Community Internally-driven Corpus Buildings. Three Examples from the Breton Ecosystem

Mélanie Jouitteau

▶ **To cite this version:**

Mélanie Jouitteau. Community Internally-driven Corpus Buildings. Three Examples from the Breton Ecosystem. 2nd Annual Meeting of the ELRA/ISCA SIG on Under-resourced Languages (SIGUL 2023), Aug 2023, Dublin, Ireland. pp.103-107, ⟨10.21437/sigul.2023-22⟩. ⟨hal-04384300⟩

HAL Id: hal-04384300

https://hal.science/hal-04384300v1

Submitted on 10 Jan 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

HAL Authorization

2nd Annual Meeting of the ELRA/ISCA SIG on Under-resourced Languages (SIGUL 2023)
18-20 August 2023, Dublin, Ireland

Community internally-driven corpus buildings

Three examples from the Breton ecosystem

Mélanie Jouitteau, IKER, UMR 5478, CNRS

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

Abstract

This paper is a position paper concerning corpus-building strategies in minoritized languages in the Global North. It draws attention to the structure of the non-technical community of speakers, and concretely addresses how their needs can inform the design of technical solutions. Celtic Breton is taken as a case study for its relatively small speaker community, which is rather well-connected to modern technical infrastructures, and is bilingual with a non-English language (French). I report on three different community internal initiatives that have the potential to facilitate the growth of NLP-ready corpora in FAIR practices (Findability, Accessibility, Interoperability, Reusability). These initiatives follow a careful analysis of the Breton NLP situation both inside and outside of academia, and take advantage of pre-existing dynamics. They are integrated to the speaking community, both on small and larger scales. They have in common the goal of creating an environment that fosters virtuous circles, in which various actors help each other. It is the interactions between these actors that create quality-enriched corpora usable for NLP, once some low-cost technical solutions are provided. This work aims at providing an estimate of the community's internal potential to grow its own pool of resources, provided the right NLP resource gathering tools and ecosystem design. Some projects reported here are in the early stages of conception, while others build on decade-long society/research interfaces for the building of resources. All call for feedback from both NLP researchers and the speaking communities, contributing to building bridges and fruitful collaborations between these two groups.

Index Terms: FAIR practices, corpus-building tools, citizen science, open science, language policies, Celtic, Breton

1. Fairness vs. farness

The growing gap between high-resource languages and low-resource languages is a well-known source of degradation of fairness in NLP development. The current success of AI for super-equipped languages will only worsen the trend, and marginalize more languages. Top-down approaches of technical solution engineering typically face implementation problems. To the extend that they require the involvement of the actual speakers, they typically lack dissemination techniques, as well as feedback channels for measuring their usability and efficiency.

The starting point of this work is the observation that the NLP development of small corpus languages needs a bridge to be engineered between two groups: the communities of minoritized languages and NLP developers. The speaking communities are in urgent need of NLP applications, and they remain the producers of quality-corpus while synthetic data is not yet available. On the other hand, NLP developers are in need of corpora and follow their best conception of linguistic fairness principles. This is where distance becomes costly. They may not share any common language with the target language speakers. They may have an inaccurate representation of how the target society can provide the fastidious work of data enrichment. They may misrepresent the needs that are self-identified by the community, and consequently miss ways to propose corpus-building strategies to fulfill them.

A Breton on-the-ground experience provides some concrete examples. In academia, everybody knows English, or has to pretend to do so. Breton is however a Celtic language whose speakers are bilingual with a non-English language (Romance French). Feedback questionnaires conceived in English receive poor engagement, and worse, yield misleading answers. The field of syntax can testify from concrete experience that written elicitation protocols in English can lead to confusing results if they rely on the English proficiency of the speakers. Knowledge of the social environment of the speakers is also key. In the context of the development of the global digitization of medical services, fairness principles can lead to prioritize the development of applications for the oldest speakers, facilitating access to medical services in their native language [1]. However, the profile of Breton elderly speakers precisely has difficulty accessing written Breton, and the standard dialect. It also remains unclear how making medical services digitally accessible solves anything if the medical staff consists exclusively of non-speakers. Finding support in society can also be difficult when access to medical care proves already difficult in the majority language. The motivation of academic literature researchers can lead them to annotate data in order to facilitate the quantitative exploration of texts. But this may happen only if the language has an academic field dedicated to it, with existing corpus quantifying habits. This is not the case in Brittany. The smaller the language, the farther its academic actors will be from an IT department. The long term effect is that language experts will face more difficulties in acquiring knowledge of basic tools and digital facilities, and access to them.

Communities of large languages can rely on a critical overlap between the two groups: the NLP developers may even form a subset of the speaking community, providing them with natural access to both the language and cultural representations. In the case of minoritized languages, the lack of a critical overlap between the two groups has to be dealt with. How to engineer virtuous ecosystems integrated into the

10.21437/SIGUL.2023-22

speaking communities, whose operating mode would be consubstantial with the speaking communities, and whose independent outcome would be NLP-usable corpora ? This paper adds to Nicolas and al. 2020 [2] and Millour and Fort 2018 [3] to raise this question. It reports on the exploration of three Breton society internal ecosystems that showed potential to feed NLP development. This paper is also a call addressed to the NLP community for feedback and collaboration to build tools that could facilitate the emergence of larger corpora for all low-resource languages.

2. Breton corpus for NLP

The aim of this preliminary section is to provide a very raw image of the current availability of Breton corpora for NLP development, and to ensure its comparability with other minoritized languages of the Global North. For a better survey of NLP development in Breton, its resources and potential, see Tyers and Howell 2021 [4], Jouitteau 2023 [5] and references therein.

We consider here only open copyright material, which NLP can build upon. Available raw written corpus is mostly to be found in the wiki suite, (wikipedia.br, wikisource for Breton *wikimammenn*), and some books and articles one can gather online, with various copyrights attached. The parallel corpus for written material is mostly the corpus of Breton/French translations made available by the Public Office of the Breton Language, amounting to about a million words [6]. Parallel corpus for oral material is mostly that of Common Voice (approximately 11 validated hours). Speech synthesis is advanced [7]. Automated translation has started [4], [8], [9], as well as text-to-speech [10]. Most face corpus shortages and distribution challenges. There is a Breton dependency treebank of 888 sentences, *Breton KEB*, which has been available since 2018 [11], whose corrected version was released in 2023.

Some priorities in development could vastly improve corpus growth. Breton has a standard writing system. There is enough material to potentially automatize translation between this writing system and the various older writing systems, whose texts are more likely to fall into the public domain. There is no efficient OCR system available yet, despite a sizeable amount of numerized written corpus that could take advantage of it. Speech raw corpora also abound in radio archives, with various rights attached. An efficient speech-to-text with minimal dialectal flexibility would unlock a sizeable amount of audio data transcripts. Finally, a campaign directed towards the content producers (media, publishers, YouTube content producers) could vastly improve the adoption of open Creative Commons labelling.

In the remainder of this article, I report on three different community internal initiatives to gather and build corpora, and to make them available for development.

3. From wikigrammar to treebank

This section reports on the creation of a Breton treebank, using the data that has been independently gathered and annotated for an online grammar of the Breton dialects. The ARBRES website is developed under wiki. The content of the website is under an open Creative Commons license. It is written in French and has been in constant development since 2009 [12].

ARBRES is now a large Breton grammar with over two thousand articles. The grammar references previous descriptions and analyses of the language, and replicates the results with new material from contemporary speakers of various dialects. It is also a resource center for formal syntax research on the Breton language, including an up-to-date article on NLP resources for Breton [5]. Additionally, it provides an extensive bibliography about the Breton language, as well as an elicitation center where researchers can ask a Breton speaking linguist to conduce elicitation for their own research purposes. The elicitation protocols can even be co-constructed with the language expert.

ARBRES is an open science and citizen science experiment. It is designed as an interface between the speaking community and the research community [13]. The wiki is open for contributions and every page has a discussion page attached. Engagement of the community on the website is rather low in volume, but fruitful and with complete traceability. However, comments, suggestions and corrections are often addressed by email to the developer. According to the *Google Analytics* statistical observation tool, in 2022, 30,490 people started a session on the site, which corresponds to an average of 84 human visits per day. Over the last 28 days of December 2022, 4962 pages were viewed, with an average session time of 3:58 minutes. The readership encompasses historic Brittany as well as urban places typical of the diaspora. It is also used by French speaking academic communities over the world, probably linguistics students, with peak attendance around exam dates in North America, Switzerland, Belgium, and French-speaking Africa. This readership approaches formal linguistics explained in French with Breton sentence examples.

At the beginning of 2023, the website counted 9,511 pages, including 4,285 pages of content: 2,616 articles on elements of Breton grammar and 322 theoretical explanation sheets. The wikigrammar is illustrated by corpus data coming from 459 corpus references produced by native speakers of all dialects (novels, newspaper articles, songs), and 1,160 research works on the Breton language (books, dictionaries, articles, dialectal blogs, etc.). The data also comes from elicitations conducted by linguists. The raw results of 42 elicitation sessions with native speakers are available online in the elicitation center, which is a part of the website.

The examples that illustrate the grammar form a high quality corpus. A linguist has selected the examples to illustrate various speech levels, a broad spectrum of dialectal variations and various styles (written, oral, colloquial, literary, journalistic). The selection represents the global diversity of the language, with a positive bias towards rare linguistic facts. Each example is attached to meta-information : a speaker profile (native/late learner/child) and a dialect dialect (geolectal dialects/standard). As of 2023, we estimate that ARBRES contains 15,000 glossed and translated examples. Duplicates probably bring this resource back to around 5,000 original sentences. Interestingly for NLP uses, the wikigrammar has a system of clickable glosses, from which POS tag information can be retrieved.

The data on ARBRES thus constitutes a reasonably sized database of quality corpus, that can be extracted automatically from [13]. There is for Breton a first treebank, *Breton KEB,* a prototype dependency treebank for Breton of 888 sentences [11]. The idea is to extend this resource.

The **Breton treebank II** project has started in 2022. It aims at building an annotated *Universal Dependencies* corpus based on the existing data annotated in the *ARBRES* wikigrammar. It is part of the ANR funded *Autogramm* program (2022-2025) led by Sylvain Kahane in U. Paris Nanterre (Modyco, CNRS). A team led by Kim Gerdes (LISN!, CNRS) extracts the data in the wikigrammar and organizes them under a Conll-U format [15]. The coding is in SUD format with an automatic switch to UD managed by Grewmatch (Bruno Guillaume, LORIA, INRIA), and Arboratorgrew [16]. The remaining work, in progress, is to finish filling the Conll format by annotating the dependencies. This last step is partially automatized by a pre-annotation tool, building on Breton KEB [11]. Loic Grobol (Modyco, U. Paris Nanterre) coordinates the development of a parser. The writer of the Breton wikigrammar serves as a language expert throughout the project. The extracted data is deposited in a GitHub repository under ConLL format, ensuring that this reusable corpus is findable, accessible, and interoperable.

The feasibility of the data extraction out of the wikigrammar serves as a proof of concept. It is possible for communities to create a grammar of the language, an educational application for the speaking community, or a research application for the academic world. At the same time, it can serve as a platform for gathering data from within the community and generating a recoverable annotated corpus. The recuperation process is not cost-free, but it can be improved upstream by using more adequate annotation guidelines. This pilot project provides an example where the needs of a speaking community that are not related to NLP outputs (pedagogical material) are addressed in a way that grows enriched NLP-ready corpora. Digital dictionaries have been identified as applications that offer essential material for NLP development, while also serving immediate pedagogical purposes within the communities. Wikigrammars demonstrate a similar potential for replicating the results of coding efforts. The wiki environment is also rather accessible for language experts without coding knowledge. It is designed for low-cost and easy collaboration, and requires minimal IT maintenance over time.

4. Transcription loops

This section reports on a much smaller but successful initiative internal to the Breton speaking community for speech-to-text corpora building. In this case, the corpus is not consciously grown by the speaking community, but the data and the enrichment work are not stolen from them either. Instead, an ecosystem is installed where each actor operates for their own benefit, and is conscious of the global realization.

A **transcription loop** is a chain of actors that provides aligned sound and written corpora to NLP development. I present two micro-chains where each actor acts for his own interest. It is the entire ty of the chain that provides aligned NLP-ready corpora. The two examples are prototype micro-ecosystems I engineered in interface with the NLP speech-to-text developer Duval-Guennoc [10].

The first micro-chain consists of five individual actors. The observed loop is the following: A viewer of a utube video in Breton ask for subtitles (actor 1). The Breton speaking content producer (actor 2) expresses his lack of resources to provide the work needed for the required subtitles. A Breton teacher (actor 3) identifies a student (actor 4) among her students whose learning profile matches the required transcription task. The teacher (actor 3) corrects the pedagogical exercise for the student. She sends it to the Breton speech-to-text developer in need of corpora (actor 5). The developer (actor 5) integrates the corpus into his training set and sends back a time-aligned file for it. The srt file is offered to the content developer (actor 2). He is thanked for the open copyright of his content, which promotes FAIR practices.

Upscaling this micro-ecosystem is certainly a challenge, but note that each actor is operating for his/her own benefit. Actor 1 and actor 2 express frustration, and signal their needs on a social platform. The former will obtain the required subtitles, and the latter will have this work done for him. The content producer also expresses symbolic retribution from knowing he aided NLP development. The teacher (actor 3) freely uses the content of actor 2 for her pedagogical practice. It is important not to overstate the self-interest of the teacher, but her required efforts could be lessened. She could be provided a list of audio files to be transcribed with the associated dialect and an estimate of the level of transcription difficulty. With a transcription task, the student (actor 4) exercices her language skills under profesional supervision that guarantees that the exercice is adapted to her linguistic level and her current learning challenges. She gains both the transcription correction and a symbolic retribution: she may not feel completely secure speaking the language yet, but she already contributed to the speaking community. This alleviates for her the issue of learner's legitimacy. The NLP developer (actor 5) gets a small aligned corpus for the training of his speech-to-text tool, in exchange for the srt file. The process of corpus enrichment allows for its large distribution in terms of copyright licensing. All actors in the chain of corpus enrichment, of course, are to be acknowledged for their contributions.

Finally, I present experiences of pre-annotation of linguists' elicitations with the same NLP developer, much in the spirit of Le Ferrand and al. (2023) for Kréyòl Gwadloupéyen [17]. In this case, the raw elicitations files were pre-annotated by the speech-to-text tool in development [10], which is a hybrid model (deep neural network for the audio model and N-Gram type language model) in Vosk format, trained using the Kaldi framework. The VOSK-br-0.7 model (May 2023 release) shows a performance of 36.4% WER (word error rate) on the Mozilla Common Voice V11 test data set. Its automatic transcription is next corrected by the linguists and sent back as an NLP training set. This corrected dataset is then deposited in an online archive, *Cocoon* [18], which ensures that the reusable corpora created are findable, accessible, and interoperable.

This exact ecosystem could be replicated for the derushing by filmmakers, under the condition that they agree to return at least a portion of the corrected transcript.

The observed ecosystem of the transcription loop is not directly transferable to translation subtitles, because these can be semantically approximative and have to be shortened because of their reading time. As for dubbing files, they are plenty available in Breton, and they consist of rather natural texts read by actors, but they trigger copyright issues. Different copyrights are owned by actors, translators, original writers, lip synchronizers, and dubbing companies.

Upscaling could take the form of a public funded structure offering transcription services to all content producers putting an open license on their content. This service would only be

required until automatization could replace it. Alternatively, a structure could curate a list of corpora to be transcribed, and thus provide teachers with accurate material for pedagogical training.

5. Building a corpus mixer

This section presents a project at the conception stage. The project is to build a tool for a speaking community that is already aware of its own needs for NLP-usable corpora. The proposed scenario enables the speaking community to seek support from the linguistic policy structures. Preliminary interviews in Brittany suggest that it could find a supportive and dynamic environment within this speaking community.

The **corpus mixer** is a corpus-building tool for low-resource languages. It is designed to address both the needs for quality corpora and FAIR practices. It is a web interface for depositing written corpora. It allows the deposit of writings under proprietary copyright. The distribution under a form reconstructible by humans remains prohibited. This acceptation of the proprietary copyright aims at preserving the fragile economy of edition in low resource languages. The depositor however consents to the distribution of his content inasmuch as the distributed form lost its meaning at the text level. The sentences are NLP cleaned and enriched (disambiguation of acronyms, points, etc), and mixed with a mass of corpus. Its output can be freely used by developers of digital tools for automatic language processing. I will now describe step by step the corpus mixer's input, processing operations and output.

This tool is a text mixer with a web interface. It mixes sentences from multiple sources. It is a universal tool, with one parameterization per language. The mixer produces text 'au kilomètre' by removing the difficulty of copyright insofar as no human could easily reconstitute a complete work from it. It remains prohibited to automatically reconstitute and distribute any of the works of the compound, which still fall under proprietary copyright.

The **input of the mixer** is text that can be uploaded online by speaking communities. It must be able to accommodate docx, odf, html, word, and ideally pdf. The input interface consists of two pages. In the first, the applicant completes a minimum form consisting of an identification field:

- last name, first name, or name of the structure
- a drop-down menu for tagging corpuses
- author(s) of the corpus
- dialectal specification of the corpus
- date of writing
- + free field
- two check boxes

With the first checkbox, the depositor certifies to be in possession of the rights, and assigns them under the express condition that the text is only distributed in mixed form to a set of sentences of at least one million words. The reconstruction and distribution of the initial corpus are strictly prohibited. The second checkbox alternatively certifies that the corpus is free of rights.

The input interface also includes a second page where it is possible to enter:

- the list of words in the language containing spaces
- the list of words containing dots (abbreviations)
- the list of proper names

The **internal processing operations** are as follows. The texts are cut into sentences. Each of them is assigned an identifier. Sentences are counted to reach a predetermined threshold, estimated to one million words for example, in order to make it possible to build a parser. Phrases are randomly shuffled between the different sources (text phrase 1, text phrase 14, text phrase 6, etc.). Phrase IDs are kept but will not be distributed.

The **output of the corpus mixer** is a sentence-to-sentence mix of the various deposited corpora. It can be downloaded from the interface in an easily processable format in NLP. It is legally distributable and publicizable to developers. The loading interface includes several fields:

- file format
- tags associated with the corpus to extract

If a particular tag (for example, the dialect #Vannetais Breton) uploads a corpus set too small to operate the mix, a message warns the user.

5.1. A tool for the communities, by the communities

With such a mixing deposit system, individuals, editors of newspapers, or publishing houses, assign their rights to one or more texts, and are automatically mentioned on the deposit interface for the amount of their contribution (for example "*Al Liamm* editions contributed 34,067 words to the common pot for NLP development"). For a publisher, providing their texts is an image success vis-à-vis the speaking community. The shuffling of sentences ensures on the other hand that the distribution networks in bookstores that are already very precarious will not be harmed. Image success is an important factor because minority language print publishers are heavily dependent on public financial support, and such structures may want to ensure that they are perceived by their fund providers as playing collectively. This effect will increase because in the coming years, policy makers will be more and more aware of the importance of the corpus outcomes of their funding choices.

In the local ecosystem around Breton, the contacts made so far react rather positively to such a solution (academics from language departments, institutional language workers, holders of archives of meeting minutes, and even artists producing written corpora who do not seem to see in it a sacrilege to their art). Overall, the idea is received as strange but playful. It quite simply creates emulation: people of different dialects want to make sure that their traditional or standardized dialect, or their choice of spelling, will be well represented. Inasmuch as the web interface could be decently user-friendly, it seems possible to independently carry promotion campaigns to feed the corpus mixer, raising awareness about FAIR practices in the process. The mixed data constitute a reusable corpora that are findable online, in a format that ensures its accessibility and interoperability.

6. Conclusion

Low-resource languages need both quality corpora to compensate for their scarce resources, and FAIR practices to safeguard their ecosystem of NLP development and its bridge to international academic research. This paper has reported on three distinct community internal initiatives to gather such enriched corpora of Breton, at various scales and stages of development. I have presented their potential to contribute to

the engineering of corpus-building ecosystems for low-resource languages.

There may well be simple tools that can dramatically help the speaking communities of low-resource languages to raise adequate resources. These tools will become more equitable and efficient if they answer self-identified needs of the communities, and leverage the internal dynamics within the community.

7. Acknowledgments

Additionnally to the above mentioned actors of Breton NLP development, special thanks are due to the NLP developers who provided their time and efforts to open the discussion, especially Reun Bideault, Gweltaz Duval Guennoc, Johannes Heinecke and Loic Grobol, as well as the actors of the two ANR programs *Autogram* and *Divital* for preminary feedback on the corpus mixer. Thanks also to the Breton community members, experts and workers of the langage who provided feedback from various perspectives, especially Huguette Gaudart, Malo Adeux, Manon Jouitteau, Tudi Kernalegenn, Brendan-Budok Durand-Le Ludec, as well as Cedric Choplin, Stefan Moal, Erwan Hupel (CELTIC-BLM, U. Rennes II) and Annie Forêt (U. Rennes I). Five first year master students of the University of Paris Nanterre have worked so far on the extraction of the data of the wikigrammar, Katharine Jiang, Salomé Chandora, Aurelien Said Housseini (2022), and Yingzi Liu and Yidi Huang (2023). Finally, I wish to thank three anonymous reviewers who provided useful references and helped clarify the paper. English polishing in the paper was done with the help of ChatGPT 3,5 (Open AI) and GoogleTranslate (Google).

8. References

[1] Hicks, Davyth. "Breton – a digital language ?", *The Digital Language Diversity Project*, Erasmus +, 2017.

[2] Nicolas, L. & al. "Creating Expert Knowledge by Relying on Language Learners: a Generic Approach for Mass-Producing Language Resources by Combining Implicit Crowdsourcing and Language Learning", *Proceedings of The 12th Language Resources and Evaluation Conference*, 2020. pp. 268-278.

[3] Millour, A. & K. Fort. "À l'écoute des locuteurs : production participative de ressources langagières pour des langues non standardisées", *Revue TAL*, ATALA (Association pour le Traitement Automatique des Langues), 2018.

[4] Tyers, Francis M. & Nicholas Howell. "Morphological analysis and disambiguation for Breton", *Language Resources and Evaluation*, 2021, pp. 431-473.

[5] Jouitteau, M. "Traitement automatique des langues – Breton", *ARBRES, wikigrammaire des dialectes du breton et centre de ressources pour son étude linguistique formelle*, IKER, CNRS, https://arbres.iker.cnrs.fr/index.php?title=TAL, 2009-2023.

[6] Kerbrat, D. *Ar brezhoneg en oadvezh an niverel, diagnostik ha strategiezh diorren*, [La langue bretonne à l'ère du numérique, diagnostic et stratégie de développement], ms. OPLB, 2021.

[7] Guennec, D., H. Hajipoor, Gw. Lecorvé, P. Lintanf, D. Lolive, A. Perquin, G. Vidal. "BreizhCorpus: a Large Breton Language Speech Corpus and its use for Text-to-Speech Synthesis", *The Speaker and Language Recognition Workshop*, 2022. pp. 263-270.

[8] OPLB, Alan Entem, Brendan-Budok Durand-Le Ludec. 2022. *Traducteur automatique breton-français / français-breton*. https://troer-emgefre.web.app.

[9] Grobol, Loïc. 2022. 'Troer v0', https://huggingface.co/spaces/lgrobol/troer.

[10] Duval-Guennoc, Gw. 2023. "Anaouder, a VOSK model for the Breton language", https://github.com/gweltou/our-voices-model-competition/blob/vosk-br/submit/Open_Category/vosk-br/README.md.

[11] Tyers, Francis M. & Vinit Ravishankar. "A prototype dependency treebank for Breton", *Actes de la conférence Traitement Automatique de la Langue Naturelle*, TALN 2018, pp. 197-204.

[12] Jouitteau, M. (ed.). *ARBRES, wikigrammaire des dialectes du breton et centre de ressources pour son étude linguistique formelle*, IKER, CNRS, http://arbres.iker.cnrs.fr. 2009-2023.

[13] Jouitteau, M. "La linguistique comme science ouverte; Une expérience de recherche citoyenne à carnets ouverts sur la grammaire du breton", *Lapurdum* XVI, Charles Videgain (dir.), 2013, pp. 93-115.

[14] Jouitteau, M. and R. Bideault. "Outils numériques et traitement automatique du breton", Annie Rialland, Michela Russo & Catherine Schnedecker (éds.), Société de Linguistique de Paris. forthcoming. Available at https://hal.science/hal-03918268.

[15] Kahane & al. "Breton Conlls", https://github.com/Autogramm/Breton/tree/main/bretonconlls.

[16] Guibon, G., M. Courtin, K. Gerdes, B. Guillaume. "When Collaborative Treebank Meets Graph Grammar: Arborator With a Grew Back-End", Actes de LREC, 2020.

[17] Éric Le Ferrand, Fabiola Henri, Benjamin Lecouteux, Emmanuel Schang. "Application of Speech Processes for the Documentation of Kréyòl Gwadloupéyen", *The second workshop on NLP applications to field linguistics*, Oleg Serikov; Ekaterina Voloshina; Anna Postnikova; Elena Klyachko; Ekaterina Neminova; Ekaterina Vylomova; Tatiana Shavrina; Eric Le Ferrand; Francis Tyers, May 2023, Dubrovnik, France. hal-04128574.

[18] Jouitteau, M., E. Elfner and F. Torres-Tamarit. "The prosody of Breton dialects and the syntax-phonology interface", IKER, CNRS. 2023. https://doi.org/10.34847/COCOON.3626DBEC-0905-4A59-A6DB-EC09054A59F7.