Italian-Ligurian Machine Translation in its Cultural Context

Christopher Haberland[◇], **Stefano Lusito**[♣], **Jean Maillard**[♥]
[◇] University of Washington [♣] University of Innsbruck [♥] Council for Ligurian Linguistic Heritage
info@conseggio-ligure.org

Abstract

Large multilingual machine translation efforts are driving improved access and performance for under-resourced languages, but often fail to translate culturally specific and local concepts. Additionally, translation performance from practically relevant input languages may lag behind that of languages that are comparatively over-represented in the training dataset. In this work, we release a new corpus, ZenaMT, containing 7,561 parallel Ligurian-Italian sentences, nearly a fifth of which are also translated in English. This corpus spans five domains: local and international news, Ligurian literature, Genoese Ligurian linguistics concepts, traditional card game rules, and Ligurian geographic expressions. We find that a translation model augmented with ZenaMT improves a baseline by 20%, and by over 25% (BLEU) compared to NLLB-3.3B, which is over 50 times the size. Our results demonstrate the utility of creating data sets for MT that are tailored for local cultural contexts by target language speakers. We freely release ZenaMT and expect to periodically update the corpus to improve MT performance and domain coverage.

Keywords: machine translation, Ligurian, Genoese, low-resource

1. Introduction

Large multilingual translation models from well-resourced tech companies (NLLB Team et al., 2022; Bapna et al., 2022; Siddhant et al., 2022) have included a much greater number of languages compared to prior model releases. For many communities, these models often represent a form of digital recognition of their heritage language and may even attain high translation performance. However, the training data for under-resourced languages fed as input to these large multilingual releases does not always include culturally relevant language data (Buscaldi and Rosso, 2023; Ramponi, 2024), or lacks a sufficiently strong parallel signal between language pairs that are crucial for the target language community. The datasets compiled by these centralized efforts can be insufficient to achieve high performance for localized translation contexts that are encountered by communities of under-resourced and minority languages. In this work, we document how intentional collation of a parallel dataset with participation and direction from the target language community improves culturally pertinent machine translation performance for Genoese Ligurian.

2. Background

2.1. Linguistic Background

Genoese Ligurian is a Romance variety[1] originating from Liguria, a coastal region in northwestern Italy. Genoese is the prestige variety of Ligurian (Forner, 1988; Petracco Sicardi, 1995; Toso, 2002), a group of mutually intelligible varieties that evolved from Latin independently from Italian (Toso, 1995, pp. 29-46).

Genoese is spoken today mainly in the central part of Liguria, in an area roughly between Noli and Moneglia on the coast and much of its hinterland (Toso, 1992). However, several sites outside this area are still oriented towards Genoese, and this variety is understood almost universally by other Ligurian speakers. Other Ligurian varieties are spoken in Monaco (Arveiller, 1967), where Monégasque is considered the principality's national language (Frolla, 1977), in Carloforte and Calasetta in Sardinia (Toso, 2003, 2004), where it is still used by the vast majority of pre-school-aged children (Sitztia, 1998, pp. 53-81; Spiga, 2007, pp. 69-74), and in Bonifacio in Corsica (Comiti and Di Meglio, 2021). In the past, Ligurian communities spread throughout the Mediterranean and Black Sea via Genoese maritime commercial enterprises (Toso, 2020).

Thanks to its uninterrupted written usage from the 13th century to the present day, Genoese graphemic sequences correspond to phonemes in a different way than those of neighboring languages, such as Italian (Toso, 2009b). However, Ligurian is not recognized under Italian law and is not officially standardized, remaining largely absent from the educational environment.[2] For these reasons, Genoese lacks a regulated spelling system, and "spontaneous spellings" (Iannàccaro and Dell'Aquila, 2008) are common in the Ligurian lin-

[1] We use the term 'varieties' to bridge different communities' reference systems for linguistic entities, following Ramponi (2024).

[2] The only notable exception is Monégasque, taught in schools since the 1970s (Stefanelli, 2000; Lusito, 2022b).

SIGUL2024 Workshop, pages 168–176

guistic landscape and on social networks. These writings largely emerge in informal settings, draw upon Italian spelling rules, and exhibit a high degree of variability. This situation is shared by many other Romance languages spoken in Italy without institutional prerogatives, such as Lombard (Miola, 2015), Neapolitan (Leoni, 2015) or Piedmontese (Miola, 2021).

The Genoese data we present in this work are written in a codified form of the traditional spelling (Acquarone, 2015b; Lusito, 2022c; Maillard et al., 2023b), itself a simplification of the rules proposed by Toso (1997, pp. 25-46). This spelling model represents the *de facto* standard for news media – such as the weekly page in Genoese in the main daily newspaper of Liguria (Acquarone, 2015a) – as well as for literary (Toso, 2015–2019; Acquarone, 2018–present; Roveda, 2023–present), didactic (Lusito, 2022a), and academic work (Toso, 2015; Guasoni, 2019; Autelli et al., 2019; Lusito, 2023; Lusito et al., 2023; Toso, 2023; Jones et al., 2023). Other orthographic standards have also been proposed by language enthusiasts, such as those offered by Petrucci (1984), Costa (1993), Gambetta (2009), and Durante (2014), yet these proposals exhibit varying degrees of completeness and specificity, presenting challenges for their uniform application across all Ligurian linguistic varieties. The system proposed by Bampi (2009) attempts to closely align the written form to its pronunciation. Although this strategy captures nuanced variations in pronunciation, it inherently leads to a diverse array of spellings for the same word, reflecting individual speech patterns and judgments. Consequently, this system results in a spectrum of spellings rather than a single, standard orthography.

2.2. Related Work

The first translation system for Ligurian (targeting Genoese, like the present work) was NLLB (NLLB Team et al., 2022), coinciding with the release of the evaluation benchmark FLORES-200 and some seed training datasets, which also covered Ligurian. We make use of both these datasets in our work. In a follow-up paper, Maillard et al. (2023a) train a translation model covering several languages of Italy, and show the effectiveness of the seed training dataset in bootstrapping machine translation (MT) systems.

Buscaldi and Rosso (2023) analyze the performance of NLLB and find that it performs poorly on a test set built from texts that are culturally relevant to Ligurian speakers. They identify two key issues with previous work on Ligurian MT. First, NLLB Ligurian training data is only present in the form of English-Ligurian aligned text, even though most Ligurian speakers are likely to prefer translating from and into Italian. Second, most of the training data is translated content sampled from English Wikipedia, a corpus that omits concepts of special relevance to Ligurian speakers. The present work most closely aligns with Buscaldi and Rosso's in acknowledging the importance of culturally-relevant, Italian-Ligurian training and evaluation data, and aims to make progress towards the issues they highlight.

Our work is among several recent efforts to build MT and NLP tools for linguistic varieties of Italy. We refer readers to Ramponi (2024) for an overview of recent language technology tools that have been built for minority linguistic varieties in Italy.

3. Ligurian Machine Translation

Despite the marginalization of Ligurian in most spheres of society, the Ligurian speaking community demands translation tools. This is evinced by the numerous comments soliciting translation assistance that are frequently posted to social media sites, which have emerged as primary spaces for asserting linguistic agency for members of minority language communities, where hybrid language usage is often encouraged (Belmar and Glass, 2019).[3] One of the authors who manages the website for the Council for Ligurian Linguistic Heritage[4] reports that the vast majority of traffic arrives via Google after searching for a "Ligurian translator" (as reported by Google Search Console). The group receives regular emails soliciting translation consultation between Italian and Ligurian.

All of the models we train are Italian to Ligurian bilingual translation systems, trained exclusively on Italian-Ligurian parallel data. Our choice to focus on translation from Italian to Ligurian reflects preferences expressed by the community. Our decision to not train a large multilingual system, using, for example, English-aligned data, is based on a desire to concentrate on smaller, more efficient models that could more easily be trained and deployed by language community members on widely available and cheaper infrastructure.

In developing our machine translation system, we deliberately only train on data written in the traditional codified Genoese orthography described in §2.1. This decision stems from the fact that mixing orthographies would affect the spelling of nearly every word in Genoese, which would render the model incapable of learning by introducing irreconcilable linguistic inconsistencies during the train-

[3] We found several requests for translation tools in popular Ligurian Facebook groups *Gruppo de discuscion in sciâ lengua zeneise* and *Amici del dialetto ligure*.

[4] *Conseggio pe-o patrimònio linguistico ligure*, a nonprofit association for the promotion of Ligurian: https://conseggio-ligure.org.

Subset	Ligurian Sentence	English Gloss
linguistics	A-o comenso ò pensou ch'o voeiva ingannâme, ma dapeu me son dæto conto ch'o l'ea scinçeo	At first I thought he wanted to trick me, but then I realized he was sincere.
news	L'inflaçion a chiña ma, segondo i economisti, a l'arrestià ancon tròppo erta pe tròppo tempo.	Inflation is falling but, according to economists, it will remain too high for too long.
literature	O l'à fondou o Comitao de Tradiçioe Monegasche e do 1927 o l'à pubricou A legenda de Santa Devota, poemma naçionale monegasco.	He founded the Committee of Monégasque Traditions and in 1927 he published A legenda de Santa Devota, the Monégasque national poem.
games	A biscambiggia inta trei a l'é squæxi do tutto pægia a-o zeugo inta doî.	Three-handed biscambiggia is almost identical to the two-handed game.
entities	Begæ o dà o nomme à un di fòrti de Zena.	Begato gives its name to one of the forts of Genoa.

Table 1: Example sentences and translations in ZenaMT by data subset.

ing phase. Mixing spellings is also inadvisable for target-side evaluation, as even a perfect translation model would be presented with the impossible task of guessing, for each token, the correct spelling variation to use in a particular test sentence. A high degree of spelling variation is observed, for instance, in the dataset by Buscaldi and Rosso (2023), where even common function words are affected by irregular and unpredictable variations.[5] Therefore, when using this dataset in this work, we normalize its spelling manually.

We emphasize that our work is inclusive of the community for which it benefits, in line with calls for "participatory AI" (Birhane et al., 2022). In this regard, our work is inspired by other participatory machine translation initiatives for local language communities, such as Masakhane (Nekoto et al., 2020). By tailoring training data for the Genoese Ligurian-speaking community by including culturally relevant data, or data on domains that are useful to the community, we aim to test the performance of dependent machine translation systems for domains that are likely to be of greater importance to actual users. We also solicit data submissions by active community members themselves. We expect that improved machine translation in domains more pertinent to the Ligurian community will increase the relevance of MT as a tool not only for adapting content for Ligurian speakers, but for helping less confident speakers to practice and learn the language. For these reasons, we see a participatory approach in collecting data and developing solutions for the Ligurian community as vital to support the goal of linguistic revitalization.

[5]We note for instance, the presence of conflicting spellings for the Genoese preposition *into* ("in the"), which is also variously written as *'ntou*, *'nt'u* and *'nt'ou* in an unpredictable way.

3.1. Corpus Construction

We compile a corpus of Italian-Ligurian parallel sentences across 5 subsets according to domain. Ligurian training examples are shown in Table 1. The authors consulted with Ligurian community members affiliated with the Council for Ligurian Linguistic Heritage to identify domains that would balance domain diversity and linguistic representation, and would minimize the cost imposed by the data collection process. A **linguistics** subset is comprised of 1,066 sentences that are drawn from the interactive Genoese Ligurian dictionary published on the official website of the Council for Ligurian Linguistic Heritage. **News** is drawn from the weekly online newspaper *O Zinâ*.[6] The **literature** subset is drawn from the published anthology of Ligurian literature by Guasoni (2023–present). A **games** subset contains parallel sentences from a website documenting the rules of several traditional Ligurian card games.[7] Finally, geographic **entities** are compiled in a separate subset comprised of sentences pertaining to regional toponyms (mapped in Figure 1). With the exception of a small fraction of sentences from the **literature** subset, all ZenaMT sentences were originally written in Ligurian and translated to Italian by native speakers. The size of train, validation, and test splits for all corpus subsets are shown in Table 2. Validation and test splits were made only for the **news**, **literature**, and **entities** splits to reflect fairer evaluations by not privileging models trained on specialized domains (such as the **linguistics** and **games** subset domains).

[6]https://ozina.org/

[7]https://www.sbiro.eu/

Corpus	Languages	Train	Valid	Test
linguistics	lij, ita	3,497		
news	lij, ita	1,884	130	264
literature	lij, ita, eng	724	135	207
games	lij, ita, eng	297		
entities	lij, ita	282	70	71
Total		*6,684*	*335*	*542*

Table 2: Number of parallel sentences by subset, set of languages, and data split of the newly contributed ZenaMT corpus.

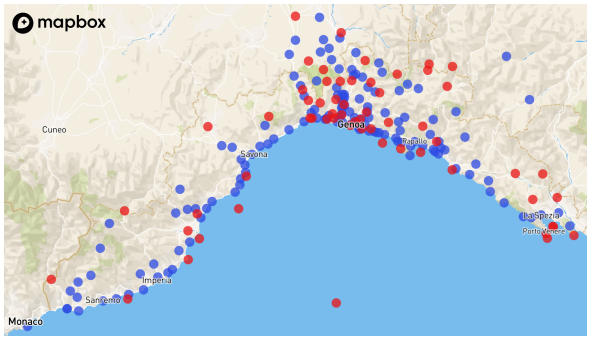

Figure 1: Geocoded toponyms from the **entities** subset of ZenaMT. Red points represent natural geographic features, blue points represent urban features. © Mapbox, © OpenStreetMap.

3.2. Experimental Setup

We conduct our experiments on a Google Colab notebook backed by a single NVIDIA V100 16GB GPU. We use Sentencepiece (Kudo and Richardson, 2018) to train a single unigram language model tokenizer (Kudo, 2018) with a vocabulary size of 1k tokens for both Italian and Ligurian.

The translation models are trained using Fairseq (Ott et al., 2019), and use an encoder/decoder transformer architecture (Vaswani et al., 2017) with 6 encoder and 6 decoder layers, 512 hidden size and 8 attention heads, equating to roughly 65 million parameters. We train with a batch size of 16,384 tokens using the AdamW optimizer (Loshchilov and Hutter, 2019), with 1000 warmup iterations, inverse square root decay, a maximum learning rate of 0.001 and 0.5 dropout. Models are trained until convergence as determined by BLEU score (Papineni et al., 2002) on the combined FLORES and ZenaMT validation sets.

We train a **Baseline** system with the aim of measuring achievable performance with data that had been available before our corpus collection efforts. Namely, we use 1,520 Italian-Ligurian parallel sentences from the Tatoeba project[8] and 6,193 Italian-Ligurian parallel sentences, which we obtain by machine-translating the English NLLB seed data (Maillard et al., 2023a) to Italian with OPUS-MT (Tiedemann and Thottingal, 2020)[9] and manually post-editing it. We evaluate on the ZenaMT test set and on the FLORES-200 devtest set. We also evaluate on the test set by Buscaldi and Rosso (2023), which we normalize to our target orthography to avoid the issues described in §3. Data statistics for these corpora are available in Table 3.

Corpus	Train	Valid	Test
Seed	6,193		
Tatoeba	1,520		
FLORES		997	1,012
Norm. B&R			283

Table 3: Additional Italian-Ligurian translation datasets beyond ZenaMT used in the **Baseline** and **New** experiments.

Our **New** system is trained on the above data, with the addition of ZenaMT, described in §3.1.

3.3. Results

Test Set	NLLB-3.3B	Baseline	New
FLORES	13.9 / 40.6	14.5 / 42.9	17.4 / 45.8
Norm. B&R	9.9 / 35.4	10.3 / 37.6	16.0 / 43.3
ZenaMT	24.0 / 51.9	25.4 / 53.6	47.9 / 69.7

Table 4: Italian-Ligurian translation performance of our models and NLLB-3.3B measured with BLEU (Papineni et al., 2002) and chrF++ (Popović, 2017).[11]

Table 4 shows translation performance for our two sub-100M-parameter models and the 3.3B-parameter version of NLLB. We investigate the Italian to Ligurian translation direction, since this is by far the most requested by the community.

The first trend to emerge is the impact of training on Italian-Ligurian data. Compared to our two models, NLLB is a much larger, massively multilingual model, trained on far more text. It does however lack direct Italian-Ligurian data, and despite the benefits of cross-lingual transfer, we see that it is already outperformed by our baseline model.

Second, our model trained on the additional ZenaMT data achieves a clear boost in translation

[8] https://tatoeba.org/, retrieved 2024-02-05.

[9] https://huggingface.co/Helsinki-NLP/opus-mt-en-it/, accessed January 2024.

[11] SacreBLEU (Post, 2018) signatures `nrefs:1|case:mixed|eff:no|tok:13a|smooth:exp|version:2.4.0` and `nrefs:1|case:mixed|eff:yes|nc:6|nw:2|space:no|version:2.4.0`.

performance across all three test sets, attesting to its effectiveness. Unsurprisingly, we see a much larger increase in performance compared to the baseline on the ZenaMT test set, as it has been drawn from the same sources that make up the additional training data.

Finally, we see that performance on the FLORES and normalized Buscaldi and Rosso (2023) test sets are much lower compared to the ZenaMT test set. This can likely be attributed to the origins of these datasets. While ZenaMT is largely Ligurian-original text, both the Italian and Ligurian versions of FLORES were translated from English, so the effects of *translationese* (Riley et al., 2020) are likely impacting both sides. The Buscaldi and Rosso test set, while culturally relevant to Ligurian contexts, does also suffer from some of the same issues, as the majority of the data (over 80% by character count) comes from the writings of Charles Dickens, originally written in English, translated into Ligurian, and then machine-translated into Italian. Some of the remaining data are lyrics of celebrated singer-songwriter Fabrizio De André, which, although originally written in Ligurian, are known to be unrepresentative of general language use (Toso, 2009a).

4. Conclusions

We have described the construction of ZenaMT, a parallel Italian-Ligurian corpus for training machine translation models.[12] Its over 7,000 sentences were collected from sources which are culturally relevant to Ligurian speakers. We train an Italian to Ligurian translation model by combining this data and existing corpora (including a newly derived Italian-Ligurian seed corpus based on data provided by the NLLB project). Our model consists of fewer than 100M parameters but outperforms the 3.3B-parameter NLLB model on multiple benchmarks, attesting to the importance of using Italian-Ligurian, culturally-relevant data. Our approach exemplifies the downstream performance benefits and increased relevance of digital translation tools that are achievable through intentional dataset creation in partnership with a target minority language community.

ZenaMT constitutes a living corpus compiled with direct participation from the Ligurian speaking community that we intend to update periodically to improve domain and language coverage, as well as translation performance. We hope to significantly expand it in the future with more news coverage, weather forecasts, and sentences that include other named entities such as international toponyms, local geographic features, and important figures.

5. Acknowledgements

We extend our heartfelt gratitude to those who have generously contributed to this project. Our thanks go to Fabio Canessa, for his contribution of news articles; Alessandro Guasoni, for sharing his Anthology of Ligurian Literature (Guasoni, 2023–present); and Claudio Rezzoagli, for his invaluable assistance in translating named entities. Their support not only enriched our project but also enhanced the quality of our evaluation.

6. Ethical Considerations and Limitations

Our work focuses on traditional Genoese orthography. Some Ligurian speakers may prefer alternative spelling systems. A similar concern was elicited by Haroutunian (2022) from a panel of speakers of Armenian, a language with multiple orthographic conventions, who saw harm in one orthographic alternative potentially supplanting another via the standardizing effect of a proliferated machine translation system. In cases where Ligurian is an input language – such as for Ligurian to Italian MT – robustness to spelling variation could be achieved via data augmentation strategies using approaches similar to the one described by Karpukhin et al. (2019). As discussed in §3, using multiple spelling systems of Ligurian for the target output data presents a different set of challenges, since doing so in a single model would introduce inconsistencies in the training signal. One solution could involve training completely separate models for different spelling systems, therefore treating them as if they were separate languages. A better solution could make use of a text adaptation layer as a post-processing step, since effective transliteration models have already been demonstrated in prior work (Lusito et al., 2023). The value of our work can therefore be realized by proponents of any spelling system.

Finally, we note that Ligurian and Italian are both members of the Romance language family, and consequently, translation between these two languages is generally easier than between more distant language pairs. The relatively high translation performance we were able to achieve in this study in spite of the small size of our training datasets would likely not be reproducible for arbitrary translation directions.

[12] We make this data available under CC BY-4.0 at https://github.com/ConseggioLigure/data/. The models described in this paper were trained on the version of the data at commit hash 52ed7b6

7. Bibliographical References

Andrea Acquarone. 2015a. Creusa o creuza? Ecco come si scrive in lingua genovese. *Il Secolo XIX,* Nov 6, 2015, page 31.

Andrea Acquarone. 2015b. Scrivere la lingua. In Andrea Acquarone, editor, *Parlo Ciæo. La lingua della Liguria*, pages 87–94. De Ferrari and Il Secolo XIX, Genova, Italy.

Andrea Acquarone, editor. 2018–present. *Biblioteca zeneise*. De Ferrari, Genova, Italy.

Raymond Arveiller. 1967. *Étude sur le parler de Monaco*. Comité national des traditions monégasques, Monaco.

Erica Autelli, Konecny Christine, and Stefano Lusito. 2019. GEPHRAS: il primo dizionario combinatorio genovese-italiano online. In Fiorenzo Toso, editor, *Il patrimonio linguistico storico della Liguria: attualità e futuro. Raccolta di Studi*. InSedicesimo, Savona, Italy.

Franco Bampi. 2009. *Grafîa ofiçiâ*. S.E.S., Genova, Italy.

Ankur Bapna, Isaac Caswell, Julia Kreutzer, Orhan Firat, Daan van Esch, Aditya Siddhant, Mengmeng Niu, Pallavi Baljekar, Xavier Garcia, Wolfgang Macherey, et al. 2022. Building machine translation systems for the next thousand languages. *arXiv preprint arXiv:2205.03983*.

Guillem Belmar and Maggie Glass. 2019. Virtual communities as breathing spaces for minority languages: Re-framing minority language use in social media. *Adeptus*, (14).

Abeba Birhane, William Isaac, Vinodkumar Prabhakaran, Mark Diaz, Madeleine Clare Elish, Iason Gabriel, and Shakir Mohamed. 2022. Power to the people? Opportunities and challenges for participatory AI. *Equity and Access in Algorithms, Mechanisms, and Optimization*, pages 1–8.

Davide Buscaldi and Paolo Rosso. 2023. How good is NLLB-200 for low-resource languages? A study on Genoese. In *CLiC-it 2023: 9th Italian Conference on Computational Linguistics*.

Jean-Marie Comiti and Alain Di Meglio. 2021. Le bonifacien, un isolat linguistique ligure en Corse. In Claude Passet, editor, *Gênes et la langue génoise: expression de la terre et de la mer, langue d'ici et langue d'ailleurs*, pages 499–513. Éditions ECG / Académie des langues dialectales, Monaco.

Carlo Costa. 1993. *Grammatica del genovese*. Tigullio-Bacherontius, Santa Margherita, Italy.

Nino Durante. 2014. *Grammatica genovese curiosa e intrigante. Grafia tradizionale. Proverbi, frasi celebri, modi di dire*. ERGA, Genova, Italy.

Werner Forner. 1988. Italienisch: Areallinguistik I. Ligurien. In Christian Schmitt Günter Holtus, Michael Metzeltin, editor, *Italienisch, Korsisch, Sardisch*, volume IV of *Lexicon der Romanistischen Linguistik*, pages 453–469. Max Niemeyer Verlag, Tübingen.

Louis Frolla. 1977. Monaco. Son idiome national. In *Annales monégasques*, pages 67–77. Publication des archives du Palais Princier, Monaco.

Enrico Gambetta. 2009. *Piccola grammatica del genovese*. ERGA, Genova, Italy.

Alessandro Guasoni. 2019. *Poesia in ligure fra Novecento e Duemila*. Cofine, Roma, Italy.

Alessandro Guasoni. 2023–present. Antologia da lettiatua ligure. Council for Ligurian Linguistic Heritage.

Levon Haroutunian. 2022. Ethical considerations for low-resourced machine translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 44–54, Dublin, Ireland. Association for Computational Linguistics.

Gabriele Iannàccaro and Vittorio Dell'Aquila. 2008. Per una tipologia dei sistemi di scrittura spontanei in area romanza. *Estudis Romànics*, 30:311–331.

Alex Jones, Isaac Caswell, Ishank Saxena, and Orhan Firat. 2023. Bilex Rx: Lexical data augmentation for massively multilingual machine translation. *arXiv:2303.15265*.

Vladimir Karpukhin, Omer Levy, Jacob Eisenstein, and Marjan Ghazvininejad. 2019. Training on synthetic noise improves robustness to natural noise in machine translation. In *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, pages 42–47, Hong Kong, China. Association for Computational Linguistics.

Taku Kudo. 2018. Subword regularization: Improving neural network translation models with multiple subword candidates. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 66–75, Melbourne, Australia. Association for Computational Linguistics.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Federico Albano Leoni. 2015. Carmniell o' srngar. Osservazioni sulla ortografia selvaggia del napoletano. In *Elaborazione ortografica delle varietà non standard*, Esperienze spontanee in Italia e all'estero, pages 51–78. Bergamo University Press / Sestante Edizioni, Bergamo.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Stefano Lusito. 2022a. *Dizionario italiano-genovese. O diçionäio ch'o mostra o zeneise d'ancheu*. Programma, Treviso, Italy.

Stefano Lusito. 2022b. L'insegnamento scolastico del monegasco dagli esordi al panorama attuale: presenza nei programmi di istruzione, metodologie pedagogiche, strumenti didattici e aspetti linguistici. volume 46 of *Bollettino dell'Atlante linguistico italiano*, pages 181–213. Istituto dell'Atlante Linguistico Italiano, Torino, Italy.

Stefano Lusito. 2022c. Prefaçion. In *Dizionario italiano-genovese. O diçionäio ch'o mostra o zeneise d'ancheu*, pages 14–15. Editoriale Programma, Treviso, Italy.

Stefano Lusito. 2023. *Stefano De Franchi. Ro mêgo per força*, Zimme de braxa, chapter Glossario. Zona, Genoa, Italy.

Stefano Lusito, Edoardo Ferrante, and Jean Maillard. 2023. Text normalization for low-resource languages: the case of Ligurian. In *Proceedings of the Sixth Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 98–103. Association for Computational Linguistics.

Jean Maillard, Cynthia Gao, Elahe Kalbassi, Kaushik Ram Sadagopan, Vedanuj Goswami, Philipp Koehn, Angela Fan, and Francisco Guzmán. 2023a. Small data, big impact: Leveraging minimal data for effective machine translation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2740–2756, Toronto, Canada. Association for Computational Linguistics.

Jean Maillard, Stefano Lusito, and Alessandro Guasoni. 2023b. Ligurian (Genoese) orthography.

Emanuele Miola. 2015. Chì pòdom tucc scriv come voeurom. Scrivere in lombardo online. In Iannàccaro G. Dal Negro S., Guerini F., editor, *Elaborazione ortografica delle varietà non standard. Esperienze spontanee in Italia e all' estero*, pages 79–96. Bergamo University Press / Sestante Edizioni, Bergamo.

Emanuele Miola. 2021. *Taking a Closer Look at Spontaneous Writing in Piedmontese*, Studies in World Language Problems, chapter 8, Contested Orthographies. John Benjamins Publishing Company.

Wilhelmina Nekoto, Vukosi Marivate, Tshinondiwa Matsila, Timi Fasubaa, Tajudeen Kolawole, Taiwo Fagbohungbe, Solomon Oluwole Akinola, Shamsuddeen Hassan Muhammad, Salomon Kabongo, Salomey Osei, et al. 2020. Participatory research for low-resourced machine translation: A case study in african languages. *arXiv preprint arXiv:2010.02353*.

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. *arXiv:1902.01382*.

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 the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53, Minneapolis, Minnesota. Association for Computational Linguistics.

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

Giulia Petracco Sicardi. 1995. Italienisch: Areallinguistik I. In Christian Schmitt Günter Holtus, Michael Metzeltin, editor, *Die einzelnen romanischen Sprachen und Sprachgebiete vom Mittelalter bis zur Renaissance*, volume II of *Lexicon der romanischen Sprachen*, pages 111–124. Max Niemeyer Verlag, Tübingen, Germany.

Vito Elio Petrucci. 1984. *Grammatica sgrammaticata della lingua genovese*. Sagep, Genova, Italy.

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.

Alan Ramponi. 2024. Language varieties of Italy: Technology challenges and opportunities. *Transactions of the Association for Computational Linguistics*, 12:19–38.

Parker Riley, Isaac Caswell, Markus Freitag, and David Grangier. 2020. Translationese as a language in "multilingual" NMT. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7737–7746, Online. Association for Computational Linguistics.

Anselmo Roveda, editor. 2023–present. *Zimme de braxa. Colleçion de lettiatua ligure*. Editrice Zona and Council for Ligurian Linguistic Heritage, Genova, Italy.

Aditya Siddhant, Ankur Bapna, Orhan Firat, Yuan Cao, Mia Xu Chen, Isaac Caswell, and Xavier Garcia. 2022. Towards the next 1000 languages in multilingual machine translation: Exploring the synergy between supervised and self-supervised learning. *arXiv preprint arXiv:2201.03110*.

Paola Sitztia. 1998. *Le comunità tabarchine della Sardegna meridionale: un'indagine sociolinguistica*. Condaghes, Cagliari, Italy.

Riccardo Spiga. 2007. I codici delle aree linguistiche. In *Le lingue della Sardegna. Una ricerca sociolinguistica*, pages 65–74. Regione Autonoma della Sardegna, Cagliari, Italy.

René Stefanelli. 2000. Le parler de Monaco à l'école. *Annales Monégasques*, 24:151–185.

Jörg Tiedemann and Santhosh Thottingal. 2020. OPUS-MT — Building open translation services for the World. In *Proceedings of the 22nd Annual Conferenec of the European Association for Machine Translation (EAMT)*, Lisbon, Portugal.

Fiorenzo Toso. 1992. Unità e varietà delle parlate liguri. Problemi di definizione areale e di classificazione sociolinguistica del genovese. In *Travaux du Cercle linguistique de Nice*, volume 13, pages 23–41.

Fiorenzo Toso. 1995. *Storia linguistica della Liguria. Vol. 1. Dalle origini al 1528*. Le Mani, Recco (Genova), Italy.

Fiorenzo Toso. 1997. *Grammatica del genovese. Varietà urbana e di koinè*. Le Mani, Recco, Italy.

Fiorenzo Toso. 2002. La Liguria. In Nicola De Blasi e Gianrenzo P. Clivio Manlio Cortelazzo, Carla Marcato, editor, *I dialetti italiani: storia, struttura, uso*, pages 196–225. UTET, Torino, Italy.

Fiorenzo Toso. 2003. *I tabarchini della Sardegna. Aspetti linguistici ed etnografici di una comunità ligure d'oltremare*. Le Mani, Recco.

Fiorenzo Toso. 2004. Il tabarchino. Strutture, evoluzione storica, aspetti sociolinguistici. In Augusto Carli, editor, *Il bilinguismo tra conservazione e minaccia. Esempi e presupposti per interventi di politica linguistica e di educazione bilingue*, pages 21–235. FrancoAngeli, Milano, Italy.

Fiorenzo Toso. 2009a. De Andrè, il genovese. *Insula Europea*.

Fiorenzo Toso. 2009b. *La letteratura ligure in genovese e nei dialetti locali*. Le Mani, Recco (Genova), Italy.

Fiorenzo Toso. 2015. *Piccolo dizionario etimologico ligure. L'origine, la storia e il significato di quattrocento parole a Genova e in Liguria*. Editrice Zona, Genova, Italy.

Fiorenzo Toso, editor. 2015–2019. *E restan forme*. Zona, Genova, Italy.

Fiorenzo Toso. 2020. *Il mondo grande. Rotte interlinguistiche e presenze comunitarie del genovese d'oltremare. Dal Mediterraneo al Mar Nero, dall'Atlantico al Pacifico*. Edizioni dell'Orso, Alessandria, Italy.

Fiorenzo Toso. 2023. *Desgel. Dizionario etimologico storico genovese e ligure. Volume di saggio. Lettera N*. Edizioni dell'Orso, Alessandria, Italy.

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.