Make wikigrammars!

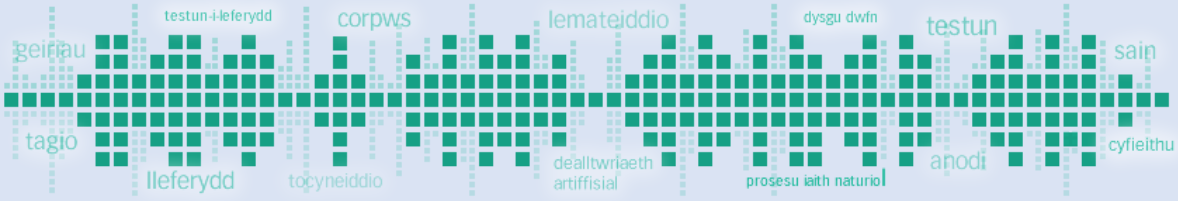
Mélanie Jouitteau, Loïc Grobol

Language and Technology in Wales: Volume II

Watkins, Gareth; Prys, Delyth; Prys, Gruff; Jones, Dewi; Cooper, Sarah; Williams, Meinir; Vangberg, Preben; Ghazzali, Stefano; Gruffydd, Ianto; Farhat, Leena; Grobol, Loïc; Jouitteau , Mélanie; Morris, Jonathan; Ezeani , Ignatius; Young , Katharine; Davies, Lynne; El-Haj , Mahmoud; Knight , Dawn; Jarvis , Colin; Barnes, Emily

Published: 01/11/2024

Publisher's PDF, also known as Version of record



Language and Technology in Wales: Volume II

Editor: Gareth Watkins

PRIFYSGOL BANGOR UNIVERSITY

Ariennir gan Lywodraeth Cymru
Funded by Welsh Government

This ebook was first published in 2024 by

Bangor University, College Road, Bangor, Gwynedd LL57 2DG

www.bangor.ac.uk

International Book Number (ebook):

ISBN 978-1 84220-207-4.

Design and proofreading assistance from Prof. Delyth Prys and Stefano Ghazzali. This book is also available in Welsh under the title *Iaith a Thechnoleg yng Nghymru: Cyfrol II*, ISBN number 978-1 84220-208-1.

Make Wikigrammars!

MÉLANIE JOUITTEAU

l'Université Bordeaux Montaigne, France and l'Université de Pau et des Pays de l'Adour, France

LOÏC GROBOL

Université Paris Nanterre, France

This chapter presents an evaluation from a natural language processing (NLP) perspective on the concept of wikigrammars, using the Breton ARBRES wikigrammar as a case study. It explores the utilisation of a wiki-based platform for documenting the syntactic diversity of a low-resource Celtic language, with an interactive component aimed at community engagement. It constitutes a comprehensive, annotated corpus that supports both theoretical linguistics and NLP. We advocate for the adoption of such platforms by communities speaking minoritized languages, arguing that they provide corpora with rich syntactic, orthographic, and stylistic diversity. The artificially selected diversity of wikigrammars may mitigate the scarcity of extensive, freely available corpora in low-resource language contexts.

Keywords: Wikigrammars, NLP, Breton, Corpus

1 INTRODUCTION

A wikigrammar is a wiki-based platform describing a language that is open to contributions and discussions, and whose examples are annotated and automatically retrievable.

Wikigrammars provide a very specific type of resource for language technology development: a corpus that is by definition a concentrate of linguistic diversity. In this article, we present some features of the ARBRES wikigrammar of the dialects of Breton [1], a low-resource Celtic language. We recommend the adoption of this solution by communities of minoritized languages to foster the development of their digital resource ecosystem for language technologies.

2 LINGUISTIC DIVERSITY BY DESIGN

The primary goal of ARBRES is to provide a comprehensive description of Breton, capturing its diversity, complexity and regular features, accessible in its online and written forms for the speech community. Such a grammar should not only mention and describe the most common structures, but also exceptions and infrequent phenomena. The statistical distributions of words and structure are therefore much more diverse than they would be in a random sample of the language of a similar size. Another effect contributes to the diversity of the data: for copyright reasons, the author could only take a modest percentage of the sentences for each published corpus. The effect is a widening of the variety of sources for printed free corpora (literature, newspaper articles, novels, songs, poems, collections of popular expressions, political leaflets, town hall presentation sites, posts on social networks, etc.).

A second goal of ARBRES is to be a documented resource for ongoing debates in theoretical linguistics. In that way, it is akin to a regular research notebook for its main author. This has also had an influence on the resulting data: it includes the somewhat artificial sentences typical of grammars and research papers. However, these are significantly outweighed by more natural examples. This source of data includes minimal pairs and negative evidence.

Another major source of examples is fieldwork elicitation data, where native speakers have been subjected to protocols of questions, translations or descriptive tasks of images. The raw results of the elicitations are posted online and feed the grammar. These protocols also include tasks of judgments of grammaticality of sentences, resulting in ungrammatical example that serve as contrastive negative evidence. This source of data thus also includes minimal pairs and negative evidence.

2.1 Dialectal and historical diversity

ARBRES is a grammar of dialects and has by design a high dialectal diversity. It is a descriptive grammar, where standard Breton is merely one dialect among others. The dialectal spectrum is therefore quite broad, with the notable exception of the Gwenedeg dialect, which is linguistically the furthest from the others, and is currently underrepresented in ARBRES. Its analysis requires expertise that the main editor is sometimes lacking, and as a result less data represents this dialect.

Aside from this particular caveat, we can consider that quantitatively, rare dialectal features are over-represented in the data. Indeed, common linguistic features are only illustrated with a few examples for each major dialect. On the contrary, to be able to precisely describe a rare feature, its dialect distribution and the parameters of their context of appearance, each existing occurrence will be carefully integrated. Rare features 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, while ARBRES is questionable for quantitative studies, it is very well suited for qualitative ones.

Finally, while ARBRES is not strictly speaking a diachronic work, it still includes data from Middle Breton to 21st century Breton. The presence of written corpus data from these periods implies, especially for the twentieth century, the presence of several competing orthographic systems. The source data has not been altered, and examples appear in their original printing spellings. The result is a multi-orthographic corpus.

2.2 Size

The ARBRES website has been developed since 2007, and started having an online presence in 2009. In the last five years, it has received more than 100 human visits per day. As of February 2024, the site consists of 10,238 pages, including 4,804 pages of content, 19 pages of presentations, and a number of redirection pages. The content pages consist of 3,094 articles on elements of Breton grammar and 325 sheets which each provide an explanation of a linguistic term or concept. Overall, this amounts to about 15,000 original Breton sentences, glossed and translated into French, coming from 1,208 research works on the Breton language (books, dictionaries, research articles, data collection blogs), 493 corpus references produced by native speakers (mostly in written forms: novels, newspaper articles, songs) and 44 elicitation sessions with native speakers.

3 A DATA SOURCE FOR NLP

3.1 Morphosyntactic annotations

The examples are provided under the form of wikitables, tables in the markup language of the MediaWiki software [2] that powers ARBRES. Each of these tables provide for a single sentence alignment of the original word forms and their glosses, the global translations of the sentence, the name of the dialectal variety, and the reference of the

source. Each word form gloss is connected via a hyperlink to a dedicated page, including at least its standard lemma and its grammatical category. Given the variety of possible spellings, this allows a high consistency in the data without being detrimental to diversity.

This system also makes it possible to reach all the word forms for a given lemma, which is crucial in this Celtic language, where inflexions are not only suffixes, but also modifications of the initial consonant depending on the syntactic context (consonant 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 c'hrokodiled* (the crocodiles) and *war grokodileta* (about to look for crocodiles), all these occurrences pointing to the page for the lemma *krokodil*. Conversely, disambiguation pages provide clickable lists of morphemes and words with more than a single meaning.

From a language technology point of view, this means that the glosses on ARBRES are already a morphosyntactically annotated corpus: a set of sentences, with lemmas and part-of-speech tags for every word and additional morphological features. It also makes it a very good seed for growing [3]. For additional details on the recoverable grammatical annotations, see Jouitteau and Bideault [4].

3.2 Parallel data

All the glosses also include translations in French, either sourced from their original publication or provided by the author, but in all cases by fluent speakers of Breton. While these translations were originally provided merely to help non-speakers make sense of the source material, they can also be seen as a parallel corpus of sentences.

This corpus is of a modest size, but it is of a very high quality and has a much larger diversity than a random sample of an equivalent size would have. Its quality simply comes from the origin of the data: all sentences have been manually selected, translated by fluent speakers, and validated carefully to ensure their relevance as illustrations of linguistic phenomena. The high diversity is ensured by the function of the glosses: since they are meant to illustrate as many linguistic phenomena as possible while taking dialectal variations into account, rare phenomena will be over-represented relatively to their organic occurrences.

Currently ongoing experiments in developing machine translation system using an early extraction of this data (around 5,000 deduplicated sentences, after removal of negative data and instances of failed data extraction) tend to confirm that these characteristics make ARBRES a very valuable dataset. Indeed, its inclusion in the training data of off-the-shelf systems result in performance gains that are comparable to those obtained with an order of magnitude more data [5].

3.3 Cost estimates

Using wikigrammars as sources of linguistic data is expensive in that it requires one or more people trained in the language, with a certain dialectal flexibility, and a social network suitable for reaching speakers of different linguistic profiles. It also requires technical support to design and maintain the website, and ensure its accessibility. Data extraction also requires qualified workers. The more laborious task is the extensive coding of examples for their appropriate presentation within wikitables. The complexity of this task is evolving fast, due to improvements in natural language generation. For the Breton wikigrammar, it took 15 years for a single annotator, working about half-time on it, to barely reach annotation of 15,000 sentences.

Chatbots now enable the automation of a significant portion of the annotation work. For instance, as of 2024, with a suitably detailed prompt providing seven examples of structured data, Chat GPT 3.5 is capable of distributing tokens across tables, aligning glosses, encoding a large fraction of clickable links, offering translations (inaccurate,

but correctly aligned), and properly organising source references. Manual interventions by a language expert are still indispensable, but they have been significantly simplified, to the point where a single person can easily enter 300 examples per month. ChatGPT 4 enhances this process even further with superior translations. Of course, this last capability is contingent upon the volume and quality of the targeted language data within the ChatGPT training dataset. These systems have well-known downsides, most notably in term of social impact and inefficiency (see Solaiman et al. [6] and references therein), but their value as assistive tools for this task is a good indication of how much systems developed specifically for this task could achieve (while avoiding said downsides).

The novelty of the solution is that all of these necessary resources and goals might exist outside the scope of NLP research. Investment may be driven entirely by internal goals at the community level, or by a linguist for scientific purposes. Moreover, ARBRES was created by a formal linguist, but it doesn't have to be: as long as the grammar is written to teach humans about the language, the required amount of diversity will be found in the data. The resource can then be incrementally built as an educational and/or scientific resource in a form adapted to its audience. On the scale of small language communities, this avoids monopolising experts to create resources that would not be usable by the general public. The more specialised annotations of the data (grammatical categorisation, lemmatisation, coding of consonant mutations) remains inconspicuous, and just serves navigation of the human reader.

The development of wikigrammars is particularly recommended for the construction of pilot project resources on languages with restricted corpora, since even where tech actors fail momentarily to provide finalized tools for speakers, the investment will remain beneficial for the speaking community, which can truly continue to improve the wikigrammar for itself.

Descriptive and formal linguists set themselves the task of producing language analysis material, and can develop these without NLP training. The wiki syntax has a very low entry cost, which is now roughly that of a normal word processing program. In languages with restricted corpora, linguists and trained experts are often very committed to their empirical domain and to the speakers who produce the data. They usually have a precise cultural knowledge, including the diversity of live data, and this also has a beneficial impact on the chosen examples. In terms of human resources, this solution makes it possible to capture their fine-grained expertise. Finally, the wiki solution is designed for large-scale collaboration of potentially isolated contributors. This is particularly suitable for minoritized languages where linguists and trained experts are usually few in numbers, sometimes in precarious socioeconomic situations. Finally, wikigrammars allow for the corpus to be built under the review, direct and indirect, of the entire speaking community.

4 SOCIAL ENGAGEMENT IN MINORITY LANGUAGES

4.1 Public engagement

Internal statistical tools, as well as external analytics systems provide precise insights about the uses of the website, by tracking (anonymously) the more than 100 daily human visits in ARBRES. The graph in Figure 1 shows global visit statistics from October 2023 to the end of January 2024.

Studying the flow of readers makes it possible to identify and understand gaps. One can see the successful entry pages, those that receive the lesser engagement or the shortest reading times, or the particular requests made on search engines that led the readers to the grammar.

Once the website has reached a critical size and a good representation in search engines, the geographical sources of connections can be analysed to provide information on the readership. ARBRES is predominantly utilized within Brittany and among diaspora communities, as seen in Figure 2, which reports the number of visits by city in January 2024.

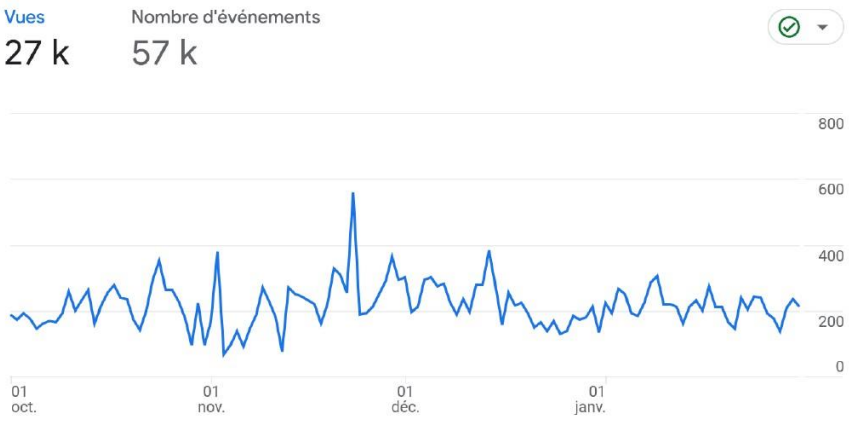


Figure 1: Number of visits on ARBRES from October 2023 to the end of January 2024.

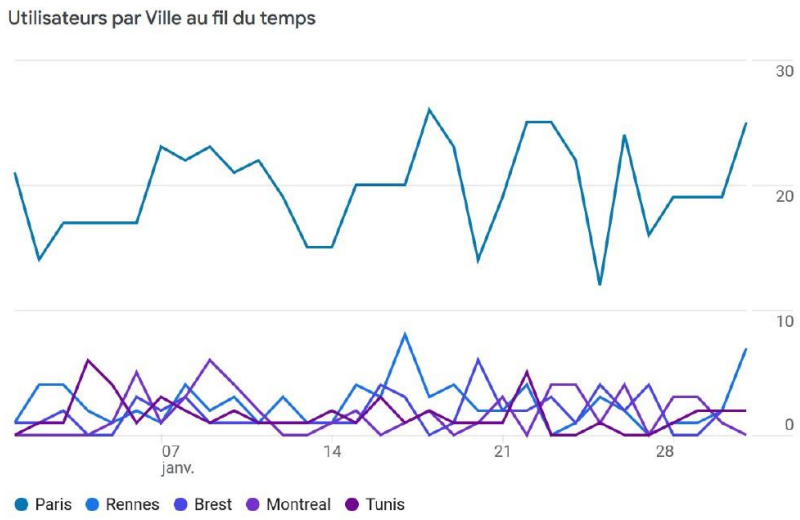


Figure 2: Number of visits on ARBRES in January 2024 split by city

The usage for the website is notably in sync with academic calendars. The sections dedicated to more complex aspects of formal linguistics, offering foundational information in French, experience significant surges in traffic

during typical examination periods in French-speaking regions (e.g., Switzerland, Morocco, Qu´ebec, Algeria, Belgium, etc.).

The precision of the geographic data allows for the observation of the resource's international usage, such as when Breton language courses refer to it. For instance, in 2010, Anna Mouradova started teaching Breton in Moscow, which sparked a spike in connections. Interestingly, one can also see where the resource is not identified (like the sporadic Breton classes in Harvard).

4.2 Equip the interface between science and society

The Breton wikigrammar ARBRES is an experiment of open and participative science (see Jouitteau [7] for an early analysis of the deployment). Wikigrammars bring the scientific process closer to the public. Like any other grammar in open access, it makes available the results of research at the end of its process at a given time. But it also does much more. In synchrony, it links the work to the used sources and to the scientific community. It also sheds light on the past of its making, and on the future of its making. We now illustrate these three dimensions.

Scientific monitoring makes it possible to feed the grammar with the results of the latest research. This effect only derives from its use as a research notebook. The external resources are summarized, referenced and, when open access allows for it, directly linked to. All of these operations bring the readership closer to scientific stakeholders, making them more understandable and more accessible. In 2014, the organization of the Redadeg (Race for the Breton language event) asked for the translation of "I speak Breton, and you?" in different languages. In a few days, linguists from all over the world happily participated in contributing to the page I speak Breton, what about you?, bringing together translations of this sentence in 77 different languages. In support of the event, 1,695 Breton speakers posted self-portraits online with these sentences. The international community of linguists was rendered visible to the general Breton speaking community, and conversely the Breton language appeared very concretely as a production of living speakers to the scientists.

A wikigrammar also includes its whole history. It references the making of its own research. The wiki history function offers for each page a full traceability of the process of knowledge building and data gathering: contributions, corrections, discussions, exploration of new datasets, integration of new bibliographic sources and new hypotheses on the rise being tested. Each page is associated with a complete history giving all modifications made to it since its creation. One can trace back how science is done, how new data and new publications change our hypotheses. The diversity of contributors or lack thereof for each topic is visible. Every contribution is visible and can be duly credited.

Scientific research is the result of a methodology, and is at heart a process accessible to anyone, as long as the methodology is respected. Within these limits the wiki software is designed to allow both cumulative collaboration (massive aggregation of small contributions into a single architecture), and a distributive collaboration (with differentiated tasks). Various competences can then come in together to build a strong resource for the community. This medium raises for the reader the question of their place in the process, enabling a spectrum of engagement levels, ranging from passive activities such as reading, to active participation like commenting, correcting, providing input, writing, linking and so forth. This is particularly welcome in the case of minoritized languages, where speakers commonly report feelings of dispossession of what they consider their language.

Finally, let us discuss a marginal but beneficial effect. Society is rife with sometimes poorly informed debates about languages and especially minoritized languages, due to a lack of verifiable information, a lack of objective knowledge of the linguistic varieties, or an accumulation of inaccuracies. The wikigrammar hosts linguistic

discussion articles which provide concrete elements of analysis on these debates, and proper scientific references. The digital format of these articles makes them directly shareable on social networks, in a format open to a scientific discussion, within the limits of scientific argumentation. In ARBRES, the article about the Sapir-Whorf hypothesis is the second most frequently visited page on the website.

In turn, the traffic generated by this supports search engine optimization and maintains visibility for a work related to marginalized languages on the Internet.

REFERENCES

[1] Mélanie Jouitteau. 2009–2024. ARBRES, Wikigrammaire Des Dialectes Du Breton et Centre de Ressources Pour Son Étude Linguistique Formelle. Retrieved from http://arbres.iker.cnrs.fr

[2] Magnus Manske and Lee Daniel Crocker. 2002. MediaWiki. Retrieved from https ://www.mediawiki.org/wiki/MediaWiki

[3] Mélanie Jouitteau, Yidi Jiang, Yingzi Liu, Salomé Chandora, Kim Gerdes, Bruno Guillaume, Adrien Said-Housseini and Sylvain Kahane. 2022–2024. Autogramm/Breton II. Retrieved from https://github.com/Autogramm/Breton

[4] Mélanie Jouitteau and Reun Bideault. 2023. Outils Numériques et Traitement Automatique Du Breton. In: Langues Régionales de France: Nouvelles Approches, Nouvelles Méthodologies, Revitalisation. Société Linguistique de Paris, 37–74.

[5] Loïc Grobol, and Mélanie Jouitteau. 2024. ARBRES Kenstur: A Breton-French Parallel Corpus Rooted in Field Linguistics. In: Forthcoming.

[6] Irene Solaiman, Zeerak Talat, William Agnew, Lama Ahmad, Dylan Baker,Su Lin Blodgett, Hal Daumé III au2, Jesse Dodge, Ellie Evans, Sara Hooker, Yacine Jernite, Alexandra Sasha Luccioni, Alberto Lusoli, Margaret Mitchell, Jessica Newman, Marie-Therese Png, Andrew Strait and Apostol Vassilev. 2023. Evaluating the Social Impact of Generative AI Systems in Systems and Society. Jun. 12, 2023. arXiv: 2306.05949.

[7] Mélanie Jouitteau. 2012. La linguistique comme science ouverte. In: Lapurdum. Euskal ikerketen aldizkaria | Revue d'études basques | Revista de estudios vascos | Basque studies review 16 (16 1st Oct. 2012), 93–115. doi: 10.4000/lapurdum . 2357.