ARBRES Kenstur: a Breton-French Parallel Corpus Rooted in Field Linguistics

Loïc Grobol, Mélanie Jouitteau

MoDyCo, CNRS,
and Université Paris Nanterre ;
Lattice, ENS, CNRS,
and Université Sorbonne Nouvelle
lgrobol@parisnanterre.fr

IKER, UMR 5478, CNRS,
Université de Pau et des Pays de l'Adour,
and Université Bordeaux-Montaigne
melanie.jouitteau@iker.cnrs.fr

Abstracts

ARBRES is an ongoing project of open science implemented as a platform ("wikigrammar") documenting both the Breton language itself and the state of research and engineering work in linguistics and NLP. Along its nearly 15 years of operation, it has aggregated a wealth of linguistic data in the form of interlinear glosses with translations illustrating lexical items, grammatical features, dialectal variations… While these glosses were primarily meant for human consumption, their volume and the regular format imposed by the wiki engine used for the website also make them suitable for machine processing.

ARBRES Kenstur is a new parallel corpus derived from the glosses in ARBRES, including about 5k phrases and sentences in Breton along with translations in standard French. The nature of the original data — sourced from field linguistic inquiries meant to document the structure of Breton — leads to a resource that is mechanically more concerned with the internal variations of the language and rare phenomena than typical parallel corpora. Preliminaries experiments in using this corpus show that it can help improve machine translation for Breton, demonstrating that sourcing data from field linguistic documentation can be a way to help provide NLP tools for minority and low-resource languages.

Keywords: parallel corpus, Breton, machine translation, field linguistics

1. Introduction

While advances in Natural Language Processing in recent years have made language technologies significantly more useful for the general public, only a small fraction of the languages of the world have benefited from them (Joshi et al., 2020; Simons et al., 2022). Even when putting aside interest and profitability factors, in a domain that is dominated by for-profit industrial actors, one salient and well-known issue is the extreme scarcity of available data for most languages. Given the overwhelming domination of data-driven techniques in current NLP, it is easy to see how so many languages can be left behind and how communities can end up without support of their languages in their day-to-day use of digital tools.

However, recent efforts such as those of the Masakhane collective (Adelani et al. (2022) *inter alia*) have shown that while data availability is still crucial, leveraging massively multilingual datasets and models can significantly reduce the volume needed to obtain useful downstream NLP tools. In particular, for machine translation, a few thousand parallel sentences (down from several millions in e.g. Barrault et al. (2019)) used to fine-tune a pretrained multilingual system yielded models with very encouraging performances. However, gathering even such small amount of data can prove challenging for minority languages.

Breton (*Brezhoneg*) is a Celtic language of the Brythonic family, spoken historically in the Brittany region of northwestern France. Like all minority languages of France, it has been steadily losing territory to French during the last century. Its current status is "severely endangered" according to UNESCO's Atlas of the world's languages in danger (Mosely and Nicolas, 2010) and Joshi et al. (2020) have it in their category 1 "the scraping-bys" in terms of available resources and academic interest. While some resources and tools do exist (Jouitteau and Bideault, 2023), they are not necessarily easily available or at a level of performance that let them be useful for the general public.

In this work, we describe our efforts to tap into an open grammar to provide resources for Breton-French machine translation. To this end, we leverage the ARBRES project (Jouitteau, 2023; **?**) a public wiki documenting both the Breton language and its ecosystem. In particular, we extract data from the interlinear glosses it includes to illustrate various phenomena and transform them into a parallel corpus: ARBRES Kenstur. The resulting corpus includes several thousand pairs of sentences and using it to train a machine learning system results in significant improvements, compared to both the only publicly available (rule-based) MT system and an experimental NMT system trained on previously available data.

2. Existing Resources

Most of the existing work in machine translation for Breton come from the efforts of Tyers (2009, 2010), with development of a Breton plugin for the Apertium (Forcada et al., 2011) platform backed by a morphological analyzer and a bilingual dictionary. While its performances are not sufficient for autonomous day-to-day usage, it can usually function as a "gisting" system, providing hints as to the meaning of a Breton sentence, albeit not a complete or faithful translation. Most importantly, Tyers (2009) was accompanied by the release of a parallel corpus of 63.8 ksentences, sourced from translation memories of the *Ofis Publik ar Brezhoneg* — a public organization coordinating translation efforts for the Brittany regional government — and therefore of very reliable quality. Extensions of this work have been proposed (Sánchez-Cartagena et al., 2015, 2020), but the resulting datasets and systems have never been publicly released.

Since then, MT systems for Breton have mostly come in the form of multilingual system including Breton, most notably OPUS-MT (Tiedemann and Thottingal, 2020) and M2M100 (Fan et al., 2021). However the performances of these systems for Breton are abysmal: see table 1 for a quantitative report, qualitative assessment is also trivial, these models usually producing nothing that resembles Breton at all. While a full investigation of the causes of this situation is out of scope for this work, a cursory look at the main training source for the models, the WikiMatrix (Schwenk et al., 2021a) and CCMatrix (Schwenk et al., 2021b) corpora shows that their quality for Breton is also very bad, most sentence pairs not being related at all. This could have in turn been foreseen, since these corpora come from parallel sentence mining backed by the LASER sentence embedding (Artetxe and Schwenk, 2019), which also have very bad performances for Breton. Therefore, even if they do include Breton in their supported languages, these systems are not usable for Breton and the successor to M2M100, NLLB200 (NLLB Team et al., 2022) removed it from their supported languages.

As for corpora, beyond those already mentioned, the OPUS repository (Tiedemann, 2012) lists a few other resources, most which come from software translation efforts and as such only list translation of very short phrases in very specific domains.

3. ARBRES

3.1. Wikigrammar

The data from the ARBRES wikigrammar were collected and annotated by a linguistics researcher to conduct fundamental research in formal linguistics. In this sense, the data are those of a research notebook. The data was then organized and significantly augmented with the aim of creating a descriptive grammar, usable in its online form by the speaking community. The goal is therefore twofold: to produce a description of a natural language in its diversity, complexity and regularities, and to provide new data relevant to fundamental research debates in generative linguistics.

In the corpus that this constitutes, one finds free corpus data, extracted from oral interviews or various cultural products (newspaper articles, novels, songs, poems, collections of popular expressions, political leaflets, town hall presentation sites, posts on social networks, etc.). It contains the somewhat artificial sentences typical of grammars, but they are outweighed by other more natural ones, of varied informational structure.

The corpus also includes elicitation data, a result of fieldwork for linguistic description purposes. The linguist has subjected native speakers to a protocol of questions, translations, descriptive tasks of images, or tasks of judgments of grammaticality of sentences which were proposed to them. Copyright on these sources is respected because the speakers provide informed consent on the dissemination of the results of the surveys, or where applicable, on the online distribution of their voice.

The presence of written corpus data from the 20th century means, in the case of Breton, the presence of several competitive spellings. The source data has not been altered, and examples appear in their original printing spellings. However, each form is connected to its standard equivalent.

ARBRES is a grammar of dialects. Its corpus has a high dialectal diversity by design. This is a descriptive grammar, not a prescriptive grammar. Standard Breton is treated as a dialect among the others. The dialectal spectrum is therefore quite broad. The Gwenedeg dialect is specifically underrepresented, with a relative deficit of data in this dialect which is also linguistically the furthest from the others. Its analysis requires expertise where the main editor is sometimes lacking, and as a result less data represents this dialect. Aside from this particular deficiency in the Gwenedeg dialect, we can consider that quantitatively, rare dialect facts are over-represented in the data.

The linguistic facts that are in the language will be illustrated once for each major dialect, but not beyond. On the contrary, to be able to precisely describe a rare fact, its dialect distribution and the parameters of their context of appearance, its examples will be provided for each existing occurrence. Rare facts are also more likely to be the subject of thematic elicitation research, which provides more data where they occur. For the same purpose of describing the variation, the forms of different styles will co-exist within the corpus, with a quantitative over-representation of this variation compared to any single corpus. In this sense, the ARBRES corpus is not well adapted for quantitative studies, but it offers a concentrate of grammatical diversity.

3.2. Glosses

Interlinear glosses are a tool for linguistic description and language education, where linguistic material in the language to be described is analyzed — typically at a more or less fine morphosyntactic level — and translated, often with both rough word-by-word equivalents and a global translation. For instance (1) is a gloss of a sentence illustrative the singulative phenomenon in Breton.

(1) *eur mell gwezenn glas he deliou.*
a big tree.SG green his^{2} leaf.PL
'a big tree with green leaves'

The current standard guidelines for gloss are the Leipzig Glossing Rules (Comrie et al., 2015), however it is common for authors to adapt them to their needs, such as in (1), where the non-standard superscript 2 indicates a consonant mutation trigger.

In particular, in ARBRES, words in glosses are made clickable for the user of the interface. Their redirect addresses are the spellings of their standard forms. The multiplicity of spellings present, combined with the systematic linking of each occurrence to its standard lemma, allows for a high diversity in the data without being detrimental to their consistency. This very system that redirects the tokens towards their respective lemmas also makes it possible to connect the various word forms. This is key in this Celtic language, which not only show inflections by suffixation, but also modifications of the initial consonant depending on the syntactic contexts in which they appear (mutations).

The lemma *krokodil* can thus be automatically linked to its occurrences in "*krokodil Maia*" "the crocodile of Maia", "*ar c'hrokodil*" "the crocodile", "*ar grokodilez*" "the female crocodile", "*war grokodileta*" "about to look for crocodiles". In the wiki, all these occurrences point to the same page dedicated to the lemma "krokodil". This page being categorized as a page concerning a noun, its grammatical category is also automatically recoverable. For a detailed description of the recoverable grammatical annotations, see Jouitteau and Bideault (2023) for the details of the data extraction project by the AUTOGRAMM Breton treebank II project.

All the examples include translations, either sourced from their original publication or provided by the author, but in all cases by fluent speakers of Breton. While translations are provided merely to help non-speakers to make sense of the source material, in our case, they are of tremendous interest, since they enable building corpora of parallel sentences from this linguistic field work material.

4. Corpus

Extracting parallel sentences from ARBRES was made relatively simple by the consistent use throughout of the wiki of the `prettytable` environment (see fig. 1) which allows both an easy identification and a convenient extraction, of which parallel sentences are only the most simple part.

The extraction was realized with a few simple scripts[1] operating on an archive of the wiki. Some manual work of polishing and uniformization was required on the wiki side to ensure a reliable extraction, but it was not prohibitively long and if anything, should help make the gloss coding more consistent in the future. This pipeline, with no post-hoc intervention on the parallel data ensures that subsequent versions of the wiki will be easily transposable to new and improved version of the corpus.

The resulting corpus consists of 5192 deduplicated sentence pairs, which, while an order of magnitude less than the Ofis Publik corpus, is still a significant sum of data. A cursory exploration revealed that some noise was still present (e.g. some gloss comments ended up in the translations and some deduplication imperfections), that should be fixed in subsequent releases after corrections on the wiki side, but overall, the sentences are of great quality and consistency.

5. Use for Machine Translation

Since the amount of gathered data is modest compared to existing parallel corpora for Breton, we assess the usefulness of this collection work by using it to train NMT models. Our hypothesis is that the diversity of this dataset, particularly compared to the Ofis Publik corpus should compensate for its size. To this end, we compare three Breton-French MT systems:

- The Breton-French Apertium plugin introduced by Tyers (2010) and made publicly available by the Apertium project.
- M2M100 (Fan et al., 2021), in its "418M" setting, which claims Breton among its supported languages.
- Two versions of the same M2M100-418M model, fine-tuned respectively on subset of the Breton-French parallel data available on OPUS (Tiedemann, 2012) and on this same subset augmented with ARBRES Kenstur.

[1] Available at `https://anonymized.for.review`

```
{| class="prettytable"
|(1)|| eur || mell || gwezenn
|-
||| [[un|un]] || [[mell|grand]] || arbres.[[-enn|SG]]
|-
||| colspan="15" | 'un grand arbre'
|}
```

Figure 1: Wikitext code for the gloss of “*eur mell gwezenn*”

The inspiration for the fine-tuning of M2M100 comes from Adelani et al. (2022), and we use the same hyperparameters as them[2].

Since the quality of the OPUS dataset for Breton is not good in general, we use the heuristic filtering tools proposed by Opus Tools (Aulamo et al., 2020) to filter out obviously wrong sentence pairs[3], resulting in training material that consists almost exclusively of the Ofis Publik corpus and some of the software translation data mentioned earlier.

The evaluation of the resulting models was conducted using sacreBLEU (Post, 2018), using a test dataset provided by the *Ofis Publik ar Brezhoneg* and used internally to evaluate their own private translation tools. So far our calls for a public release of this dataset have yet to be answered by the Ofis, but we hope that it could be the foundation of a standard for the evaluation of Breton-French MT systems in the future.

Model	**BLEU**	**ChrF++**	**TER**
Apertium	24.15	50.23	63.93
m2m100-418M	0.58	11.85	114.49
+OPUS	30.01	50.16	55.37
+ARBRES	37.68	56.99	48.65

Table 1: Evaluation results on the Ofis test dataset. See the complete sacreBLEU signatures in Appendix.

Table 1 reports the results of the various models in terms of BLEU, ChrF++ and TER scores. We can observe that the older and rule-based Apertium model does not fare too badly compared to even the state-of-the-art M2M100 model fine-tuned on the same data, but that adding the ARBRES data results in a clear and significant improvement across all scores, suggesting that even this limited amount of data is a very valuable boon for improving Breton-French MT.

6. Conclusion and Perspectives

In this work, we have shown that with very little efforts, it is possible to extract parallel corpus data from a grammar, provided that it is already available in digital form. Moreover, the data collected this way, while modest in size, can lead to significant improvement of machine translation systems. Interlinear glosses, designed as a tool for human consumption can thus become a convergence point between field linguistics, academic NLP and digital linguistic tools development for the general public[4].

This solution of data gathering is expensive in that it requires one or more people trained in the language with minimal dialectal flexibility, a social surface suitable for reaching speakers of different linguistic profiles, ways for them to find a non-monetary advantage in passing a linguistic protocol. This work also represents a long time of coding the examples and their adequate presentation in the grammar for a human readership. It requires technical support for the design and general maintenance of the site and its updates, and technical monitoring of its accessibility on screen for various users. However, all of these necessary resources exist outside the scope of NLP.

At the community level, investment may be driven entirely by internal goals. The database incrementally builds an educational and/or scientific resource in a form adapted to its audience. On the scale of small language communities, this avoids monopolizing experts to create databases which would not be usable by the general public.

The development of wikigrammars is particularly recommended for the construction of pilot project resources on languages with restricted corpus, because if the IT field fails to provide finalized tools for speakers, the investment will remain beneficial for the speaking community, which can truly continue to improve it for itself. In terms of human resources, descriptive and formal linguists set themselves the task of producing language analysis material. They are generally few in number in languages with a restricted corpus, but often have profiles that are very committed to their empirical domain and the speakers who produce it, with a de-

[2] Using Zelda Rose `https://zeldarose.readthedocs.io` for the actual training.

[3] See detailed settings at `https://anonymized.for.review`

[4] A demonstrator for the translation model is available at `https://anonymized.for.review` and has received a positive welcome from the Breton-speaking community.

tailed cultural knowledge of interactions with them. The wiki solution, for its part, is directly designed for large-scale collaboration of potentially isolated contributors, which is particularly suitable for minoritized languages.

7. Bibliographical References

David Adelani, Jesujoba Alabi, Angela Fan, Julia Kreutzer, Xiaoyu Shen, Machel Reid, Dana Ruiter, Dietrich Klakow, Peter Nabende, Ernie Chang, Tajuddeen Gwadabe, Freshia Sackey, Bonaventure F. P. Dossou, Chris Emezue, Colin Leong, Michael Beukman, Shamsuddeen Muhammad, Guyo Jarso, Oreen Yousuf, Andre Niyongabo Rubungo, Gilles Hacheme, Eric Peter Wairagala, Muhammad Umair Nasir, Benjamin Ajibade, Tunde Ajayi, Yvonne Gitau, Jade Abbott, Mohamed Ahmed, Millicent Ochieng, Anuoluwapo Aremu, Perez Ogayo, Jonathan Mukiibi, Fatoumata Ouoba Kabore, Godson Kalipe, Derguene Mbaye, Allahsera Auguste Tapo, Victoire Memdjokam Koagne, Edwin Munkoh-Buabeng, Valencia Wagner, Idris Abdulmumin, Ayodele Awokoya, Happy Buzaaba, Blessing Sibanda, Andiswa Bukula, and Sam Manthalu. 2022. A Few Thousand Translations Go a Long Way! Leveraging Pre-trained Models for African News Translation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3053–3070, Seattle, United States. Association for Computational Linguistics.

Mikel Artetxe and Holger Schwenk. 2019. Massively Multilingual Sentence Embeddings for Zero-Shot Cross-Lingual Transfer and Beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610.

Mikko Aulamo, Umut Sulubacak, Sami Virpioja, and Jörg Tiedemann. 2020. OpusTools and Parallel Corpus Diagnostics. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3782–3789, Marseille, France. European Language Resources Association.

Loïc Barrault, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, Christof Monz, Mathias Müller, Santanu Pal, Matt Post, and Marcos Zampieri. 2019. Findings of the 2019 Conference on Machine Translation (WMT19). In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 1–61, Florence, Italy. Association for Computational Linguistics.

Bernard Comrie, Martin Haspelmath, and Bickel. 2015. The Leipzig Glossing Rules: Conventions for interlinear morpheme-by-morpheme glosses.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Edouard Grave, Michael Auli, and Armand Joulin. 2021. Beyond english-centric multilingual machine translation. *The Journal of Machine Learning Research*, 22(1):107:4839–107:4886.

Mikel L. Forcada, Mireia Ginestí-Rosell, Jacob Nordfalk, Jim O'Regan, Sergio Ortiz-Rojas, Juan Antonio Pérez-Ortiz, Felipe Sánchez-Martínez, Gema Ramírez-Sánchez, and Francis M. Tyers. 2011. Apertium: A free/open-source platform for rule-based machine translation. *Machine Translation*, 25(2):127–144.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The State and Fate of Linguistic Diversity and Inclusion in the NLP World. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Mélanie Jouitteau. 2023. ARBRES, wikigrammaire des dialectes du breton et centre de ressources pour son étude linguistique formelle.

Mélanie Jouitteau and Reun Bideault. 2023. Outils numériques et traitement automatique du breton. In Catherine Schnedecker, Annie Rialland, and Michela Russo, editors, *Forthcoming*. Société Linguistique de Paris.

Christopher Mosely and Alexandre Nicolas. 2010. *Atlas of the World's Languages in Danger*, 3 edition. UNESCO, Paris.

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.

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.

Víctor M. Sánchez-Cartagena, Mikel L. Forcada, and Felipe Sánchez-Martínez. 2020. A multi-source approach for Breton–French hybrid machine translation. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 61–70, Lisboa, Portugal. European Association for Machine Translation.

Víctor M. Sánchez-Cartagena, Juan Antonio Pérez-Ortiz, and Felipe Sánchez-Martínez. 2015. A generalised alignment template formalism and its application to the inference of shallow-transfer machine translation rules from scarce bilingual corpora. *Computer Speech & Language*, 32(1):46–90.

Holger Schwenk, Vishrav Chaudhary, Shuo Sun, Hongyu Gong, and Francisco Guzmán. 2021a. WikiMatrix: Mining 135M Parallel Sentences in 1620 Language Pairs from Wikipedia. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1351–1361, Online. Association for Computational Linguistics.

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

Gary F. Simons, Abbey L. L. Thomas, and Chad K. K. White. 2022. Assessing Digital Language Support on a Global Scale. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4299–4305, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

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

Jörg Tiedemann and Santhosh Thottingal. 2020. OPUS-MT – Building open translation services for the World. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 479–480, Lisboa, Portugal. European Association for Machine Translation.

Francis Tyers. 2010. Rule-based Breton to French machine translation. In *Proceedings of the 14th Annual Conference of the European Association for Machine Translation*, Saint Raphaël, France. European Association for Machine Translation.

Francis M. Tyers. 2009. Rule-Based Augmentation of Training Data in Breton-French Statistical Machine Translation. In *Proceedings of the 13th Annual Conference of the European Association for Machine Translation*. European Association for Machine Translation.