT for less-resourced and morphologically rich languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Matīss Rikters, Marili Tomingas, Tuuli Tuisk, Valts Ernštreits, and Mark Fishel (2022). Machine translation for Livonian: Catering to 20 speakers. In *Proceedings*

of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 508–514, Dublin, Ireland. Association for Computational Linguistics.

Dana Ruiter, Cristina España-Bonet, and Josef van Genabith (2019). Self-supervised neural machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1828–1834, Florence, Italy. Association for Computational Linguistics.

Dana Ruiter, Josef van Genabith, and Cristina España-Bonet (2020). Self-induced curriculum learning in self-supervised neural machine translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2560–2571, Online. Association for Computational Linguistics.

Louisa Sadler (1992). Rule-based translation as constraint resolution. In *International Workshop on Fundamental Research for the Future Generation of Natural Language Processing (FGNLP)*, Manchester, UK.

Iqbal H. Sarker (2021). Machine learning: Algorithms, real-world applications and research directions. *SN Computer Science*, 2(3):160.

Holger Schwenk, Guillaume Wenzek, Sergey Edunov, Edouard Grave, Armand Joulin, and Angela Fan (2021). CCMatrix: Mining billions of high-quality parallel sentences on the web. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6490–6500, Online. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch (2016). Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.

Rico Sennrich and Biao Zhang (2019). Revisiting low-resource neural machine translation: A case study. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 211–221, Florence, Italy. Association for Computational Linguistics.

Gary F. Simons, Abbey L. L. Thomas, and Chad K. K. White (2022). Assessing digital language support on a global scale. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4299–4305, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Ralf Steinberger, Bruno Pouliquen, Anna Widiger, Camelia Ignat, Tomaž Erjavec, Dan Tufiş, and Dániel Varga (2006). The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy. European Language Resources Association (ELRA).

Dario Stojanovski and Alexander Fraser (2018). Coreference and coherence in neural machine translation: A study using oracle experiments. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 49–60, Brussels, Belgium. Association for Computational Linguistics.

Thomas Stolz (1991). *Sprachbund im Baltikum? Estnisch und Lettisch im Zentrum einer sprachlichen Konvergenzlandschaft*. Universitätsverlag Brockmeyer.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le (2014). Sequence to sequence learning with neural networks. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'14, page 3104–3112, Cambridge, MA, USA. MIT Press.

Maali Tars, Taido Purason, and Andre Tättar (2022a). Teaching unseen low-resource languages to large translation models. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 375–380, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Maali Tars, Andre Tättar, and Mark Fišel (2021). Extremely low-resource machine translation for closely related languages. In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 41–52, Reykjavik, Iceland (Online). Linköping University Electronic Press, Sweden.

Maali Tars, Andre Tättar, and Mark Fishel (2022b). Cross-lingual transfer from large multilingual translation models to unseen under-resourced languages. *Baltic Journal of Modern Computing*, 10.

Yi Tay, Mostafa Dehghani, Dara Bahri, and Donald Metzler (2022). Efficient transformers: A survey. *ACM Comput. Surv.*, 55(6).

Jörg Tiedemann (2012). Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218, Istanbul, Turkey. European Language Resources Association (ELRA).

Lisa Torrey and Jude Shavlik (2009). Transfer learning. In Emilio Soria, Josefa Martin, Rafael Magdalena, Manuel Martinez, and Armando Serrano, editors,

Handbook of Research on Machine Learning Applications and Trends: Algorithms, Methods, and Techniques. IGI Global.

Andre Tättar, Taido Purason, Hele-Andra Kuulmets, Agnes Luhtaru, Liisa Ratsep, Maali Tars, Marcis Pinnis, Toms Bergmanis, and Mark Fishel (2022). Open and competitive multilingual neural machine translation in production. *Baltic Journal of Modern Computing*, 10(3):422–434.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin (2017). Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6000–6010, Red Hook, NY, USA. Curran Associates Inc.

Viveka Velupillai (2012). *An Introduction to Linguistic Typology*. Typological Studies in Language. John Benjamins B.V., Amsterdam, The Netherlands. Reprinted 2013 with corrections.

Anna Verschik (2005). The language situation in estonia. *Journal of Baltic Studies*, 36(3):283–316.

Yequan Wang, Minlie Huang, Xiaoyan Zhu, and Li Zhao (2016). Attention-based LSTM for aspect-level sentiment classification. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 606–615, Austin, Texas. Association for Computational Linguistics.

Ze Yang, Wei Wu, Jian Yang, Can Xu, and Zhoujun Li (2019). Low-resource response generation with template prior. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1886–1897, Hong Kong, China. Association for Computational Linguistics.

JiaJun Zhang and ChengQing Zong (2020). Neural machine translation: Challenges, progress and future. *Science China Technological Sciences*, 63(10):2028–2050.

Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight (2016). Transfer learning for low-resource neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1568–1575, Austin, Texas. Association for Computational Linguistics.