Graph-based RAG for Low-Resource Aromanian–Romanian Translation

Laurențiu Ghețoiu and **Sergiu Nisioi**
Human Language Technologies Research Center
Faculty of Mathematics and Computer Science
University of Bucharest
ghetoiu.laurentiu@gmail.com, sergiu.nisioi@unibuc.ro

Abstract

Aromanian, a linguistically and culturally significant yet low-resource Romance language, poses substantial challenges in computational linguistic research due to its limited NLP resources and non-standardized orthography. In this paper, we present an experimental study aimed at translating Aromanian texts into Romanian using a variety of modern NLP methodologies. We leverage two key resources: a parallel corpus consisting of approximately 3,000 sentence-aligned short stories and a dictionary of over 28,000 Aromanian-Romanian word pairs. Our approaches include Retrieval-Augmented Generation (RAG) supported by a graph-based alignment database, fine-tuning multilingual transformer models (specifically Meta's No Language Left Behind (NLLB)), and parameter-efficient fine-tuning techniques such as Low-Rank Adaptation (LoRA) applied to LLaMA-derived models. Evaluations using standard metrics such as Bilingual Evaluation Understudy (BLEU) and Character n-gram F-score (chrF) demonstrate varied effectiveness across these methodologies, highlighting the strong performance of NLLB for general translation tasks, while RAG excels in translating frequently seen content. Our findings underline the complexities inherent in low-resource language translation and provide valuable insights into effective digital preservation and Natural Language Processing (NLP) adaptation strategies for underrepresented languages.

1 Introduction

Aromanian is a low-resource language with significant cultural and linguistic value. Aromanian is a Romance language related to Romanian, but rooted in the Latin spoken south of the Danube during the Romanization period, in contrast with Romanian which is rooted in the Latin spoken north of the Danube (Sala, 2012). There are ongoing debates regarding its classification as a dialect of Romanian or as an independent language. Beyond its linguistic interest, Aromanian serves as a cultural identifier for the Aromanian people, who numbered between 400,000 and 500,000 across the Balkan Peninsula, Western Europe, North America and Australia (Saramandu, 2023). Among the countries with a significant number of speakers, in arbitrary order, we count Greece, Albania, North Macedonia, Bulgaria, Serbia, and Romania. Yet, the language remains underrepresented in NLP resources and lacks standardized writing conventions, making it both technically challenging and deeply important for digital preservation efforts.

Despite shared origins, Aromanian's unique vocabulary, syntax, and cultural significance present notable challenges in natural language processing.

Even after experimenting with a wide range of modern NLP models and fine-tuning approaches, our results show that truly usable Aromanian-Romanian translation remains out of reach when only extremely low-resource data is available.

2 Previous Work

The problem of translating low-resource languages has been studied in multiple ways. The most relevant and transferable to our task are Meta's NLLB models (Dale, 2023). Although there are better performing models, we stick to the ones that offered efficiency and worked on limited computational resources such as NLLB-200 Distilled (NLLB Team et al., 2024). Those models are trained on a dataset consisting of 3001 parallel sentences between 200 languages, including Romanian. By mapping the words in a shared spatial representation it allows the system to determine if a sentence is a translation of another, without the need of resourceful data. Thus, making this a good model for our low-resource problem because it can leverage knowl-

Proceedings of Recent Advances in Natural Language Processing,pages 388–394
Varna, Sep 8–10, 2025
https://doi.org/10.26615/978-954-452-098-4-047

edge from more resourceful languages with similar structure (like Romanian). We further fine-tune this model by adding the Romanian-Aromanian pairs of words and sentences, training in both translation directions.

Previous work by Jerpelea et al. (2025) already reports encouraging results on Aromanian-Romanian translation using larger datasets and similar multilingual models, including NLLB. Unlike these approaches, our aim is to train a model on severely scarce data of only 3,000 sentence pairs and a dictionary.

The other work that inspires our experiments is the attempt to translate a new language using only a grammar book (Tanzer et al., 2024). This also addresses multiple questions about LLMs such as their capacity to learn to perform new tasks when adapted, or if adaptation only brings up what the LLM had already learned. It seems that the task of translating a low-resource language is a good way of looking into some subjects that lack clarity. We also draw our own conclusions on some of these questions. Their work explores fine-tuning with Low-Rank Adaptation (LoRA) (Hu et al., 2021). LoRA is a method of fine-tuning neural network based models, by freezing some of their layers and adding a small number of new trainable parameters. The efficiency of this method is based on the fact that newly introduced parameters can be learned using low-rank matrices, which can then be used to adapt the original weights. This technique is part of the broader group of Parameter Efficient Fine-Tuning (PEFT). The authors also investigate different ways of prompting during fine-tuning, experimenting within the range of prompt lengths, giving more or less context. They also explore multiple retrieval strategies, including sentence overlap techniques.

3 Dataset

There are two main data sources on which the results of all the models are based. The short stories dataset (Petrariu and Nisioi, 2024) comprises approximately 3,000 sentence-aligned Aromanian texts with translations into Romanian, English, and French, providing a structured resource that captures Aromanian's linguistic particularities and facilitates machine translation tasks. We only use Aromanian and Romanian texts from this. The other source is an Aromanian-Romanian dictionary consisting of 28,647 words (Papahagi, 1974). We estimate the total word count to be around 100,000. Although the short-story corpus is already standardized across different writing forms, we still have to make some minor changes to ensure that both data sources use the same writing system. For example, this includes eliminating the "-mi" particle from some dictionary words and replacing letters or groups of letters in some words (e.g., *γumarlu* (with Greek γ) is transliterated as **yumarlu**; the cluster *ts* is simplified to a single **t**). However, diacritics are removed, accepting the risk of potential meaning loss, to ensure that all characters have consistent representations across different LLMs. An exception is made for the RAG variant, where diacritics are kept, as the specific LLMs used there often return responses containing diacritics. These can lead to inaccurate translations if the prompt words lack them. For example, the Romanian word *fata* ("girl") is not recognized at all in Aromanian and the LLM returns the Romanian word *față* (face), but if written with the correct diacritic *fată* it is correctly translated to Aromanian as *feată*.

In order to make our work easier, we create a dictionary representation that offers more possibilities for retrieving aligned words. In addition to direct translations, we add methods to the dictionary for obtaining similar words based on Levenshtein distance. We also test other metrics on cross-lingual embeddings, such as Euclidean and Cosine distances between the Romanian word and its Aromanian counterpart, to obtain alignments, but these do not provide better results. When querying the dictionary, multiple translation methods are available. If a direct translation exists in the dictionary, it will be provided. Alternatively, the dictionary can offer a translation based on a customizable similarity threshold. This threshold ensures that the returned word has a high degree of similarity to the queried word, minimizing the chance of returning dissimilar words. The similarity threshold can be adjusted to better suit specific needs, allowing for more flexible or stricter translation results depending on the context.

4 Methodology

4.1 Retrieval-Augmented Generation

Retrieval-Augmented Generation (RAG) is a technique that enhances language models by combining them with a retrieval mechanism. This approach uses external data sources, such as databases or graphs, to generate more accurate responses by fetching relevant information based on a query (NVIDIA, 2023). For our project, we use a graph-based database to assist with the translation of Aromanian to Romanian. This approach facilitates access to various alignment types and provides the large language model (LLM) with context to explore a wider array of indirectly connected, yet relevant, nodes (Edge et al., 2024). Before designing the RAG-based translation pipeline, we initially seek to establish direct correspondences between Aromanian and Romanian words. This was motivated by the availability of the Tache Papahagi dictionary (Papahagi, 1974), which provides reliable word-level alignments. We hypothesize that these can be leveraged directly for translation, allowing partial mappings of parts of speech and phrases. Building on this idea, we align dictionary entries and sentence pairs, providing them to a large language model via a RAG system (Lewis et al., 2021). After this, we began experimenting with transformer-based models using few-shot prompting or parameter-efficient fine-tuning techniques.

We construct a graph with three types of nodes: Romanian segments, Aromanian translations, and words appearing in both languages. The relationship between Romanian and Aromanian sentences is represented by a "translates_to" link. Each Aromanian word in a sentence has a relationship of "appears_in" that connects it to the corresponding sentence node. An example of the resulting architecture is illustrated in Figure 1.

Each Aromanian word node is enriched with several attributes: its part of speech, its translation into Romanian (from the dictionary), and a similarity score if the translation is not an exact match. The similarity score helps measure how closely the Aromanian word corresponds to a Romanian word, accounting for potential variations that are not directly found in the dictionary. This graph-based approach allows for more flexible and accurate translations by considering sentence structure, word alignment, and orthographic similarities across both languages.

When a new sentence is proposed for translation, it is split into words. The graph database searches for these words or, if they don't exist, for similar words. The corresponding sentence nodes are retrieved, along with other words in those sentences. These nodes and the information they contain are assembled into a prompt that can be well-understood by LLMs to translate new input. We test this only for Romanian-to-Aromanian translation, as in the other direction the LLMs do not seem able to generate any coherent answers. Table 1 reports the BLEU and chrF scores for different LLMs under both English and Romanian prompts, showing that Romanian prompts yield better results.

Model	BLEU Score	ChrF Score
English Prompt		
3.5 - Turbo	1.8989	14.6433
4o-mini	0.6895	11.4524
4o	1.0438	14.2354
Romanian Prompt		
3.5 - Turbo	3.6485	27.1208
4o-mini	8.0872	27.7183
4o	8.0988	33.7438

Table 1: BLEU and ChrF scores for different LLMs using a RAG system on Aromanian to Romanian translation. Results are shown for English and Romanian prompts.

4.2 NLLB

The next approach we try is using NLLB models (Costa-jussà et al., 2022). These models use transformer architecture (Vaswani et al., 2023; Yildirim, 2024). NLLB covers low-resource languages and was initially trained on multilingual dataset, providing a strong foundation in understanding language and translation. Unlike the RAG approach, for this one, we split the short stories dataset into sentence-level parallel data. If initially, some parallel entries in the dataset consisted of more than one sentence, now, for the short stories dataset, all entries are sentence-to-sentence. The dictionary data remains unchanged from the first version. This change was made due to poor results in initial training. Additionally, by reducing the length of the training samples, we are able to reduce the need to split a sample if it exceeds the maximum token sequence length.

This model is trained in both directions (Romanian -> Aromanian and Aromanian -> Romanian) for about 20,000 steps, using 27,000 training examples (Dale, 2023). 92% of the dataset entries are composed of dictionary words, with the remaining 8% being sentences. For testing, we use a dataset with a slightly higher proportion of sentences, about 10%, in order to evaluate the model's performance in more naturally occurring scenarios.

We only use NLLB200-600M, the distilled version provided by (Ott et al., 2019). The distilla-

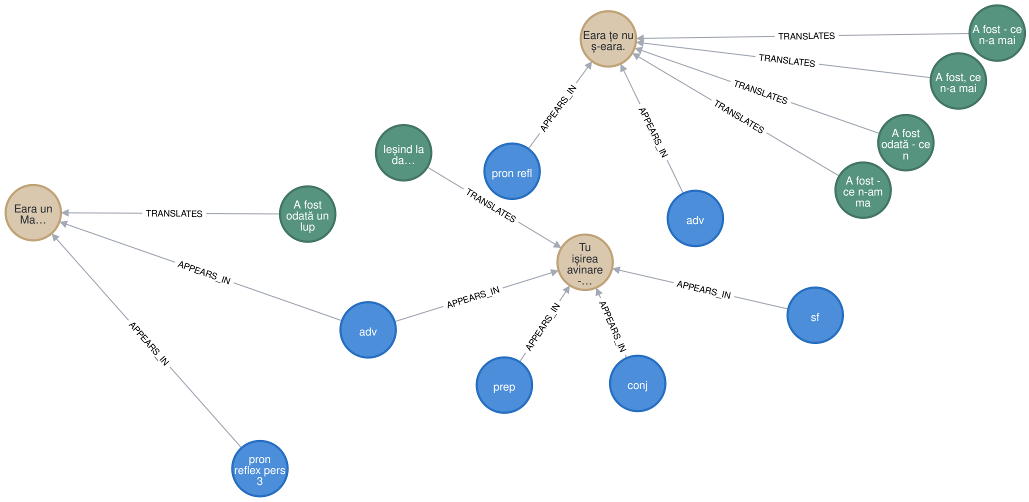

Figure 1: An image illustrating a graph database architecture used for RAG (visualization produced in Neo4j, Inc. (2024))

tion technique compresses a model into a smaller and more efficient one, while transferring as much knowledge as possible from the original (Hinton et al., 2015). This keeps resource usage low without needing additional resources during model training. The results are summarized in Table 2.

Direction	BLEU	chrF2
Aromanian → Romanian	5.62	21.91
Romanian → Aromanian	5.18	22.08

Table 2: Results for NLLB in translating from Aromanian to Romanian (first row) and from Aromanian to Romanian (second row).

4.3 Pretrained Transformers and LLMS

Several sources inspire our approach to adapting large language models on our dataset, notably the English-Kalamang translation initiative. This initiative focuses on translating between English and Kalamang, a low-resource language, using a grammar book. The book includes, among other resources, a bilingual word list and a small set of parallel sentences. (Tanzer et al., 2024) (Visser, 2022). This intersects with our dataset, which includes the parallel sentences from the short stories and the bilingual word list from the dictionary. In contrast to this project, we lack access to a structured language-learning textbook. Moreover, we impose specific constraints on processing power throughout our experimentation.

We primarily rely on Meta's LLaMA models (LLaMA2 (Touvron et al., 2023), LLaMA3 and LLaMA3.1 (et al., 2024), and LLaMA3.2), as well as their Romanian variants, RoLLaMA (Masala et al., 2024). The fine-tuning is parameter-efficient, where we train only the linear layers of the models. Using LoRA (Hu et al., 2021) and Parameter-Efficient Fine-Tuning (PEFT) (Mangrulkar et al., 2022), we train quantized versions (4-bit) of the models. The best results are obtained by RoLLaMA3-8b-Instruct with a BLEU score of 1.61 and a chrF of 18.26. However, the differences across models were minimal, indicating limited significance in model selection. For instance, RoLLaMA2, which is based on LLama2-7B yielded a BLEU score of 1.32 and a chrF of 15.02. All models are exclusively trained for the Aromanian-to-Romanian (rup -> ro) direction. Training is conducted for one or three epochs, using varying learning rates. For the best-performing model, we use r=16 and lora_alpha = 32, while for RoLLaMA2, r= 16 and lora_alpha=16. We experimented with the parameters, cross-checking them between models and adjusting them, but none led to improved performance.

As the large parameter models trained with LoRA and PEFT do not perform well, we also try smaller models. However, the dataset was not sufficient even for the smaller models to learn using parameter-efficient methods. For LLaMA3.2 1B, the BLEU score is 0.59 and chrF2 is 8.83. We observe that even before the end of one epoch, the loss begins fluctuating significantly and does not decrease meaningfully, as shown in Figure 2.

As a last resort, we explore other transformer-based language models with even fewer parameters. For this, we use Meta's BART model (Lewis

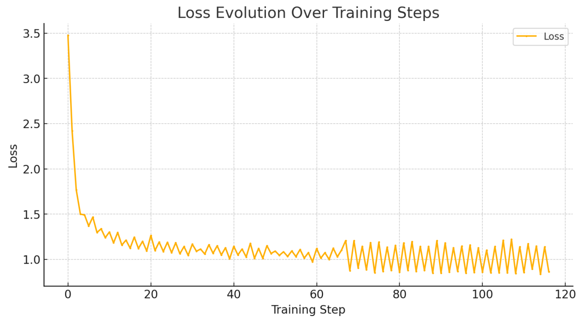

Figure 2: Loss fluctuation observed during training, showing significant instability before the end of one epoch.

et al., 2019). Given its reduced parameter count and simpler architecture, it performs comparably to larger models, even exceeding the 1B version. Training is done on the same dataset for 3 epochs with a 5e-5 learning rate, obtaining BLEU=1.28 and chrF=19.90. A comparison across different transformer models can be found in Table 3.

Model	BLEU Score	chrF Score
RoLLaMA3-8b	1.61	18.26
RoLLaMA2	1.32	15.02
LLaMA3.2 1B	0.59	8.83
BART	1.28	19.90

Table 3: BLEU and chrF scores for various models fine-tuned for Aromanian-to-Romanian translation

5 Evaluation

We measure the quality of our translations mainly by using BLEU (Papineni et al., 2002) and chrF (Popović, 2015) scores. Although in sheer numbers RAG appears to perform the best, NLLB might produce better actual results. During the testing of RAG, we observed that some texts were recognized and translated correctly, seemingly because they were popular texts, such as The Lord's Prayer. This could be an indicator of data contamination in the NLLB pre-training.

6 Conclusions

In this study, we investigate and compare various modern NLP approaches for translating between Aromanian, an underrepresented and linguistically distinctive low-resource language, and Romanian. Leveraging a carefully curated dataset composed of a sentence-aligned corpus and a comprehensive dictionary, we experiment with Retrieval-Augmented Generation (RAG), multilingual transformer architectures like Meta's NLLB, and fine-tuning techniques involving parameter-efficient methods such as LoRA applied to LLaMA-based models.

Our experiments indicate that while NLLB, with its multilingual pretraining and relatively straightforward fine-tuning, consistently provides the most robust results among the methods tested, translation quality remains far from usable when only extremely limited data is available. RAG showed promise, particularly for translating known or highly frequent texts, yet its performance highlighted a susceptibility to memorization over generalization. Transformer-based models fine-tuned with LoRA provided limited success, reflecting the challenges imposed by dataset constraints typical in low-resource contexts.

Overall, our results make it clear that with such limited resources (3000 sentence pairs and a dictionary), none of the evaluated methods can deliver practical or reliable translations for dialectal Romanian - Aromanian machine translation.

Acknowledgments

This research is partially supported by the project "Romanian Hub for Artificial Intelligence - HRIA", Smart Growth, Digitization and Financial Instruments Program, 2021-2027, MySMIS no. 334906 and partially by InstRead: Research Instruments for the Text Complexity, Simplification and Readability Assessment CNCS - UEFISCDI project number PN-IV-P2-2.1-TE-2023-2007.

References

Abhimanyu Dubey et al. 2024. The llama 3 herd of models.

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 Dale. 2023. How to fine-tune a nllb-200 model for translating a new language. Accessed: 2024-11-02.

Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, and Jonathan Larson. 2024. From local to global: A graph rag approach to query-focused summarization.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models.

Alexandru-Iulius Jerpelea, Alina Radoi, and Sergiu Nisioi. 2025. Dialectal and low resource machine translation for Aromanian. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 7209–7228, Abu Dhabi, UAE. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2021. Retrieval-augmented generation for knowledge-intensive nlp tasks.

Sourab Mangrulkar, Sylvain Gugger, Lysandre Debut, Younes Belkada, Sayak Paul, and Benjamin Bossan. 2022. Peft: State-of-the-art parameter-efficient fine-tuning methods. `https://github.com/huggingface/peft`.

Mihai Masala, Denis C. Ilie-Ablachim, Alexandru Dima, Dragos Corlatescu, Miruna Zavelca, Ovio Olaru, Simina Terian-Dan, Andrei Terian-Dan, Marius Leordeanu, Horia Velicu, Marius Popescu, Mihai Dascalu, and Traian Rebedea. 2024. "vorbeşti româneşte?" a recipe to train powerful romanian llms with english instructions.

Neo4j, Inc. 2024. Neo4j Graph Database (version 5.15.0).

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 Mejía 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. 2024. Scaling neural machine translation to 200 languages. *Nature*, 630(8018):841–846.

NVIDIA. 2023. What is retrieval-augmented generation? Accessed: 2024-10-29.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of NAACL-HLT 2019: Demonstrations*.

Tache Papahagi. 1974. *Dicționarul dialectului aromân: general și etimologic*. Editura Academiei Republicii Populare Romane, Bucharest, Romania. Free copy available from Institutul de Lingvistică al Academiei Române „Iorgu Iordan - Al. Rosetti”.

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.

Iulia Petrariu and Sergiu Nisioi. 2024. A multilingual parallel corpus for Aromanian. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 832–838, Torino, Italia. ELRA and ICCL.

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.

Marius Sala. 2012. *De la Latină la Română*. Editura Pro Universitaria, București, România.

Nicolae Saramandu. 2023. Aromânii. istorie, literatură. scrieri despre dialectul aromân. *Fonetică și dialectologie*, XLII.

Garrett Tanzer, Mirac Suzgun, Eline Visser, Dan Jurafsky, and Luke Melas-Kyriazi. 2024. A benchmark for learning to translate a new language from one grammar book.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2023. Attention is all you need.

E. Visser. 2022. *A grammar of Kalamang*. Comprehensive Grammar Library. Language Science Press.

Savas Yildirim. 2024. Fine-tuning transformer-based encoder for turkish language understanding tasks.