Paragraph-Level Machine Translation for Low-Resource Finno-Ugric Languages

Dmytro Pashchenko and **Lisa Yankovskaya** and **Mark Fishel**
Institute of Computer Science
University of Tartu, Estonia
{dmytro.pashchenko,lisa.yankovskaya,mark.fisel}@ut.ee

Abstract

We develop paragraph-level machine translation for four low-resource Finno-Ugric languages: Proper Karelian, Livvi, Ludian, and Veps. The approach is based on sentence-level pre-trained translation models, which are fine-tuned with paragraph-parallel data. This allows the resulting model to develop a native ability to handle discource-level phenomena correctly, in particular translating from grammatically gender-neutral input in Finno-Ugric languages. We collect monolingual and parallel paragraph-level corpora for these languages. Our experiments show that paragraph-level translation models can translate sentences no worse than sentence-level systems, while handling discourse-level phenomena better. For evaluation, we manually translate part of FLORES-200 into these four languages. All our results, data, and models are released openly.

1 Introduction

The existence of massively multilingual pre-trained translation models (e.g. m2m100, NLLB, and MADLAD-400: Fan et al., 2021; NLLB Team et al., 2022; Kudugunta et al., 2023) has made work on machine translation significantly easier by eliminating the need for training large models from zero. Nevertheless, even the largest of these models still leave many low-resource languages out—mainly due to lack of or difficulty to acquire textual data (monolingual or parallel) in those languages.

Moreover, these translation models approach translation by handling each sentence independently and thus do not handle discourse-level phenomena well[1]. Ignoring the discourse-level phenomena has been shown to pose problems for translation quality and its assessment (Bawden et al., 2018; Läubli et al., 2018). Even though decoder-only language models (e.g. GPT4, OpenAI et al., 2024) are an easy way to approach document-level translation, the availability of pre-trained open multilingual language models and their language coverage are even narrower than for translation models. Also, translation is more efficiently solved with sequence-to-sequence models when emergent abilities are not a requirement and the main purpose is to solve translation, not other tasks.

In this paper, we focus on developing machine translation for the Finno-Ugric family of languages, which is a good fit for addressing both aforementioned issues, namely support for low-resource languages and discourse-level phenomena ignorance:

- the majority of pre-trained models only support three languages from this family (Finnish, Estonian and Hungarian), with MADLAD-400 also including a few more, still leaving out dozens of languages, and
- Finno-Ugric languages have no grammatical category of gender and use gender-neutral pronouns. This increases their dependence on document-level context, see an example in Figure 1.

We narrow down our scope to four under-resourced members of the Finno-Ugric language family: Proper Karelian, Livvi, Ludian, and Veps. All four are low-resource languages and are not included in m2m100, NLLB, or MADLAD-400; they are also not supported by Google Translate[2] or DeepL[3], as of January 2025.

[1] Although MADLAD-400 (Kudugunta et al., 2023) is pre-trained on full documents, this is done without cross-linguality but rather via monolingually denoising documents in several languages; translation is later taught to the model on sentence level.

[2] https://translate.google.com

[3] https://deepl.com

Proceedings of the Joint 25th Nordic Conference on Computational Linguistics and 11th Baltic Conference on Human Language Technologies (NoDaLiDa/Baltic-HLT 2025), pages 458–469
March 3-4, 2025

Text in Veps:	Naine tuli kodihe. **Hänen** mašin jäi garažas.
English translation:	The woman came home. **Her** car remained in the garage.

Figure 1: Example of translation challenges related to gender-neutral pronouns in Finno-Ugric languages: the Veps text includes the pronoun hänen, which can be translated both as "her" and "his"; resolving this ambiguity requires looking at the first sentence and the word naine (woman) as the antecedent.

With the issues listed above in mind, we collect paragraph-level corpora and develop paragraph-level machine translation models by simply fine-tuning sequence-to-sequence models on parallel paragraph pairs, comparing the results to sentence-level approach. In order to fit the paragraph into the context window of the model, we limit its length to five sentences at most—our experiments show that such a bounded context still allows the model to learn extrasentential dependencies.

Our key contributions are thus the following:

- We collect and release paragraph-level corpora for Proper Karelian, Livvi, Ludian, and Veps: monolingual, as well as parallel with Russian (Section 4).
- In order to evaluate the results, we extend part of the translation benchmark FLORES-200 by manually translating it into the new languages, as well as manually correct existing Russian translations for paragraph-level consistency (Section 4).
- We train both sentence-level and paragraph-level translation systems on the collected data and show that the latter has the same or better quality when applied to paragraphs as well as learns to translate discourse-level phenomena correctly (Sections 5 and 6).

The collected data[4], trained models[5], and created benchmarks[6] are released openly.

Next, we outline the related work in Section 2 and present the methodology in Section 3.

[4] https://huggingface.co/datasets/tartuNLP/pale-madlad-data

[5] https://huggingface.co/tartuNLP/pale-madlad-mt

[6] https://huggingface.co/datasets/tartuNLP/smugri-flores-testset

2 Related Work

Document-level translation Elaborating on the importance of considering the extrasentential context in machine translation (MT), Bawden et al. (2018) describe major discourse-level phenomena that present problems for most MT systems: coreference, lexical cohesion, and lexical disambiguation. Taking into account the context beyond a single sentence is essential for correct translation. Throughout the history of MT, researchers tried to address this problem from different perspectives—from rule-based to statistical to corpus-based approaches—creating various document-level systems (Hardmeier, 2012; Hardmeier et al., 2013).

Currently, attempts have been made to incorporate context in the attention-based models' scope by modifying their architecture. The researchers offered methods such as hierarchical attention (Miculicich et al., 2018) or memory networks (Maruf and Haffari, 2018) among others. However, the most straightforward strategies, like passing an entire text to the model, proved also the most effective. Sun et al. (2022) trained the Transformer model (Vaswani et al., 2017) on documents, repeatedly dividing them into parts to vary input lengths. Although this approach has shown a big leap in translation quality, it does not remedy another important problem: long processing times of large documents. The time and memory consumption of Transformer-based systems scales quadratically with the input length. We try to avoid this issue by splitting documents into small, fixed-size paragraphs rather than translating documents fully.

MT for low-resource Finno-Ugric languages Machine translation for low-resource Finno-Ugric languages has been explored in a number of works. To name but a few, Tyers et al. (2009) examined rule-based and statistical MT systems when translating between North and Lule Sámi; Pirinen et al. (2017) employed rule-based MT in their North Sámi-Finnish system; Rikters et al. (2022) designed a neural MT system for Livonian. The languages studied in this work were presented in MT systems developed by Yankovskaya et al. (2023) and Purason et al. (2024), but unlike our approach, their systems do not take the document or paragraph context into account.

3 Methodology

In this chapter, we briefly describe our approach to dealing with paragraph-level data, ways to extract paragraphs from documents and evaluate paragraph-level translations. We chose MADLAD-400 as the basis for our experiments, since, in addition to being a small, powerful, and open-source model, it has the potential for paragraph-level translation as it was pre-trained with document-level monolingual data.

3.1 Splitting Documents into Paragraphs

With our primary task being to test whether including the extrasentential context improves the performance of MADLAD, we need to decide on how many sentences to use as the model's input. On the one hand, the more sentences we take from a document, the more likely the model is to capture the necessary context for translating each sentence. On the other hand, passing the document as a whole as the model's input may be impractical for two reasons:

- **Time and memory consumption.** The attention mechanism inside Transformers has quadratic computational complexity $\mathbf{O}(n^2)$, since the attention is calculated between each pair of tokens. Therefore, computation time and memory consumption increase quadratically with the input size. Shorter input sequences would ensure much faster model training.

- **Overfitting by length.** Varis and Bojar (2021) show that Transformers generalize badly to out-of-distribution input lengths. This means that loosening the restrictions on input length would require more training with diverse data (short and long) to avoid underfitting some lengths and overfitting the others. The stricter the restrictions—the easier the training.

We overcome the two aforementioned issues at once by splitting documents into smaller paragraphs of fixed, reasonable length. Since MADLAD was trained on sequences whose length did not exceed 256 tokens, we set a similar length limit. We abandoned the idea of forming paragraphs from as many sentences as possible to get close to the size limit, for this would have led to a low variance of data lengths. Instead, we combine a fixed number of sentences. If the paragraph length exceeds 256 tokens, we split the paragraph in two; if the paragraph is still too long but consists of a single sentence, we trim the paragraph to the maximum length.

Through experimentation, we have found that, on average, five sentences are enough to fit into the context window of 256 tokens on our training data without resorting to unnecessary splitting or truncation of paragraphs. Where the number of sentences is not divisible by 5, we take the remainder as a separate paragraph. We emphasize that there is no optimal choice of paragraph length and it should instead be chosen empirically or based on the model's context length and the available data.

3.2 Evaluating Paragraph-Level Translations

The most popular surface-level metrics, BLEU (Papineni et al., 2002) and chrF++ (Popović, 2017), were designed to evaluate sentences. Applying them to paragraphs could compromise correlation with human judgments. Deutsch et al. (2023) have proved the opposite: BLEU scores for paragraphs not only align with those of humans but also become more accurate as paragraph size increases. This finding allows us to adopt BLEU as a paragraph-level metric without the need to train custom scoring models, which is problematic due to the resource-constrained setting.

We also use chrF++ as it is more suited for morphologically rich languages, such as the ones from the Finno-Ugric family. Drawing on the formal similarity and correlation of the BLEU and chrF++ metrics, we apply the latter directly to paragraphs as well.

3.3 Managing Language Tokens

MADLAD-400 requires a language token to be manually prepended to the user's input sequence. These tokens take the form `<2xx>`, where `xx` stands for a target language code. For instance, the sequence "`<2en> Mitä kuuluu?`" indicates that the Finnish sentence "`Mitä kuuluu?`" needs to be translated into English. Thus, we prepend four language indicators to the input sequences: `<2krl>` for Proper Karelian, `<2lud>` for Ludian, `<2olo>` for Livvi, and `<2vep>` for Veps. The codes are taken from the ISO 639-3[7] code set. As for the Russian language, MADLAD encodes it as `<2ru>`.

However, in this work, we do not expand MADLAD's vocabulary with new language tokens. In-

[7] https://iso639-3.sil.org/code_tables/639/data

stead, to save effort and time, we do nothing and expect the model to learn the tokens solely based on their textual representations, which we prepend to inputs. Pilot experiments showed us that this approach is as effective as specifying language tokens explicitly.

4 Data

In this paper, we focus on (i) two dialects of Karelian: Livvi (olo) and Proper Karelian (krl)[8]; (ii) Ludian (lud), which is closely related to Karelian, but is considered a language in its own right (Pahomov, 2017); and (iii) Veps (vep). These are all endangered Finnic languages, mainly spoken in Finland and Russia.

4.1 Data sources for training

The majority of the training data was parsed from the following resources: two media portals omamedia.ru[9] and yle.fi[10], open corpus of Veps and Karelian languages VepKar (Boyko et al., 2022), and Wikipedia. We were unable to utilize other published datasets, as they primarily comprised sentences rather than documents (although sentence-level data can still help improve the overall translation quality).

A preliminary analysis of translations revealed that the MT system was mixing Livvi and Proper Karelian. A possible reason for this mixing could be the incorrect assignment of language labels to the source data. After studying the sources and consulting linguists, we discovered that the texts from the media portal "Omamedia" were not only written in Livvi, as we previously thought, but also in other varieties of Karelian language, mainly Proper Karelian. Using the language identification tool `GlotLID` (Kargaran et al., 2023), we redistributed the texts according to the new language labels.

We did minor preprocessing steps aimed at normalizing characters and removing redundant elements (e.g., useless Wikipedia sections) to extract coherent texts from the sources.

Table 1 presents the composition of the final dataset.

[8]Proper Karelian comprises Northern (Viena) Karelian and Southern Karelian. In this study, we use both varieties to train our MT system, but we test the output only in Northern (Viena) Karelian

[9]`https://omamedia.ru/en`

[10]`https://yle.fi/t/18-44136/fi`

4.2 Benchmark dataset

The benchmark dataset of low-resource Finno-Ugric languages published by Yankovskaya et al. (2023) contains Livvi, our language of interest. We extended this dataset by adding three more languages: Proper Karelian (Viena), Ludian[11], and Veps. Like Yankovskaya et al. (2023), we translated the first 250 rows of the FLORES dataset (NLLB Team et al., 2022); the translations from Russian were done by native speakers of these languages who have extensive translation experience.

Another important step was to modify the existing FLORES-200 test set, transforming it from a sentence-level set into a paragraph-level one. Fortunately, the FLORES-200 benchmark (NLLB Team et al., 2022) is a collection of short excerpts from Wikipedia, where sentences are sequential. All we had to do was isolate these paragraphs. When their length exceeded the maximum allowable, we manually divided them into smaller paragraphs in such a way as to avoid incurring a significant loss of context. Thus, the original 250 rows transformed into 87 paragraphs. However, when verifying the consistency of paragraphs, we noticed that the sentences in the data set were probably translated separately, out of context. Therefore, we manually edited the paragraphs, ensuring the correct and consistent use of pronouns, names, terms, etc. in the Russian segment of FLORES.

We shall refer to these benchmarks sets as "Smugri FLORES benchmark."

5 Experimental Setup

To investigate the effect of paragraphs on the quality of translation of Proper Karelian, Livvi, Ludian, and Veps, we fine-tune two MADLAD models: one on sentence-level data and the other on paragraph-level data. We translate the languages into Russian and vice versa. Russian was chosen as a translation objective (among other high-resource languages available in MADLAD) because most of the openly available parallel texts were aligned with the Russian language.

To further improve the model, we perform back-translation making use of our monolingual data. We back-translate in a single direction—from Finno-Ugric languages to Russian—and thus, enhance the quality of translation from Russian to Finno-Ugric languages (otherwise quite low). We

[11]using the alphabet with ü instead of y

	krl		lud		olo		vep	
data source	mono	para	mono	para	mono	para	mono	para
vepkar-sent	45.4	32.3	5.9	7.9	36.0	22.2	38.3	20.4
vepkar-par	9.6	6.9	1.2	1.6	7.6	4.8	8.1	4.5
wikipedia-sent	-	-	-	-	28.4	-	99.8	-
wikipedia-par	-	-	-	-	7.7	-	24.0	-
omamedia-sent	8.3	-	-	-	3.5	-	6.5	-
omamedia-par	2.0	-	-	-	0.8	-	1.6	-
ylefi-sent	-	-	-	-	14.2	-	-	-
ylefi-par	-	-	-	-	3.2	-	-	-
total-sent	53.7	32.3	5.9	7.9	82.2	22.2	144.6	20.4
total-par	11.7	6.9	1.2	1.6	19.4	4.8	33.7	4.5

Table 1: The distribution of sentence-level (sent) and paragraph-level (par) parallel data (para) and monolingual data (mono) by language in the final dataset. Quantities are given in thousands, rounded to the nearest tenth.

avoid back-translation between the four selected languages because low-quality synthetic data can harm the resulting performance instead of improving it (Yankovskaya et al., 2023).

Using the HuggingFace framework[12], we fine-tune both the sentence-level and paragraph-level model for 10 epochs under equal conditions. We set the hyperparameters of `Seq2SeqTrainingArguments` to their default values with the following exceptions:

- We limit the generation length to 256 tokens.
- Following the MADLAD-400 paper, we set up an inverse square root scheduler with 300 warmup steps.
- We distribute fine-tuning across 8 GPUs. To approximately equalize the number of optimization steps for both models, we adjust the batch size depending on the total amount of data: 8 examples for paragraph-level data and 32 examples for sentence-level data.

We perform fine-tuning on the LUMI[13] supercomputer with AMD Instinct MI250X GPUs.

We use both models to translate paragraphs from the modified Smugri FLORES benchmark. For generation, we set the standard beam size of 5. We evaluate translations with the BLEU and chrF++ metrics, of which we use the SacreBLEU (Post, 2018) implementations. When calculating chrF++, we count only word bigrams. To measure statistical significance and confidence intervals, we do bootstrap resampling with 1000 resamples.

6 Results

In this section, we examine the obtained results, starting with a quantitative analysis that presents translations from Proper Karelian, Livvi, Ludian, and Veps into Russian, as well as from Russian to these four languages. Next, we conduct a brief qualitative analysis. After this, we compare our results with those generated by the online machine translation engine Tartu NLP Neurotõlge[14]. Finally, we explore how well translation abilities transfer to the unseen case of English translation.

6.1 Quantitative analysis

We begin our analysis by comparing the translation quality of two MADLAD models—one trained with sentences (SL model) and the other trained with paragraphs (PL model) —as measured by the automatic metrics of BLEU and chrF++ (see Section 3.2). To translate paragraphs with the sentence-level system, we process them sentence by sentence and then merge back into a paragraph. Otherwise, when given a full paragraph, the SL system tends to translate it into a single complex sentence with multiple subordinate clauses, thus decreasing the scores.

The results are presented in Table 2, in which we also provide the scores of the base MADLAD

[12] https://huggingface.co/

[13] https://lumi-supercomputer.eu/

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

	base	SL	PL	p-value
krl-ru	$14.6 \pm 1.7 / 40.2 \pm 2.1$	$21.1 \pm 1.9 / 48.7 \pm 1.6$	$21.9 \pm 2.0 / 49.5 \pm 1.6$	0.060 / **0.036**
lud-ru	$8.8 \pm 1.6 / 31.1 \pm 2.2$	$18.0 \pm 1.8 / 45.2 \pm 1.6$	$19.5 \pm 2.0 / 46.1 \pm 1.6$	**0.004 / 0.034**
olo-ru	$9.5 \pm 1.6 / 31.9 \pm 2.3$	$22.0 \pm 2.0 / 48.2 \pm 1.7$	$22.4 \pm 2.2 / 48.9 \pm 1.7$	0.217 / 0.076
vep-ru	$8.6 \pm 1.5 / 30.8 \pm 2.1$	$21.1 \pm 1.8 / 46.5 \pm 1.7$	$21.1 \pm 1.8 / 47.0 \pm 1.7$	0.392 / 0.090
ru-krl	$0.4 \pm 0.2 / 3.0 \pm 0.8$	$13.5 \pm 1.5 / 46.8 \pm 1.3$	$13.3 \pm 1.6 / 46.9 \pm 1.2$	0.221 / 0.385
ru-lud	$0.3 \pm 0.1 / 2.5 \pm 0.5$	$4.3 \pm 1.1 / 34.1 \pm 1.1$	$3.9 \pm 1.1 / 33.8 \pm 1.0$	0.143 / 0.127
ru-olo	$0.6 \pm 0.4 / 2.9 \pm 0.7$	$8.7 \pm 1.4 / 40.7 \pm 1.2$	$8.5 \pm 1.4 / 40.2 \pm 1.7$	0.193 / 0.185
ru-vep	$0.3 \pm 0.1 / 3.0 \pm 0.7$	$12.0 \pm 1.4 / 43.1 \pm 1.5$	$12.1 \pm 1.6 / 42.4 \pm 1.9$	0.409 / 0.138

Table 2: Translation metrics for translation directions from/into Russian, BLEU and chrF++ scores (separated by slash) of base MADLAD, sentence-level (SL) MADLAD, and paragraph-level (PL) MADLAD evaluated on the paragraph-level Smugri FLORES benchmark. p-value is the probability that SL and PL models are the same with respect to each metric; p-value less than 0.05 indicates that the difference between the models is statistically significant (highlighted in bold).

model. In the last column, we provide a p-value for each translation direction. Our null hypothesis is that the two models, sentence- and paragraph-level, are the same model. In cases where the p-value is less than 0.05, we reject the hypothesis and conclude that the difference between the models is statistically significant, with one clearly outperforming the other.

First, we observe that the base MADLAD-400 model, with no fine-tuning, is able to translate Proper Karelian, Livvi, Ludian, and Veps into Russian with good initial quality. The Proper Karelian→Russian translation score goes as high as 14.6 BLEU or 40.2 chrF++. This probably indicates that the model's knowledge of related languages (Finnish, Estonian, Russian) was successfully transferred to this case.

After fine-tuning, the results improved considerably. The paragraph-level (PL) model is significantly better than the sentence-level (SL) one in the case of Ludian→Russian translation. The difference is notable, reaching 1.5 BLEU points and 0.9 chrF++ points. The chrF++ scores further confirm the superiority of the PL model in the Proper Karelian→Russian direction. At the same time, the BLEU metric shows no significant difference. Finally, in all other cases, both metrics indicate that the SL and PL models, on average, perform equally well.

Thus, the paragraph-level model is no worse and, at times, strongly better than the sentence-level model. The difference is the most pronounced in the case of translation *into* Russian, giving us reason to believe that the PL model successfully resolves some discourse-level phenomena inherent in Finno-Ugric languages, such as gender-neutral pronouns. These phenomena occur rarely (yet they are important for high-quality coherent translation), and automated metrics do not necessarily reflect the extent to which they have been handled. To further investigate the issue, we qualitatively analyze translated texts.

6.2 Qualitative analysis

Next we present the results of manual qualitative analysis of paragraphs translations from the FLORES-200 benchmark. Although the number of discourse-level phenomena in the test set is quite limited, we managed to discover cases where (i) lexical cohesion must be preserved to translate terminology and proper nouns and (ii) where pronouns in different sentences must be aligned via coreference resolution. A detailed descriptions of errors presented in Table 3 and a summary is presented below.

The first part of the qualitative analysis addresses lexical cohesion, which refers to the consistent translation of terminology. The PL model translates terminology and names more consistently than the SL model across all languages and directions (*from* Russian and *to* Russian). While the PL model occasionally produces incorrect translations of names and terms, it typically does so consistently. In contrast, the SL model is inconsistent, translating a term or name correctly in one sentence but incorrectly in another, or generating incorrect translations with slight variations ("Simonioff" and "Simoninov").

→	SL	PL	Comments
krl-ru	Ранее, генеральный директор Ring, Джейми Симинофф, отметил, что компания получила своё начало от того, что он не услышал, как в его гараже зазвонил звонок из магазина. Он рассказал, что сделал Wi-Fi-звонок. Симинофф рассказала, что продажи выросли после того, как она появилась в 2013 году в шоу «Shark Tank», где судьи отказались финансировать её выступление. В конце 2017 года Симинов появился на покупательном канале QVC. Кроме того, Ring заключил соглашение с конкурирующей компанией по обеспечению безопасности ADT Corporation.	Ранее генеральный директор компании Ring Джейми Симинофф отметил, что компания получила своё начало, когда он не услышал звонок в дверь из магазина в своем гараже. Он рассказал, что сделал Wi-Fi дверной звонок. Симинофф сказал, что признание выросло после того, как он появился в 2013 году в шоу Shark Tank, где судьи отказались финансировать его инициативу. В конце 2017 года Симинофф появился на торговом канале QVC. Кроме того, Ring достиг соглашения с конкурирующей компанией безопасности ADT Corporation в судебном процессе.	The PL system preserves the company name and the person's surname across the paragraph. The SL system preserves the company name but translates the person's surname differently; moreover, it changes the person's gender from sentence to sentence.
lud-ru	Раньше генеральный директор Ringo, Джейми Симинов, заметил, что эта фирма ушла, потому что он не услышал звон дверей в своей гаражной мастерской. Он рассказал, как сделал дверной колокольчик с варежкой. Симинов рассказал, что продажи выросли после его выступления в программе «Шаркий танк» в 2013 году, где судьи не стали признавать его стартап. В конце 2017 года Симинов появился на покупном TV-канале QVC. Кинг также договорился о судебном процессе с компанией по охране прав конкурентов ADT Corporation.	Ранее генеральный директор Ring Джейми Симинов заметил, что эта фирма пошла по тому пути, что он не слышал дверные колокольчики в своей гаражной мастерской. Он сказал, что сделал колокольчик с помощью скатерти. Симинов сказал, что продажи выросли после его появления в программе Shark Tank в 2013 году, где судьи не стали понимать его стартап. В конце 2017 года Симинов проиграл дело на QVC-телеканале. Ring договорился также о судебном процессе с компанией по охране конкурентов ADT Corporation.	The PL system preserves the company name and the person's surname across the paragraph. The SL system translates the company name first in Latin, changing it, then in Cyrillic, getting it wrong once again. The person's surname is translated consistently.
vep-ru	Раньше начальник компании «Круг» Зами Симинов заметил, что эта кампания началась, когда он не услышал стучащихся дверей на своем дворе. Он сказал, что сделал Wi-Fi-установку. Симинов сказал, что продажи улучшились, когда он появился в 2013 году в телепередаче «Shark Tank», в которой члены жюри согласились выделить деньги на его проект. В конце 2017 года Симинов появился на передаче QVC. Кроме того, компания «Ринг» подала в суд на своего конкурента – подпольную компанию «ADT Corporation».	Ранее глава компании «Ring» Жами Симинов заметил, что эта кампания началась, когда он не услышал дверной замк на своем автосалоне. Он сказал, что сделал Wi-Fi замк. Симинов сказал, что продажи улучшились, когда он появился в 2013 году в телепередаче «Shark Tank», в которой единогласное жюри решило дать деньги его проекту. В конце 2017 года Симинов появился на телепередаче QVC. Кроме того, компания «Ring» устроила судебные разбирательства со своей конкурентк̆ой – компанией-покровителем «ADT Corporation».	The PL system preserves the company name and the person's surname across the paragraph. The SL system translates the company name in two different ways: first, it is a literal translation (Ring—Круг), then it is a transliteration of the English title (Ring—Ринг). The surname is translated in the same fashion across the paragraph.

Table 3: Translations of the same paragraph from FLORES-200 performed by the sentence-level (SL) MADLAD and paragraph-level (PL) MADLAD in three translation directions, demonstrating the preservation of proper nouns. Underlined with a straight line comes a company name (Ring), underlined with a wavy line comes a person's surname (Siminoff).

	PL	Neurotõlge
krl-ru	$22.0 \pm 2.0/49.1 \pm 1.8$	$23.4 \pm 2.0/50.4 \pm 1.6$
lud-ru	$19.3 \pm 1.8/46.2 \pm 1.5$	$21.7 \pm 2.0/48.2 \pm 1.4$
olo-ru	$22.1 \pm 2.2/48.5 \pm 1.7$	$25.9 \pm 2.4/51.4 \pm 1.8$
vep-ru	$20.7 \pm 1.8/46.3 \pm 1.8$	$26.5 \pm 2.5/51.2 \pm 1.8$
ru-krl	$13.4 \pm 1.5/46.8 \pm 1.3$	$10.6 \pm 1.3/43.5 \pm 1.2$
ru-lud	$4.0 \pm 1.0/33.3 \pm 1.0$	$3.6 \pm 1.0/31.6 \pm 1.2$
ru-olo	$8.4 \pm 1.4/40.6 \pm 1.2$	$7.0 \pm 1.3/36.2 \pm 1.2$
ru-vep	$12.0 \pm 1.7/43.0 \pm 1.5$	$12.1 \pm 1.5/42.9 \pm 1.4$

Table 4: Comparison between our paragraph-level (PL) translation system and Neurotõlge for translation directions from/into Russian. BLEU and chrF++ scores (separated by slash) of Neurotõlge and paragraph-level (PL) MADLAD as evaluated on the paragraph-level Smugri FLORES benchmark.

We also identified several types of errors specific to translations into Russian. For instance, the same word may appear in translation in its original form ("Ring"), as a literal translation into Russian from English ("Круг"), or as a transliteration into Cyrillic script ("Ринг"). The SL model more frequently combines these three forms inconsistently within the same text compared to the PL model.

The second part of the analysis focuses on coreference resolution, specifically examining the use of pronouns. While many paragraphs in the benchmark dataset mention people, most of them are about men. Both the SL and PL models translated gender-related structures correctly in most cases, typically defaulting to the male gender. However, we found examples where both models struggled with gender, although the PL model made fewer mistakes overall.

To illustrate our findings, we present a paragraph containing examples of lexical cohesion and coreference resolution. Table 3 provides translations of this paragraph generated by the SL and PL systems. It is translated into Russian from Proper Karelian, Ludian, and Veps, with translations from Livvi omitted to save space. English reference of the paragraph is provided below:

Previously, Ring's CEO, Jamie Siminoff, remarked the company started when his doorbell wasn't audible from his shop in his garage. He built a WiFi door bell, he said. Siminoff said sales boosted after his 2013 appearance in a Shark Tank episode where the show panel declined funding the startup. In late 2017, Siminoff appeared on shopping television channel QVC. Ring also settled a lawsuit with competing security company, the ADT Corporation.

A detailed explanation of the mistakes made by the systems is presented in Table 3. As we can see, the results highlight the PL model's ability to effectively handle discourse-level phenomena.

6.3 Comparison with previous results

We compared the results of our paragraph-level model with translations generated by the online machine translation engine Tartu NLP Neurotõlge. The online system demonstrates significantly better performance when translating *into* Russian (Table 4). However, our model outperforms Tartu NLP Neurotõlge when translating *from* Russian to Proper Karelian and Livvi and shows comparable results for Ludian and Veps. For example, in the Russian→Proper Karelian direction, the PL model beats Neurotõlge by 2.8 BLEU or 3.3 chrF++.

6.4 Zero-shot English translation

In this final experiment, we investigated how well the translation abilities of the models transferred to unseen pairs of languages in the example of English. We translated the FLORES-200 benchmark from Proper Karelian, Livvi, Ludian, and Veps to English and back. The results are shown in Table 5.

First, we notice that the original model without fine-tuning already has high scores for translation *into* English. This probably means that the model transferred its knowledge of Finnish and Estonian to their low-resource relatives. Besides, MADLAD-400 has seen much more data in English than in any other language, which may account for the scores being bigger than for zero-shot translation into Russian.

Next, we observe the boost in accuracy after fine-tuning, which tells us that the knowledge has

	base	SL	PL	p-value
krl-en	$20.8 \pm 2.1 / 49.4 \pm 1.5$	$25.5 \pm 1.8 / 53.7 \pm 1.3$	$27.1 \pm 2.1 / 55.0 \pm 1.4$	**0.025 / 0.003**
lud-en	$11.4 \pm 1.9 / 37.5 \pm 1.8$	$19.2 \pm 1.8 / 47.6 \pm 1.4$	$20.1 \pm 2.1 / 48.8 \pm 1.5$	**0.037 / 0.005**
olo-en	$10.3 \pm 1.6 / 36.7 \pm 1.7$	$17.9 \pm 1.6 / 45.9 \pm 1.4$	$18.1 \pm 1.7 / 46.8 \pm 1.4$	0.240 / **0.007**
vep-en	$6.4 \pm 1.5 / 31.6 \pm 1.8$	$14.8 \pm 1.8 / 43.4 \pm 1.5$	$13.4 \pm 1.6 / 42.9 \pm 1.4$	**0.003** / 0.071
en-krl	$0.9 \pm 0.5 / 4.7 \pm 0.8$	$15.3 \pm 1.8 / 48.0 \pm 1.4$	$13.5 \pm 1.6 / 46.5 \pm 1.3$	**0.002 / 0.001**
en-lud	$0.3 \pm 0.1 / 2.8 \pm 0.4$	$3.2 \pm 1.2 / 31.6 \pm 0.9$	$3.7 \pm 1.2 / 32.1 \pm 0.9$	0.051 / **0.031**
en-olo	$0.6 \pm 0.4 / 3.1 \pm 0.5$	$7.3 \pm 1.3 / 37.3 \pm 1.1$	$6.8 \pm 1.1 / 36.7 \pm 1.1$	0.101 / **0.012**
en-vep	$0.5 \pm 0.3 / 3.6 \pm 0.5$	$7.7 \pm 1.3 / 37.6 \pm 1.3$	$7.9 \pm 1.2 / 37.7 \pm 1.3$	0.259 / 0.184

Table 5: Zero-shot performance for translation from/into English. BLEU and chrF++ scores (separated by slash) of base MADLAD, sentence-level (SL) MADLAD, and paragraph-level (PL) MADLAD evaluated on the paragraph-level Smugri FLORES benchmark. p-value is the probability that SL and PL models are the same with respect to each metric; p-value less than 0.05 indicates that the difference between the models is statistically significant (highlighted in bold).

been successfully transferred to the unseen case of English translation. The scores for translation *into* English exceed those for translation *into* Russian and go up to 27.1 BLEU and 55.0 chrF++ in the case of Proper Karelian→English translation. As for the translation *from* English, the scores remain nearly equal to those for translation *from* Russian.

The ratio of capabilities of the sentence-level and paragraph-level models changes from case to case, with both BLEU and chrF++ metrics sometimes indicating the significant superiority of the PL system (Proper Karelian→English, Ludian→English) and sometimes the superiority of the SL system (English→Proper Karelian). As no direct fine-tuning, there is no wonder that the results oscillated so much.

However, the key indicator for us is the ability of the models to handle discourse-level phenomena. As all the languages in question have Latin script, the issue with translating proper names becomes less pronounced. Yet, the distinction between the models is apparent when it comes to gender consistency. For the example explored in Subsection 6.2, the SL model inconsistently shifts gender when translating sentences from any studied language into English. The PL model, unlike the SL, consistently and accurately translates gender across the paragraph for all languages.

7 Conclusion

In this paper, we developed a machine translation system for four low-resource Finno-Ugric languages: Proper Karelian, Livvi, Ludian, and Veps. Unlike previous MT systems that cover the same languages, ours is paragraph context-aware. The analysis showed that the model consistently translates names and terminology, though, it still encounters difficulties with coreference resolution.

The developed system has been trained only on parallel corpora with Russian. Nevertheless, the system is also capable of translating to and from English, despite not being trained to do so, with paragraph-level abilities being successfully transferred to this case.

Additionaly, we presented a FLORES-based benchmark dataset for Proper Karelian (Viena), Ludian, and Veps. The collected paragraph-level corpora are released as HuggingFace scripts that will allow one to re-collect the data.

We leave for future work experiments with more Finno-Ugric languages, including creating a paragraph-level benchmark that enables a more thorough evaluation of discourse-level phenomena handling. It would also be interesting to compare our results with multilingual decoder-only models, as many of these are starting to emerge.

Acknowledgments

This work was partially supported by the Estonian Research Council grant PRG2006 as well as the National Programme of Estonian Language Technology grant EKTB67. All computations were performed on the LUMI Supercomputer through the University of Tartu HPC center. The authors also thank the University of Eastern Finland and the Karelian language revival project for their valuable consultation and support.

References

Rachel Bawden, Rico Sennrich, Alexandra Birch, and Barry Haddow. 2018. Evaluating discourse phenomena in neural machine translation. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), pages 1304–1313, New Orleans, Louisiana. Association for Computational Linguistics.

Tatyana Boyko, Nina Zaitseva, Natalia Krizhanovskaya, Andrew Krizhanovsky, Irina Novak, Nataliya Pellinen, and Aleksandra Rodionova. 2022. The open corpus of the veps and karelian languages: Overview and applications. KnE Social Sciences, 7(3):29–40.

Daniel Deutsch, Juraj Juraska, Mara Finkelstein, and Markus Freitag. 2023. Training and meta-evaluating machine translation evaluation metrics at the paragraph level. In Proceedings of the Eighth Conference on Machine Translation, pages 996–1013, Singapore. 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, et al. 2021. Beyond english-centric multilingual machine translation. Journal of Machine Learning Research, 22(107):1–48.

Christian Hardmeier. 2012. Discourse in statistical machine translation: A survey and a case study. Discours, (11).

Christian Hardmeier, Sara Stymne, Jörg Tiedemann, and Joakim Nivre. 2013. Docent: A document-level decoder for phrase-based statistical machine translation. In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations, pages 193–198, Sofia, Bulgaria. Association for Computational Linguistics.

Amir Hossein Kargaran, Ayyoob Imani, François Yvon, and Hinrich Schuetze. 2023. GlotLID: Language identification for low-resource languages. In Findings of the Association for Computational Linguistics: EMNLP 2023, pages 6155–6218, Singapore. Association for Computational Linguistics.

Sneha Kudugunta, Isaac Caswell, Biao Zhang, Xavier Garcia, Derrick Xin, Aditya Kusupati, Romi Stella, Ankur Bapna, and Orhan Firat. 2023. Madlad-400: A multilingual and document-level large audited dataset. In Advances in Neural Information Processing Systems, volume 36, pages 67284–67296. Curran Associates, Inc.

Samuel Läubli, Rico Sennrich, and Martin Volk. 2018. Has machine translation achieved human parity? a case for document-level evaluation. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 4791–4796, Brussels, Belgium. Association for Computational Linguistics.

Sameen Maruf and Gholamreza Haffari. 2018. Document context neural machine translation with memory networks. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1275–1284, Melbourne, Australia. Association for Computational Linguistics.

Lesly Miculicich, Dhananjay Ram, Nikolaos Pappas, and James Henderson. 2018. Document-level neural machine translation with hierarchical attention networks. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 2947–2954, Brussels, Belgium. Association for Computational Linguistics.

NLLB Team, 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.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Hee-

woo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. Gpt-4 technical report.

Miikul Pahomov. 2017. Lyydiläiskysymys: kansa vai heimo, kieli vai murre? Ph.D. thesis, Humanistinen tiedekunta, Suomi.

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.

Tommi Pirinen, Francis M. Tyers, Trond Trosterud, Ryan Johnson, Kevin Unhammer, and Tiina Puolakainen. 2017. North-Sámi to Finnish rule-based machine translation system. In Proceedings of the 21st Nordic Conference on Computational Linguistics, pages 115–122, Gothenburg, Sweden. Association for Computational Linguistics.

Maja Popović. 2017. chrF++: words helping character n-grams. In Proceedings of the Second Conference on Machine Translation, pages 612–618, Copenhagen, Denmark. 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 EAMT 2024 Products and Projects track.

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.

Zewei Sun, Mingxuan Wang, Hao Zhou, Chengqi Zhao, Shujian Huang, Jiajun Chen, and Lei Li. 2022. Rethinking document-level neural machine translation. In Findings of the Association for Computational Linguistics: ACL 2022, pages 3537–3548, Dublin, Ireland. Association for Computational Linguistics.

Francis M. Tyers, Linda Wiechetek, and Trond Trosterud. 2009. Developing prototypes for machine translation between two Sami languages. In Proceedings of the 13th Annual Conference of the European Association for Machine Translation, Barcelona, Spain. European Association for Machine Translation.

Dusan Varis and Ondřej Bojar. 2021. Sequence length is a domain: Length-based overfitting in transformer models. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In Advances in Neural Information Processing Systems, volume 30. Curran Associates, Inc.

Lisa Yankovskaya, Maali Tars, Andre Tättar, and Mark Fishel. 2023. Machine translation for low-resource Finno-Ugric languages. In Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa), pages 762–771, Tórshavn, Faroe Islands. University of Tartu Library.