Nesciun Lengaz Lascià Endò: Machine Translation for Fassa Ladin*

Giovanni Valer[1,*], Nicolò Penzo[1,2] and Jacopo Staiano[1]

[1]University of Trento, Italy
[2]Fondazione Bruno Kessler, Trento, Italy

Abstract

Despite the remarkable success recently obtained by Large Language Models, a significant gap in performance still exists when dealing with low-resource languages which are often poorly supported by off-the-shelf models. In this work we focus on Fassa Ladin, a Rhaeto-Romance linguistic variety spoken by less than ten thousand people in the Dolomitic regions, and set to build the first bidirectional Machine Translation system supporting Italian, English, and Fassa Ladin. To this end, we collected a small though representative corpus compounding 1135 parallel sentences in these three languages, and spanning five domains. We evaluated several models including the open (Meta AI's No Language Left Behind, NLLB-200) and commercial (OpenAI's gpt-4o) state-of-the-art, and indeed found that both obtain unsatisfactory performance. We therefore proceeded to fine-tune the NLLB-200 model on the data collected, using different approaches. We report a comparative analysis of the results obtained, showing that 1) jointly training for multilingual translation (Ladin-Italian and Ladin-English) significantly improves the performance, and 2) knowledge-transfer is highly effective (e.g., leveraging similarities between Ladin and Friulian), highlighting the importance of targeted data collection and model adaptation in the context of low-resource/endangered languages for which little textual data is available.

Keywords

Machine Translation, Low Resource Languages, Dialects, Ladin

1. Introduction

The growing scale of Large Language Models, based on the Transformer architecture, has led to models with surprising capabilities in a number of tasks, including Machine Translation (MT). However, most of the NLP community effort is focused on *high-resource* standardized languages, leaving behind the vast majority of local *under-resourced* languages. Recent works have demonstrated the utility of creating language-specific datasets for MT [1] and the effectiveness of relatively small quantities of high-quality translation data to teach a new language to pre-trained LLMs [2, 3]. To date, little work has addressed the Ladin language: even the most recent models that have included a great number of languages have not been trained with Ladin data [4], due to the scarcity of freely available parallel corpora (to our knowledge, only the OPUS corpora [5]), which are also poorly curated – e.g., wrong translations or mixed up Ladin varieties.[1]

Further, previous works have mainly focused on the two South Tyrolean varieties, *Gherdëina* and *Badiot* [6]: despite having a standardized written form and being officially recognized as a minority language, the Fassa variety (*Fascian*) has been mostly overlooked [7], while its speakers rightfully expect access to the same digital tools available for other languages [8].

We introduce the first dataset of parallel Fassa Ladin-Italian-English sentences, spanning over multiple domains: literature, news, laws, brochures, and game rules.

We evaluate several *out-of-the-box* translation systems, including the open (Meta AI's No Language Left Behind, NLLB-200) and commercial (OpenAI's gpt-4o) state-of-the-art models, and experiment with both *zero-shot* pivot-based and multilingual strategies to obtain satisfactory performances in bidirectional translation between Fassa Ladin and Italian/English. Figure 1 provides a schematic overview of our experiments, which are thoroughly described in Section 4.

Our results show how the collection of small quantities of parallel data is very effective in 'adding' support for a previously unsupported language to existing state-of-the-art models. More specifically, we find that the NLLB-200 model fine-tuned using a multilingual strategy can outperform even the most capable commercial LLMs (e.g., OpenAI gpt-4o).

For reproducibility purposes, we make the dataset and

CLiC-it 2024: Tenth Italian Conference on Computational Linguistics, Dec 04 – 06, 2024, Pisa, Italy

* *No Language Left Behind* translates to *Nesciun Lengaz Lascià Endò* in Fassa Ladin.

*Corresponding author.
giovanni.valer@studenti.unitn.it (G. Valer); nicolo.penzo@unitn.it (N. Penzo); jacopo.staiano@unitn.it (J. Staiano)
https://github.com/jo-valer (G. Valer); https://nicolopenzo.github.io (N. Penzo); https://www.staiano.net (J. Staiano)
0009-0002-2145-9497 (G. Valer); 0009-0006-8648-3307 (N. Penzo); 0000-0002-1260-4640 (J. Staiano)

© 2024 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

[1]See Appendix A.

CEUR-WS.org/Vol-3878/104_main_long.pdf

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

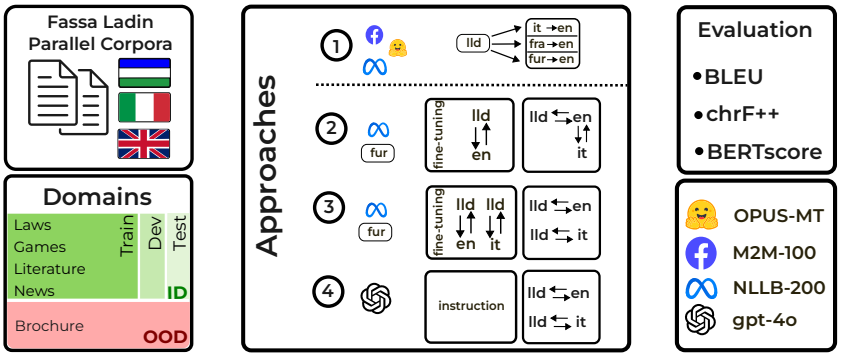

Figure 1: Our experimental setting: from the collected parallel corpora of Fassa Ladin, Italian and English we obtain training and validation data, along with both in-domain (ID) and out-of-domain (OOD) test sets; we evaluate 4 approaches: (1) use pretrained machine translation models treating the lld input as either Italian (it), French (fra) or Friulian (fur); (2) fine-tune NLLB-200 on the $lld \rightleftarrows en$ translation task, using Friulian as starting point; (3) fine-tune NLLB-200 on both $lld \rightleftarrows en$ and $lld \rightleftarrows it$; (4) zero-shot translation with gpt-4o.

code publicly available.[2]

2. Linguistic background

Ladin[3] (ISO 639-3 code: lld) is a Rhaeto-Romance language. It has numerous varieties, each one spoken in a different valley: *Anpezan* (Cortina d'Ampezzo), *Badiot* (Badia Valley), *Fascian* (Fassa Valley), *Fodom* and *Col* (Upper Cordevole Valley), and *Gherdëina* (Gardena Valley) [9]. This paper focuses on Fassa Ladin, which is spoken by approximately 8000 people and is further divided in three local varieties: *Cazét* (upper valley), *Brach* (lower valley), and *Moenat* (Moena). However, a standard variety for Fassa Ladin (named *Ladin fascian*) was established in 1999 and is currently used in official contexts; this is the variety considered in our work.

From a linguistic standpoint, Fassa Ladin is related to Italian. It also shares some linguistic phenomena with French, as the fronting of Latin /a/ to /ɛ/, e.g., PATER > fr. and lad. *père* (notice that both Ladin and French are Western Romance languages). Ladin is closely related also to Friulian, another Rhaeto-Romance language [9]. For these reasons we will consider Italian, French and Friulian for our experiments. We report in Table 1 an example of a sentence in Ladin, Italian and English.

Ladin	L porta dant azions per didèr dò la medema oportunità anter eles e ic.
Italian	Promuove azioni per favorire pari opportunità tra donne e uomini.
English	It promotes actions to foster equal opportunities between women and men.

Table 1
Parallel Ladin/Italian/English sample.

3. Data

We built the first Fassa Ladin-Italian-English parallel corpus drawing from multiple resources in 5 domains: literature, news, games, laws, and brochures. The literature subset is an excerpt of a collection of poems and stories by Galante et al. [10].

News are sourced from the Province of Trento press office releases[4] and from social networks' news.[5] The *games* subset contains parallel sentences from an online game.[6] Laws come from the *Statuto del Comune di Moena* (Statute of the Municipality of Moena)[7] and the *Statuto del Comun general de Fascia* (Statute of the 'Comun general de Fascia').[8] Finally, the *brochures* subset consists in promotional documents for tourists.[9] The latter exhibits distinct linguistic characteristics, and is characterized by poorly aligned sentences and more 'creative' translations; an example is provided in Table 2.

Thus, we used it for *out-of-domain* testing (see Section 4.3.1). The dataset compounds to 1135 parallel sentences, unevenly distributed across domains (see Table 3).

[2] https://github.com/jo-valer/machine-translation-ladin-fascian
[3] The term 'Ladin' can refer to multiple languages. In this paper we use it only in reference to the *Ladin of the Dolomites*, spoken in the so called *Ladinia brissino-tirolese*, across the provinces of Trento, Bolzano, and Belluno.
[4] https://www.ufficiostampa.provincia.tn.it/
[5] https://www.facebook.com/UalUnionAutonomistaLadina/
[6] http://avventuresuimontipallidi.it/
[7] https://it.wikisource.org/wiki/Comun_de_Moena_-_Statut
[8] https://www.consiglio.provincia.tn.it/_layouts/15/dispatcher/doc_dispatcher.aspx?app=clex&at_id=21177
[9] https://www.giornaletrentino.it/cronaca/fiemme-e-fassa/il-libro-sui-ladini-di-fascia-spacca-presto-altre-4-mila-copie-1.2242774

en: Especially in winter, when work in the fields was less intense.
it: Questi riti venivano celebrati soprattutto in inverno, quando il lavoro nei campi era meno intenso. *(These rites were celebrated mainly in winter, when work in the fields was less intense.)*
lld: Soraldut via per l'invern, ajache zacan l'era na sajon de paussa dal lurier te ciamp. *(Especially during the winter, as it used to be a season of respite from work in the field.)*

Table 2
An example of poorly aligned sentences from the brochures subset of our dataset. English translations for **it** and **lld** are provided in *italic*.

Subset	Orig. lang.	Sentences	
Laws	lld, it	742	65.4%
Games	lld, it, en	150	13.2%
Literature	lld, it, en	144	12.7%
News	lld, it	42	3.7%
Brochures	lld, it, en	57	5.0%
Total		1135	100%

Table 3
Domain distribution of sentences in our collected dataset.

When English translations were not available we used DeepL[10] to translate Italian into English.

4. Models and Methods

In our experiments we used the following machine translation model families:

- OPUS-MT, which provides unidirectional bilingual models [11];[11]
- M2M-100, a Many-to-Many multilingual model that can translate directly between any pair of 100 languages [12];
- NLLB-200, Meta AI's successor of M2M-100, supporting 200 languages [4];
- gpt-4o, the closed-source, state-of-the-art, general-purpose, instruction-tuned, multilingual model developed and commercialized by OpenAI.[12]

More implementation details are in Appendix C.

4.1. Experimental Setup

For model evaluation and validation, we prepare two held-out corpora, each of 108 aligned sentences ($\sim 10\%$ of the in-domain corpus), randomly sampled from all resources; the *brochures* subset was excluded from the training/evaluation splits and held out for out-of-domain evaluation (see Section 4.3.1). Three automatic evaluation metrics were used:

- BLEU [13], a commonly used metric based on lexical overlaps;[13]
- chrF++ [14], based on character n-gram precision and recall enhanced with word n-grams;[13]
- BERTscore [15], which uses a pretrained model (in our case, Multilingual BERT [16]) to compute pairwise token-level similarity scores between candidate and reference sentences.[14]

We chose BLEU and chrF++ metrics in line with previous work by Haberland et al. [1]. Although Multilingual BERT does not explicitly support the Ladin language, we assessed during preliminary analyses its alignment with human similarity judgments on Ladin sentences. For this reason we include it as reference for future work.

4.2. Preliminary Experiments

Firstly, we evaluate the performance of the pre-trained models in translating between Italian and English ($it \rightarrow en$ and $en \rightarrow it$), in order to have a reference for subsequent experiments. The evaluation is performed using our in-domain test set. We also evaluate the performance of the models to translate from Ladin to English, either considering Ladin sentences as if they were written in Italian, French, or Friulian. Such test allows us to have a measure of how much a given model is 'prepared' to transfer knowledge across these languages. NLLB-200 is the only model pre-trained with Friulian data, thus comparing models with this language is not possible. Nevertheless, this preliminary experiment is a viable way to investigate which language has the highest similarity to Ladin from the model's perspective.

Preliminary Results The results presented in Table 4 show how M2M-100 has lower scores for all metrics, and suggest that the best model for our experiments is NLLB-200; for this reason in the following we will consider

[10] https://www.deepl.com/
[11] https://huggingface.co/Helsinki-NLP/opus-mt-it-en
[12] The prompting strategy used for gpt-4o is presented in Appendix B.
[13] https://github.com/mjpost/sacrebleu
[14] https://github.com/Tiiiger/bert_score

Task	Model	BLEU	chrF++	BERTScore
$it \rightarrow en$	OPUS-MT	**55.61**	**73.60**	**91.68**
	M2M-100	44.18	66.39	88.40
	NLLB-200	52.93	70.65	90.33
$en \rightarrow it$	OPUS-MT	**44.35**	**67.67**	**91.33**
	M2M-100	32.40	59.06	87.24
	NLLB-200	40.13	64.51	89.48
$lld_{ita} \rightarrow en$	OPUS-MT	3.90	25.73	68.92
	M2M-100	4.84	28.81	68.52
	NLLB-200	18.52	43.83	80.05
$lld_{fra} \rightarrow en$	M2M-100	3.06	24.24	69.17
	NLLB-200	13.32	39.13	78.03
$lld_{fur} \rightarrow en$	NLLB-200	**21.76**	**46.76**	**81.74**

Table 4
Performance of the pre-trained models on different translation tasks, where lld_{ita}, lld_{fra} and lld_{fur} identify texts in Ladin, but presented to the model as if they were in Italian, French and Friulian, respectively.

this model only. We can notice a lower performance in $en \rightarrow it$, compared to $it \rightarrow en$, according to the untrained metrics; BERTscore provides instead comparable verdicts for the two tasks. This is an important finding and has to be recalled when evaluating subsequent experiments. Moreover, Friulian proves to be the most promising language for our fine-tuning purposes, even though Italian has good scores (BLEU score 21.76 vs. 18.52).

4.3. Transfer Learning Experiments

The training set consists of 862 parallel Fassa Ladin-Italian-English sentences (i.e., those remaining of the original 1135 sentences after excluding 108 for validation, 108 for in-domain test and 57 for out-of-domain test). As Ladin is not included in the pre-trained NLLB-200 model, we assign it the language code of Friulian, to leverage the similarities between these two languages. In this work we use our dataset for model fine-tuning, a relatively affordable strategy in terms of computational costs.[15] We experiment with the following approaches to add Fassa Ladin to the NLLB-200 model:

***Zero-shot* Pivot-based Transfer Learning** We fine-tune the model to only translate from English to Ladin (and viceversa), thus ignoring the Italian data. The pivot-based approach has proven to be effective for several languages [18]. We adopt a *zero-shot* pivot-based approach, meaning we do not fine-tune the model to perform $it \rightleftarrows en$, as we assume not to have the data: we investigate if such model performs well in $it \rightarrow lld$ even though it is not trained with Italian-Ladin pairs. We refer to the model fine-tuned with this approach as 'NLLB-pivot'.

Multilingual Translation We fine-tune the model for joint Ladin-Italian and Ladin-English bidirectional translation. Each batch includes a randomly selected pair of languages, in a single direction. We refer to the model fine-tuned with this approach as 'NLLB-multi'.

4.3.1. Transfer Learning Across Domains

We evaluated the model ability to generalize in different domains by testing it on our out-of-domain test set: the *brochures* subset (excluded from the training set) compounding to $\sim 5\%$ of the sentences in our entire dataset.

4.3.2. Forgetting of Previous Knowledge

Finally, we investigate whether the fine-tuned models suffer a performance drop in translating Italian to English (and vice versa), thus exploring if we encounter catastrophic forgetting [19]. We re-evaluate the models on our test set, and compare the results with the scores obtained in the preliminary experiments.

5. Results

The performances obtained by the fine-tuned models, for each translation task and for each test set, are reported in Table 5. As a strong baseline, we used gpt-4o.

5.1. Fine-tuning Approaches

The results show that both fine-tuning approaches are effective in adding Fassa Ladin to the pre-trained NLLB-200 model, increasing the BLEU score baseline of $lld_{fur} \rightarrow en$ from 21.76 to 40+, and outperforming gpt-4o (28.19). The two approaches achieve also similar results in $en \rightarrow lld$. Table 6 provides some examples of translated sentences.

We do not observe consistently higher scores by using the *zero-shot* pivot-based transfer learning approach. This might be due to the little amount of data used for fine-tuning, so that training also with Italian-Ladin parallel sentences helps by providing more data and higher diversity. Since we fixed the number of training steps for NLLB-pivot and NLLB-multi, the NLLB-multi model has seen about half of the Ladin-English batches compared to NLLB-pivot (the other half being Ladin-Italian).

This suggests that the multilingual translation approach might be preferable in the context of endangered languages for which little data is available, since it acts as a regularization method during training.

[15] Nonetheless, the increasing input context length of current LLMs allows for using many-shot in-context learning approach as shown in the concurrent work of Agarwal et al. [17], which we leave to future works.

Task	Model	*in-domain test* BLEU	chrF++	BERTScore	*out-of-domain test* BLEU	chrF++	BERTScore
$lld \rightarrow en$	gpt-4o	28.19	53.86	85.09	**26.48**	**50.51**	**86.40**
	NLLB-pivot	**41.08**	**62.68**	**88.00**	21.69	44.42	82.95
	NLLB-multi	40.17	62.11	87.82	21.70	44.34	82.86
$en \rightarrow lld$	gpt-4o	6.53	32.09	73.85	5.49	29.60	73.13
	NLLB-pivot	31.88	**56.16**	**82.95**	**12.61**	**38.97**	**76.68**
	NLLB-multi	**32.23**	56.09	82.44	10.57	38.05	**76.68**
$lld \rightarrow it$	gpt-4o	33.65	59.97	87.28	**29.10**	**52.08**	**86.13**
	NLLB-pivot	9.55	37.92	77.88	7.89	32.44	77.89
	NLLB-multi	**42.71**	**65.35**	**88.99**	20.29	44.94	82.35
$it \rightarrow lld$	gpt-4o	8.81	36.51	75.41	5.54	31.55	74.08
	NLLB-pivot	33.79	57.89	83.37	15.00	40.91	77.92
	NLLB-multi	**39.75**	**62.04**	**84.88**	**15.97**	**41.23**	**78.10**

Table 5
In-domain and Out-of-domain performances of gpt-4o and the models fine-tuned on our dataset, using Friulian as starting point.

English	GT Ladin	MT Ladin	BLEU	chrF++	BERTScore
The national communication campaign on the Electronic Health Record 2.0 also starts in the province of Trento	Ence te la provinzia de Trent pea via la campagna nazionèla de comunicazion sul Fascicol Sanitèr Eletronich 2.0	*Ence te la provinzia de Trent pea via la campagna de comunicazion nazionèla su l Fascicol Sanitèr Eletronich 2.0.*	62.44	90.22	93.87
The municipal council must adopt rules of procedure governing its functioning.	L Consei de Comun cogn se dèr n regolament che disciplinee so funzionament.	*L Consei de Comun cogn aproèr l regolament per so funzionament.*	36.96	60.82	86.25
He has an iron stomach.	L'à n stomech de fer acialinà.	*El à n stamp de fer acialà .*	13.13	42.18	85.50

Table 6
Examples of English sentences translated to Fassa Ladin using NLLB-multi, sorted by scores. We highlight the words of the machine translated (MT) sentences that differ from the ground truth (GT, whose corresponding words are underlined) using colors: completely wrong, imprecise but acceptable, substantially correct.

Turning to gpt-4o performances, it proves to better perform in $lld \rightarrow it$ task than $lld \rightarrow en$. Its scores are lower compared to our models, but the most significant finding is that it cannot generate text in Fassa Ladin ($it/en \rightarrow lld$). NLLB-multi performance in $it \rightarrow lld$ is much higher than $en \rightarrow lld$ (BLEU score 39.75 vs. 32.23), a finding calls for further analysis, left to future works, to be interpreted. We also observe NLLB-pivot performing poorly in $lld \rightarrow it$, but not in $it \rightarrow lld$. The *zero-shot* pivot-based approach appears to work in only one direction, a behavior we discuss in Section 5.3.

5.2. Domain Transfer

Unsurprisingly, a relatively lower performance on the out-of-domain test set is observed, since the original data presents less literal translations. As a consequence, the metrics matching the model output against the ground truth tend to lower scores. Still, especially for $lld \rightarrow en$, both NLLB-based models produce acceptable out-of-domain translations (BLEU scores 21+). The strong out-of-domain performance of gpt-4o, better than our models in understanding out-of-domain Ladin ($lld \rightarrow it/en$), shows how the scarcity of fine-tuning data, and its lack of linguistic diversity, has a negative impact on our models' performance. Another interpretation concerns the robustness of gpt-4o in handling grammatical errors: implicitly casting the source lld sentences to another similar language, known by the model, and then correctly translating into the en/it targets (e.g., treating Ladin words as if they were misspelled Italian words).

Model	ΔBLEU $it \rightarrow en$	ΔBLEU $en \rightarrow it$
NLLB-200	52.93	40.13
NLLB-pivot (Δ)	54.71 (+1.78)	7.72 (−32.41)
NLLB-multi (Δ)	52.95 (+0.02)	43.80 (+3.67)

Table 7
Performance shift of the fine-tuned models compared to the pre-trained NLLB-200 model (see Table 4), measured as BLEU score difference in Italian-to-English and English-to-Italian translation.

This would also explain the poor results when translating from en/it to lld.

5.3. Forgetting of Previous Knowledge

Finally, we present the performance shift in $it \rightarrow en$ and $en \rightarrow it$ of our fine-tuned models compared to the pre-trained NLLB-200 (Table 7). The idea is to evaluate the catastrophic forgetting phenomenon [20] after adding Fassa Ladin to the model, via the difference in BLEU scores. NLLB-multi produces slightly better translations after fine-tuning: this is expected, as it is better fitted to our domain. NLLB-pivot, however, has a strong drop in $en \rightarrow it$ (-32.41), but not in $it \rightarrow en$ ($+1.78$).

This suggests that after fine-tuning the model's encoder retained the ability to handle Italian inputs, while the decoder 'forgets' how to generate Italian outputs. This also explains the NLLB-pivot low performance in $lld \rightarrow it$, but relatively high scores in $it \rightarrow lld$.

The problem of 'forgetting' can be mitigated by using English-Italian sentence pairs during fine-tuning.

6. Limitations

A major limitation of this work consists in the little amount of data used for fine-tuning, and its lack of linguistic variety (most of the sentences are drawn from laws). This has a considerable impact on our MT model, which struggles on out-of-domain translations.

In general, as suggested by Ramponi [8], it would be important to assess the needs of the local community, in order to focus the efforts towards the most useful domains of application.

7. Conclusions

In this work, we show that it is possible to add a specific language variety to a pre-trained MT model using little amount of data for fine-tuning (fewer than 900 parallel sentences). To add Fassa Ladin, we fine-tune the model using as starting point a similar language included in NLLB-200: Friulian.

This approach significantly improves the performance. Moreover, in such condition, fine-tuning with parallel sentences in more than two languages proves to help regularization and to improve translations, with respect to a *zero-shot* pivot-based transfer learning approach.

Future work includes extending the dataset with new resources and domains, improving the alignment quality, and including human evaluation of translation quality. Adding data from other Ladin varieties might be a viable solution to improve the low performance caused by unknown words. Moreover, experimenting with translated words from vocabulary entries could be beneficial for Fassa Ladin, a language variety that has scarce parallel data but various publicly accessible vocabularies.

References

[1] C. R. Haberland, J. Maillard, S. Lusito, Italian-Ligurian machine translation in its cultural context, in: M. Melero, S. Sakti, C. Soria (Eds.), Proceedings of the 3rd Annual Meeting of the Special Interest Group on Under-resourced Languages @ LREC-COLING 2024, ELRA and ICCL, Torino, Italia, 2024, pp. 168–176. URL: https://aclanthology.org/2024.sigul-1.21.

[2] D. Adelani, J. Alabi, A. Fan, J. Kreutzer, X. Shen, M. Reid, D. Ruiter, D. Klakow, P. Nabende, E. Chang, T. Gwadabe, F. Sackey, B. F. P. Dossou, C. Emezue, C. Leong, M. Beukman, S. Muhammad, G. Jarso, O. Yousuf, A. Niyongabo Rubungo, G. Hacheme, E. P. Wairagala, M. U. Nasir, B. Ajibade, T. Ajayi, Y. Gitau, J. Abbott, M. Ahmed, M. Ochieng, A. Aremu, P. Ogayo, J. Mukiibi, F. Ouoba Kabore, G. Kalipe, D. Mbaye, A. A. Tapo, V. Memdjokam Koagne, E. Munkoh-Buabeng, V. Wagner, I. Abdulmumin, A. Awokoya, H. Buzaaba, B. Sibanda, A. Bukula, S. Manthalu, A few thousand translations go a long way! Leveraging pre-trained models for African news translation, in: M. Carpuat, M.-C. de Marneffe, I. V. Meza Ruiz (Eds.), Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Seattle, United States, 2022, pp. 3053–3070. URL: https://aclanthology.org/2022.naacl-main.223. doi:10.18653/v1/2022.naacl-main.223.

[3] S. Lankford, H. Afli, A. Way, adaptMLLM: Fine-tuning multilingual language models on low-resource languages with integrated LLM playgrounds, Information 14 (2023). URL: https://www.mdpi.com/2078-2489/14/12/638. doi:10.3390/info14120638.

[4] NLLB Team, M. R. Costa-jussà, J. Cross, O. Çelebi, M. Elbayad, K. Heafield, K. Heffernan, E. Kalbassi, J. Lam, D. Licht, J. Maillard, A. Sun, S. Wang, G. Wenzek, A. Youngblood, B. Akula, L. Barrault, G. M. Gonzalez, P. Hansanti, J. Hoffman, S. Jarrett, K. R. Sadagopan, D. Rowe, S. Spruit, C. Tran, P. Andrews, N. F. Ayan, S. Bhosale, S. Edunov, A. Fan, C. Gao, V. Goswami, F. Guzmán, P. Koehn, A. Mourachko, C. Ropers, S. Saleem, H. Schwenk, J. Wang, No language left behind: Scaling human-centered machine translation, 2022. arXiv:2207.04672.

[5] J. Tiedemann, OPUS – parallel corpora for everyone, in: Proceedings of the 19th Annual Conference of the European Association for Machine Translation: Projects/Products, Baltic Journal of Modern Computing, Riga, Latvia, 2016. URL: https://aclanthology.org/2016.eamt-2.8.

[6] S. Frontull, Machine translation for the low-resource Ladin of the Val Badia (2022).

[7] A. Ramponi, C. Casula, DiatopIt: A corpus of social media posts for the study of diatopic language variation in Italy, in: Y. Scherrer, T. Jauhiainen, N. Ljubešić, P. Nakov, J. Tiedemann, M. Zampieri (Eds.), Tenth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2023), Association for Computational Linguistics, Dubrovnik, Croatia, 2023, pp. 187–199. URL: https://aclanthology.org/2023.vardial-1.19. doi:10.18653/v1/2023.vardial-1.19.

[8] A. Ramponi, Language varieties of Italy: Technology challenges and opportunities, Transactions of the Association for Computational Linguistics 11 (2024) 19–38. URL: https://aclanthology.org/2024.tacl-1.2. doi:10.1162/tacl_a_00631.

[9] J. Casalicchio, Ladinia dolomitica, in: T. Krefeld, R. Bauer (Eds.), Lo spazio comunicativo dell'Italia e delle varietà italiane, München, 2020.

[10] E. Galante, C. Soraperra, M. Neri, Amer volesse, Stile Libero, 2006.

[11] J. Tiedemann, S. Thottingal, OPUS-MT – building open translation services for the world, in: Proceedings of the 22nd Annual Conference of the European Association for Machine Translation, European Association for Machine Translation, Lisboa, Portugal, 2020, pp. 479–480. URL: https://aclanthology.org/2020.eamt-1.61.

[12] A. Fan, S. Bhosale, H. Schwenk, Z. Ma, A. El-Kishky, S. Goyal, M. Baines, O. Celebi, G. Wenzek, V. Chaudhary, N. Goyal, T. Birch, V. Liptchinsky, S. Edunov, E. Grave, M. Auli, A. Joulin, Beyond English-centric multilingual machine translation, 2020. arXiv:2010.11125.

[13] K. Papineni, S. Roukos, T. Ward, W.-J. Zhu, Bleu: a method for automatic evaluation of machine translation, in: P. Isabelle, E. Charniak, D. Lin (Eds.), Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Philadelphia, Pennsylvania, USA, 2002, pp. 311–318. URL: https://aclanthology.org/P02-1040. doi:10.3115/1073083.1073135.

[14] M. Popović, chrF++: words helping character n-grams, in: Proceedings of the Second Conference on Machine Translation, Association for Computational Linguistics, Copenhagen, Denmark, 2017, pp. 612–618. URL: https://aclanthology.org/W17-4770. doi:10.18653/v1/W17-4770.

[15] T. Zhang*, V. Kishore*, F. Wu*, K. Q. Weinberger, Y. Artzi, Bertscore: Evaluating text generation with BERT, in: International Conference on Learning Representations, 2020. URL: https://openreview.net/forum?id=SkeHuCVFDr.

[16] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: J. Burstein, C. Doran, T. Solorio (Eds.), Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186. URL: https://aclanthology.org/N19-1423. doi:10.18653/v1/N19-1423.

[17] R. Agarwal, A. Singh, L. M. Zhang, B. Bohnet, L. Rosias, S. Chan, B. Zhang, A. Anand, Z. Abbas, A. Nova, J. D. Co-Reyes, E. Chu, F. Behbahani, A. Faust, H. Larochelle, Many-shot in-context learning, 2024. URL: https://arxiv.org/abs/2404.11018. arXiv:2404.11018.

[18] Y. Kim, P. Petrov, P. Petrushkov, S. Khadivi, H. Ney, Pivot-based transfer learning for neural machine translation between non-English languages, in: K. Inui, J. Jiang, V. Ng, X. Wan (Eds.), Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Association for Computational Linguistics, Hong Kong, China, 2019, pp. 866–876. URL: https://aclanthology.org/D19-1080. doi:10.18653/v1/D19-1080.

[19] Y. Luo, Z. Yang, F. Meng, Y. Li, J. Zhou, Y. Zhang, An empirical study of catastrophic forgetting in large language models during continual fine-tuning, 2024. URL: https://arxiv.org/abs/2308.08747. arXiv:2308.08747.

[20] I. J. Goodfellow, M. Mirza, D. Xiao, A. Courville, Y. Bengio, An empirical investigation of catastrophic forgetting in gradient-based neural networks, 2015. URL: https://arxiv.org/abs/1312.6211. arXiv:1312.6211.

[21] N. Shazeer, M. Stern, Adafactor: Adaptive learning rates with sublinear memory cost, in: J. Dy, A. Krause (Eds.), Proceedings of the 35th International Conference on Machine Learning, volume 80 of *Proceedings of Machine Learning Research*, PMLR, 2018, pp. 4596–4604. URL: https://proceedings.mlr.press/v80/shazeer18a.html.

Dataset	Italian	Ladin
Wikipedia	Sono usciti complessivamente tre numeri. *A total of three issues were released.*	Ie la prima plata ladina[1]. *It's the first ladin page[1].*
QED	E gli uomini delà , Meli esponilo Holly mise San , in estat' teston' *And the men delà , Meli expose it Holly put San , in estat' teston' (sic)*	Si te serf demò la lum canche la se n va , te mencia l soreie demò canche l taca a fiochèr *If you only need light when it goes out , you only miss the sun when it starts snowing*

Table 8
Two examples of non-aligned sentences from the OPUS corpora. English translations for **it** and **lld** are provided in *italic*.

A. Previous Ladin corpora

Three datasets from the OPUS corpora, namely Wikipedia, QED, and Ubuntu, contain parallel Ladin-Italian data. Unfortunately, none of these provide information about the Language variety of the sentences (e.g., the ones mentioned in Section 2). Some of them also present non-aligned sentences (see examples in Table 8).

B. Prompt for gpt-4o

Figure 2 shows the prompt used for the translation task with gpt-4o, presented in Section 4.

```
###INTRODUCTION###
You are a expert translator specialized in
low-resource languages and dialects.
Your core competence is bidirectional translation
between italian (IT), english (EN), and fassa
ladin (LLD) languages.

###INSTRUCTIONS###
You will be provided with information on the
source language (SOURCE_LANG), a textual input
(SOURCE_TEXT), and a target language (TARGET_LANG).
Your task is to accurately translate SOURCE_TEXT
from language SOURCE_LANG to language TARGET_LANG,
producing TARGET_TEXT.
Your output is a JSON file with exactly the
following schema:
{
"SOURCE_LANG": str, \\the value of SOURCE_LANG.
"TARGET_LANG": str, \\the value of TARGET_LANG.
"TARGET_TEXT": str, \\the translation output.
}
```

Figure 2: Prompt used for gpt-4o.

C. Implementation details

All experiments were conducted on Google Colab[16] using a single NVIDIA T4 15GB GPU; the fine-tuning process required approximately 1 hour.

We fine-tune the NLLB-200's distilled 600M variant[17] using the Adafactor optimizer [21], with a learning rate of $1.5 \cdot 10^{-4}$ and 500 warm-up iterations.[18] We use a batch size of 16 sentences.

[16] https://colab.google.com

[17] https://huggingface.co/facebook/nllb-200-distilled-600M

[18] https://github.com/adaptNMT/adaptMLLM