Outils numériques et traitement automatique du breton

Mélanie Jouitteau, Reun Bideault

Outils numériques et traitement automatique du breton

Mélanie Jouitteau, Reun Bideault
IKER, CNRS, UMR 5478, Université de Pau et des Pays de l'Adour, Université Bordeaux Montaigne

citer comme :
Jouitteau, Mélanie & Reun Bideault. 2023. 'Outils numériques et traitement automatique du breton', Annie Rialland, Michela Russo (dir.), *Langues régionales de France: nouvelles approches, nouvelles méthodologies, revitalisation*, Éditions de la Société de Linguistique de Paris, 37-74, ISBN-10 : 2957089424

Abstract

In this article, formal linguist Mélanie Jouitteau and web developer Reun Bideault present a synthesis of the numeric and NLP tools available or in development for Breton. They discuss the resources for its development. NLP of Breton is still objectively poorly developed, but some new tools have just been made available, which opens a real potential for development. We present a state-of-the-art of the field, and we detail how the first tree bank *Universal Dependencies*, created by Tyers & Ravishankar (2018) could be reinforced by 25000 additional glossed sentences in the databank of the wikigrammar ARBRES (Jouitteau (2009-).

Résumé

Dans cet article, la linguiste formelle Mélanie Jouitteau et le développeur web Reun Bideault présentent un état des lieux des outils numériques et des outils pour le traitement automatique du breton, et discutent les ressources à son développement. Le TAL appliqué à la langue bretonne est encore objectivement peu développé mais de nouveaux outils viennent d'être créés qui ouvrent un potentiel réel. Après un état des lieux de l'existant, nous détaillons comment la première banque d'arbres au format *Universal Dependencies* créée par Tyers & Ravishankar (2018) pourrait être alimentée de 25000 phrases glosées additionnelles provenant de la banque de données de la wikigrammaire ARBRES (Jouitteau 2009-).

1. Introduction

Le chapitre 2 présente un état des lieux des outils numériques disponibles, et est largement redevable aux synthèses pré-existantes de Aubry (2004), Foret & al. (2015), Tyers & Howell (2021:437-438), et pour l'avancement des traductions d'environnement web au rapport de l'Office Public de la Langue Bretonne (OPLB), Kerbrat (2021a,b). Nous avons complété cet état des lieux par une interview de deux chercheurs à l'IRISA, Damien Lolive (3h) pour la synthèse de la voix et Annie Foret (2h) pour la recherche fondamentale, en novembre et décembre 2021. Le chapitre 3 présente les corpus numériques existants, et c'est sous cet angle que nous présentons la wikigrammaire ARBRES (Jouitteau 2009-2021) et évaluons les pas nécessaires à son utilisation pour la construction d'une banque d'arbres en format *Universal Dependencies* pouvant consolider celle de Tyers & Ravishankar (2018). En partie 4, la conclusion synthétise les prospectives et ouvre des pistes de discussion sur les observations des usages, les pratiques de science ouverte et des considérations très concrètes pour leur développement[1].

[1] Nos plus sincères remerciements vont aux chercheurs de l'IRISA Damien Lolive, Gaëlle Vidal et Annie Foret pour le temps qu'ils nous ont consacré, ainsi qu'à Thierry Poibeau (LATTICE, CNRS), Francis Tyers (U. Bloomington, Indiana) et Stefan Moal (U. Rennes II) pour les références fournies, et enfin au centre de formation *Kelenn* à Quimper pour son accueil. L'historique de la genèse de cet article et ses plus récentes mises à jour sont disponibles en ligne dans

2. État des lieux de l'existant

2.1. Traducteurs et outils pour leur construction

Apertium fournit une interface de traduction à partir de l'analyse morphologique de Tyers (2007-2009) et d'un dictionnaire bilingue (cf. Tyers 2009, 2010a, 2010b, 2015). L'analyseur est sous licence GPL-2 (copyright Francis Tyers 2008-2011, Fulup Jakez 2009-2011, Gwenvael Jekel 2011), et disponible sur le site d'Apertium. Tyers (2010a,b) décrit le système de traduction automatique breton > français basé sur des règles.

Tyers & Howell (2021) ont évalué les résultats de l'analyseur morphologique couplé avec un désambiguïsateur morphologique basé sur une grammaire de contraintes. Ces deux derniers outils sont disponibles en logiciel open-source du projet *Apertium* (GNU GPL 3.0). L'analyseur consiste en un transducteur à états finis qui gère l'interface entre les formes de surface et les formes lexicales (les tags morphosyntaxiques et leurs lemmas). Il permet l'analyse de formes comme leur production. Les homophones sont départagés par un ensemble de règles de désambiguïsations morphologiques basé sur une grammaire de contraintes qui a été développée à partir de corrections des traductions automatisées par un brittophone[2] et Francis Tyers. L'Office Public de la Langue Bretonne diffuse une version en ligne du traducteur d'Apertium, *troer emgefre* dans le sens de traduction br > fr. Le choix de l'Office est de ne pas distribuer le sens inverse de traduction avant une perfection des traducteurs vers le breton, car le risque d'utilisation sans correction par des non-locuteurs est grand, et serait très dommageable.

Il existe d'autres projets de traducteurs. Le site Glosbe propose certaines traductions br <-> fr. Le site *Lexicool.com* regroupe les dictionnaires multilingues breton-autre langue. En utilisant la technologie des réseaux de neurones, l'équipe OPUS-MT de l'université d'Helsinki développe un traducteur automatique

Jouitteau (2009- : 'Traitement automatique des langues – Breton'), article qui comprend en plus une description des ressources humaines, des pôles de formation et des ressources de financement pour le TAL du breton.

[2] Ce locuteur n'est pas identifié clairement. Il s'agit peut-être de Fulup Jakez, remercié en note.

multilingue qui comprend un traducteur anglais-breton et breton-anglais.

2.2. Conjugateur de verbes

Le conjugateur automatique de verbes *DVB, displeger verboù brezhonek* développé par Per Morvan est en ligne depuis juin 2021.

2.3. Détecteur de langue

Foret (2018b) relève une méthode pour les langues celtiques dans Minocha & Tyers (2014) et cite deux détecteurs accessibles qui gèrent le breton: open.xerox.com et G2LI.

2.4. Outils correcteurs

Le compte rendu d'activités de l'IRISA (2001) mentionnait déjà qu'il était "désormais possible d'appeler le dictionnaire [vocal] comme outil de correction orthographique, dans une application de type traitement de texte", et il existe un rapport de projet de l'ENSSAT de 2003 sur le correcteur orthographique breton (Petit 2003). Poibeau (2014) qui fournit une formalisation des mutations consonantiques en utilisant des transducteurs à états finis suggérait leur utilisation pour un correcteur orthographique. Il en existe aujourd'hui plusieurs.

Le correcteur orthographique et grammatical *Microsoft Office 2013*, développé par l'association An Drouizig fonctionne aussi sur *MacOffice 2001*. Le correcteur orthographique *Hunspell*, aussi développé par l'association An Drouizig, fonctionne sur *Adobe Indesign, Firefox, LibreOffice* et *OpenOffice.org* et *MacOSX*. Le correcteur grammatical pour la suite bureautique *LibreOffice* développé par Dominique Pellé avec l'aide le l'OPLB utilise *LanguageTool*, testable en ligne. L'OPLB rapporte une première version de 400 règles, avec repérage des fautes de mutation. Ce correcteur est évalué dans Morvan (2019).

2.5. Dictionnaires en ligne

Il existe de nombreux dictionnaires en ligne du breton contemporain, et un agrégateur de dictionnaires. Certains sont en accès libre, mais peu sont sous licence libre. Menard & Bihan (2016-) et Favereau (1993) comportent des entrées de dialectes traditionnels mais les autres sont plutôt de breton standard.

L'aggrégateur *Geriafurch* développé par Anthony Lannuzel croise les résultats de plusieurs dictionnaires en ligne et en livre un résultat allégé. Il existe en application téléphone téléchargeable. Sa portée, en 2021, couvre le dictionnaire Brezhoneg21 = KAG (2016), ressource scolaire des sciences et techniques, le dictionnaire *Devri* de Menard & Bihan (2016-), le dictionnaire en ligne de Favereau (1993), celui de Glosbe, *Preder* et finalement *Termofis*, le dictionnaire terminologique de l'OPLB.

Le dictionnaire breton-néerlandais de Jan Deloof (2008-2010) comporte 40,000 entrées. Kevin Donnelly, qui a géré sa mise en interface, considère qu'il s'agit du plus grand dictionnaire libre (GPL) pour une langue celtique (Donnelly 2010).

Le dictionnaire Favereau (1993) comporte 40 000 entrées. La première synthèse de la voix de l'IRISA avait utilisé un algorithme pour en accepter les orthographes multiples. Il n'est pas en licence libre, raison pour laquelle Tyers ne l'utilise pas (Tyers & Howell 2021:440, fn11).

Le dictionnaire Freelang fr <-> br (disponible en ligne ou téléchargeable) de Tomaz Jacquet comporte 37.800 entrées. Tyers (2009) en a importé semi-automatiquement les classes lexicales.

Le dictionnaire br -> fr de Cornillet (2017) est disponible en ligne dans une version corrigée en 2020. Il a été utilisé pour la synthèse de la parole en 2019-2021 à l'IRISA.

Le dictionnaire de l'association *Stur* traduit 22.302 noms du français vers le breton. Il est cherchable en ligne.

Le dictionnaire Favereau (2016-évolutif) est en ligne sous format pdf, avec des dossiers séparés pour chaque lettre initiale. Le copyright propriétaire mentionné sur le site est de 2016, mais l'auteur enrichit l'ouvrage régulièrement et met en ligne les pdfs par lettre du dictionnaire. La date de dernière modification pour chaque dossier est au début de chaque pdf.

La base de données toponymique KerOfis de l'OPLB liste les noms propres noms de lieux.

Le dictionnaire multilingue *Logos* comprend le breton. Il s'agit d'un site collaboratif de traducteurs professionnels sur invitation.

Le dictionnaire multilingue *wiktionnary* comprend le breton avec *wikeriadur*.

2.6. Grammaires en ligne

Le site ARBRES (Jouitteau 2009-) offre une wikigrammaire du breton, que nous prenons le temps de décrire ici brièvement car elle informe la constitution de sa base de données, discutée en partie II.

Il s'agit d'un carnet de recherches rédigé sous forme de grammaire en ligne. Ses buts sont de fournir une description fine et théoriquement informée de la variation syntaxique en breton et un état des lieux permanent et à jour des différentes recherches en syntaxe formelle. Le projet est de créer un pont entre le milieu international de recherches linguistiques, les travailleuses et travailleurs de la langue qui cherchent plutôt une ressource pédagogique, et tout adulte et citoyen curieux de la structure de cette langue parlée par plus de 150 000 locuteurs.

Le site propose plusieurs entrées, une grammaire descriptive standard ainsi qu'une grammaire formelle qui organise une description de leur impact théorique pour la linguistique formelle. Il comporte une bibliographie générale qui se veut exhaustive pour les recherches en syntaxe, et comporte une centrale d'élicitations par laquelle la communauté internationale de recherche peut co-construire des protocoles avec Mélanie Jouitteau qui opère l'élicitation sur le terrain et poste les résultats en ligne. Fin 2021, la wikigrammaire rassemblait plus de 2000 articles thématiques. Le site est ouvert en écriture et pour les commentaires (pour l'aspect science ouverte et science citoyenne de la wikigrammaire, se reporter à Jouitteau, M. 2013b). L'OPLB a été consulté dès 2008 afin de recueillir ses vœux en termes de développements, vœux qui ont influencé la genèse de la wikigrammaire, en particulier la constitution en format récupérable pour une base de données utilisable en TAL.

Il faut ajouter à cette grammaire en ligne les ouvrages dédiés à des parlers locaux particuliers. La partie grammaticale du blog de collecte *Brezhoneg Bro-Vear* (Yekel, Georgelin & Ar C'hozh 2015-2021) est maintenant considérable. Une ressource non négligeable provient aussi des plus récentes thèses et monographies universitaires dont les textes sont disponibles en ligne. Celles-ci sont recensées dans l'inventaire des grammaires de la wikigrammaire ARBRES.

2.7 Traduction de logiciels, réseaux sociaux, jeux, etc.

Diverses applications utilisables sur internet sont traduites, souvent partiellement, en breton mais cela reste insuffisant pour créer un environnement informatique immersif. Pour le web, les interfaces utilisables sont extrêmement limitées en nombre et en pourcentage de traduction. *Wordpress* est le système de gestion de contenu (abrégé CMS pour l'anglais *Content Management System*) le plus utilisé dans le monde (40% des sites). Il n'est traduit, pour la version plus récente fin 2021 (V. 5.8.x), qu'à 18 %. Ce travail est suivi par 7 personnes. Pour comparaison, les versions en basque sont traduites à 96 %, en occitan à 53 %, pour respectivement 80 et 13 participant.e.s. Ce CMS s'appuie sur des plugins indispensables à une utilisation élargie, où le niveau de traduction est encore plus faible lorsqu'il existe. Reun Bideault, développeur web, considère que l'exemple de *Wordpress* est actuellement généralisable à tous les outils web libres et propriétaires, raison pour laquelle les professionnel.le.s du web ne peuvent actuellement fournir un produit fini et surtout évolutif permettant de travailler en breton à un coût supportable.

L'OPLB fournit la traduction en breton des données du CLDR (*Common Locale Data Repository*) d'Unicode, qui regroupe l'ensemble des paramètres régionaux à destination des applications informatiques. Lors de la publication de la version 38 du CLDR fin 2020, Kerbrat (2021a,b) estime que le breton a atteint l'avant dernier niveau de couverture (*Moderate++*). Tomaz Jacquet rend disponible en ligne sous différents formats un dictionnaire trilingue

breton, français anglais du vocabulaire utilisé dans les logiciels. Fin 2021, sont disponibles en environnement traduit :

- une suite bureautique (*LibreOffice*) qui est associable aux correcteurs orthographiques et grammaticaux décrits plus haut
- un logiciel pour la navigation web (*Firefox*)
- un logiciel pour l'échange de courriels (*Thunderbird*)
- quelques logiciels multimédia

(*VLC* pour la vidéo, *Clementine* pour la musique)

- quelques logiciels d'édition graphique

(*Inkscape, Gimp, Tuxpaint*)

Pour les réseaux sociaux, *Facebook* est utilisable en breton depuis 2014 (Ar Mogn 2015). *Mastodon*, réseau semblable à *Twitter* mais libre de droits, fait actuellement l'objet d'un projet de traduction participative.

Il existe une version bretonne pour quelques applications smartphone, en plus de l'autocorrection et la prédiction de mot en breton sur le clavier virtuel Microsoft *SwiftKey*:

- *Firefox* (iOS et Android), navigateur web
- *K-9 mai*l (Android), client de messagerie
- *Vanilla Music* (Android), lecteur musical

2.8. Synthèse vocale

La Région Bretagne à travers l'OPLB a financé à hauteur de presque 200.000 euros la construction d'un moteur de synthèse de la parole (breton KLT standard[3], un homme, une femme). Le projet était dirigé en TAL par Damien Lolive et Gwénolé Lecorvé du laboratoire *Expression* de l'ENSSAT à Lannion en collaboration avec la maison d'édition *Skol Vreizh*. Le programme de synthèse de la voix a été livré à l'OPLB en mars 2021.

Les deux locuteurs qui ont prêté leur voix ont été élevés en milieu brittophone trégorrois, à tendance plus standard pour Annaig Kervella (fille de Frañsez Kervella, auteur de la grammaire

[3] Les trois initiales bretonnes KLT réfèrent aux dialectes cornouaillais, léonard et trégorrois, ce qui exclut le dialecte vannetais. Les propriétés originales des trois dialectes majeurs du KLT sont gommées pour obtenir un standard d'usage entre les trois.

standard de référence), et plus traditionnelle pour Pascal Lintanf (avec influences léonardes pour ce dernier)[4]. Chacun des deux corpus oraux produits durent un peu plus de 20h, ils ont été constitués par tâche de lecture d'un corpus de breton standard constitué principalement de discours journalistique, et de textes littéraires (environ 10% sont des dialogues, joués avec expressivité modérée). Le corpus écrit correspondant a été normalisé (écriture en lettres des nombres et acronymes, prononciation différenciée des noms propres, etc.) puis, un panel d'experts choisi par *Skol Vreizh* et principalement le second locuteur Pascal Lintanf ont constitué un répertoire de règles de prononciation. Un lexique donnant une prononciation standard accentuée en API a été constitué par arbitrage entre plusieurs sources de lexiques phonétisés et écrits en orthographe unifiée : le dictionnaire *An Here* de Menard & Kadored (2001), le dictionnaire de Francis Favereau (2015) et sa dernière version consultable en ligne Favereau (2016-évolutif) consultables en ligne, et de Gérard Cornillet (2017). D'autres données y ont été intégrées comme celles des noms propres, fournies par l'OPLB, et celles rencontrées dans les corpus constitués. Pour dix mois, Gaëlle Vidal, ingénieure d'études, a défini et enregistré un corpus de textes, sélectionné les locuteurs, et procédé aux enregistrements et à leur découpage en phrases. Hassan Hajipoor, ingénieur de recherche, a ensuite eu 18 mois (dont un confinement) pour construire un phonétiseur, comprenant un modèle de la syllabe et de l'accentuation qui a pu être paramétrisé pour les exceptions, et entraîner un réseau de neurones sur le corpus oral et le dictionnaire. Le système en end-to-end livre le fichier son à partir de la phrase écrite. La technique ne permet pas de prendre en charge la structure informationnelle et la prosodie associée, mais l'accentuation de mot et les phénomènes

[4] Pascal Lintanf est par ailleurs l’auteur d’un mémoire universitaire sur la phonétisation du breton (An Intanv 1994).

de frontière de mot comme la mutation ou le sandhi sont pris en charge[5,6].

L'OPLB a créé un poste de chargé de développement du numérique pour sa diffusion.

3. Corpus numériques existants

La langue bretonne n'est pas une langue minorisée pour laquelle manquent les corpus, mais ils ne sont pas tous immédiatement accessibles pour des traitements automatiques de la langue (copyright restrictif, éditions épuisées, documents non-OCR, corpus numériques à URLs non-stables, etc.). Ci-dessous, sont listées les ressources a-priori disponibles au TAL, ou déjà utilisées[7,8].

3.1. Corpus non-glosés

Thierry Poibeau (c.p.) signale 23 Mo de données brutes de texte en breton, sans annotations, dans le corpus Oscar, qui sert

[5] Pour un historique détaillé de l'époque pionnière de la synthèse de la voix dans les années 90 avec Favereau, IRISA & TES. 1999, se reporter à Aubry (2004). Il semble aussi avoir existé un correcteur de prosodie (Mocquard 1999, 2001, Guillou 2000) et un entraîneur prosodique (Aubry 2000, 2004).

[6] L'OPLB, pour le projet de synthèse de la voix, n'a pas recouru à son conseil scientifique.

[7] Jouitteau (2009-:'corpus') fournit une liste plus exhaustive de corpus de breton, plus tournés vers l'apprentissage humain.

[8] Leixa & al. (2014) ont essayé de recenser les corpus utilisables en TAL pour plusieurs langues minoritaires de l'État français. L'approche est un brin parachutée. Ils comptent pour le breton 420 corpus utilisables, dont 403 corpus oraux et 17 corpus textes. "On trouve parmi ces ressources de petits enregistrements audio de quelques minutes, mais également d'importants corpus alignés pouvant servir de base à des technologies de la langue. Parmi les ressources audio, nous avons par exemple les enregistrements effectués par M. Jean Le Dû lors d'une enquête dialectologique réalisée en Bretagne, en vue de constituer le *Nouvel Atlas Linguistique de la Basse-Bretagne*" (Le Dû 2001). Cependant, à l'écoute, ces enregistrements sont difficilement utilisables car les élicitations sont effectuées à partir de gestes physiques dénotant des mots à trouver, or cette information gestuelle manque évidemment aux enregistrements. La prosodie interrogative des locuteurs est typiquement celle de quelqu'un qui cherche à deviner un mot, et la confirmation que son choix est le bon. L'identification précise de l'ensemble des corpus listés dans Leixa & al. (2014) "est disponible sur le CD qui est joint au rapport" à la DGLFLF.

actuellement pour mettre au point des modèles pour le TAL par modèles neuronaux (type *Bert*).

Les archives de traduction de l'Office constituent un corpus bilingue qui a déjà été utilisé pour le traducteur automatique (Tyers 2009). Ar Mogn (2015:15m40s), co-directeur de l'OPLB, mentionne un corpus de 43000 phrases bretonnes traduites. Kerbrat (2021a,b) l'estime à "environ 1 million de mots". Le corpus de traductions de l'OPLB, corpus de phrases en breton, et corpus de phrases en français, sont téléchargeables et libres de droit.

L'association *An Drouizig* revendique pour la construction de son correcteur orthographique *Difazier [ver 4.4]* l'analyse d'un corpus linguistique de 20 millions de mots bretons, qui comprend au moins celui de l'OPLB.

Donnelly (2010) mentionne sa création avec l'aide de Rhisiart Hincks à Aberystwyth d'un corpus parallèle de 3500 phrases en breton et gallois, organisé dans un *Breizh-Llydaw Sentence Ban*k (license GPL), et accompagné d'un dictionnaire de 1200 mots.

Il existe aussi des corpus parallèles multilingues, comme la *Déclaration des Droits Humains* de l'OHCHR et la traduction du *Petit Prince* de Saint-Exupéry.

3.2. Corpus sonores

On a vu que l'IRISA à Lannion a constitué un corpus de plus de 40h de la synthèse de la voix dont les phrases ont été individuées.

Il existe aussi différents sites de collecte de données brutes, par des collectifs associatifs à la durée de vie variable. Ces derniers n'en sont pas pour autant négligeables. Ils constituent des travaux considérables, avec traductions des données dialectales ou explicitation en standard, et des traductions en français. A notre connaissance, il n'existe aucune aide ou soutien organisée à ces travaux pionniers, même pour l'hébergement.

- les *Dictionnaires bretons parlants*

(Cheveau & Kersulec 2012-évolutif)

- la *Banque sonore des dialectes du breton*

(Desseigne & al. 2013-2018)

- *Brezhoneg Bro-Vear*

(Yekel, Georgelin & Ar C'hozh 2015-2021)

Common voice de *Mozilla* a lancé en 2018 un module de collecte de la parole en crowdsourcing, qui permet aux utilisateurs d'enregistrer leur propre parole, ou d'évaluer les enregistrements laissés par d'autres (9h d'enregistrements validés en 2021).

Les enregistrements audio de corpus libres existent dans les différents dialectes du breton, stockés dans les archives des différentes radios bretonnes, sous des formats différents allant de l'analogique au numérique. Les fichiers audio des enquêtes du *Nouvel Atlas Linguistique de la Basse-Bretagne*" (Le Dû 2001) devraient pouvoir être au moins partiellement utilisées, mais cela demanderait un tri méticuleux (voir note de bas de page numéro 8).

3.3. Banque d'arbres Universal Dependencies

Il existe pour le breton des corpus glosés traduits. Ils comportent des phrases en breton traduites mot-à-mot et traduites globalement. La traduction mot-à-mot est une glose, qui contient des informations sur l'élément linguistique en question (catégorie grammaticale, fonction, mutation déclenchée, etc.).

La notation universelle qui émerge en 2021 comme recommandation pour les banques d'arbres est celle de *Universal Dependencies* ("format UD"), qui propose un jeu de 17 parties du discours (*parts of speech*, POS) et deux douzaines de fonctions grammaticales. Certains des choix fondamentaux de ce format, comme de subordonner les catégories fonctionnelles aux catégories lexicales ne sont pas soutenus linguistiquement, mais la conversion de structures UD à des structures syntaxiques en constituants est cependant automatisable en grande partie (voir discussion par Osborne & Gerdes 2019).

Tyers & Ravishankar (2018) ont créé la première banque d'arbres réalisée au format UD. Ce corpus tree-bank breton de 10 000 tokens est annoté manuellement. L'analyseur morphologique de Tyers (2009) pour *Apertium* a été utilisé pour la tokenisation et l'annotation morphologique. Ci-dessous, un exemple de codage de la banque d'arbres de Tyers & Ravishankar (2018). On y trouve la phrase bretonne en entier et sa traduction en français, puis une glose mot-à-mots, répartie en lignes. On trouve le lemma (forme comme donnée comme pour un dictionnaire), les étiquettes de

parties du discours (POS tags) catégorielles et leurs sous-spécifications, des informations sur la structure choisie qui fait dominer le verbe lexical (noté *root*), ainsi que les traits de la morphologie flexionnelle.

```
# sent_id = apertium.vislcg.txt:1:0
# text = N'int ket aet war-raok.
# text[fra] = Ils n'ont pas progressé.
# labels = to_check
1 N'   ne    ADV adv Polarity=Neg  4 advmod _ SpaceAfter=No
2 int   bezañ  AUX  vblex  Mood=Ind|Number=Plur|Person=3|Tense=Pres|VerbForm=Fin 4 aux _ _
3 ket  ket   ADV adv    _    4   advmod _   _
4 aet  mont  VERB vblex Tense=Past|VerbForm=Part  0   root  _ _
5 war-raok war-raok ADV  adv  _ 4    advmod  _   SpaceAfter=No
6 .  .  PUNCT  sent   _   4    punct   _  _
```

Le texte du corpus consiste en 888 phrases provenant d'exemples de grammaires, de phrases tirées de *wikipedia* en breton, de textes administratifs de l'OPLB et de textes du journal *Bremaik*, avec deux chansons traditionnelles. La composition est détaillée dans Tyers & Howell (2021:450). Certains codages sont étranges, et pourraient être discutés. Par exemple dans les phrases *En em c'houlenn a ran* /se demander R fais/ ‘Je me demande’ ou *En em gannet out c’hoazh?* /se battre es encore/ ‘Tu t’es encore battu?’, le pronom proclitique réfléchi *en em* est noté / det + aux /, alors qu'en format UD les réfléchis et réciproques sont étiquetés comme des pronoms (PRON).

La banque d'arbres a d'ores et déjà servi à une expérimentation pour la construction de grammaires de dépendances afin de construire un outil de lecture augmentée (Martinet 2021), et à évaluer un analyseur morphologique et la grammaire de contraintes dans Tyers & Howell (2021:450).

3.4. Une banque de données en wikigrammaire

La wikigrammaire ARBRES Jouitteau (2009-) est sous licence creative commons CC BY-NC-SA. Elle a été prioritairement développée pour un lectorat humain en ligne, mais comporte une masse importante de données du breton localisées par leur dialecte, traduites et glosées mot à mot, organisées dans un format numérique qui est destiné à terme à alimenter un traitement automatique.

Fin 2021, le site contient plus de 75 000 tableaux de type "prettytable" qui ont servi à aligner chaque mot breton breton avec sa glose en français, et à aligner l'ensemble avec une traduction globale de la donnée en français, ainsi qu'avec une typification dialectale du locuteur source. En moyenne, nous estimons que chaque donnée originale en breton a été employée trois fois dans des endroits différents de la grammaire, ce qui donne une estimation grossière de 25 000 phrases originales en breton.

La wikigrammaire utilise des exemples tirés de plus de mille ouvrages de recherche scientifique sur le breton, des données de séances d'élicitation avec des locuteurs natifs effectuées par Mélanie Jouitteau, à son initiative ou à la demande d'autre linguistes, et dont les résultats bruts sont disponibles en ligne dans la centrale d'élicitation avant exploitation, ainsi que de 399 sources de corpus écrits différents, du vieux breton aux dialectes bretons modernes, breton standard y compris. Les dialectes y sont mentionnés comme tels, et la typification dialectale est associée à chaque donnée, donc il serait possible pour un traitement automatique de mettre de côté les états anciens de la langue, et les quelques données comparatives tirées de langues autres (hébreu, basque, occitan, etc.). Les données du breton ont servi à l'établissement d'une grammaire donc elles ont été sélectionnées pour représenter la plus grande variété possible de structures. La graphie est riche car l'orthographe des sources diverses a été respectée - les gloses, elles, sont en orthographe unifiée *peurunvan*. Certaines données de ARBRES, plutôt rares à l'échelle du corpus, comportent en plus une ligne de code donnant la tokenisation de la donnée en API, ou dans des orthographes originales renseignant la prononciation. Ces scripts peuvent être

mis de côté car ils sont signalés par une balise de mise en couleur verte (< (/) font color=green >). Les traductions en français viennent soit de corpus déjà bilingues, soit sont effectuées par Mélanie Jouitteau (native français, breton L2).

Pour comparaison avec le treebank UD de Tyers & Ravishankar (2018), je code ci-dessous le même exemple donné plus haut, dans sa forme visible aux utilisateurs et le code que cela nécessite. Le codage dans ARBRES obtient une visualisation comme ci-dessous pour les utilisateurs.

(1) N'int ket aet war-raok.
ne[1] sont pas allé sur-avant
'Ils n'ont pas progressé.'
Dialecte, source référencée de la donnée

Chaque exemple est donné en breton, glosé et traduit. La ligne de gloses fournit la traduction littérale, mot-à-mots en français. Elle comprend une mention des mutations consonantiques en superscript sur son élément déclencheur (ici, l'adverbe négatif *ne* qui provoque une lénition dans tous les dialectes, codée 1 en superscript. La mutation est notée même si, en l'occurrence, elle ne peut pas avoir ici d'effet car l'initiale du verbe qui suit n'est de fait pas mutable. Les gloses en français ne montrent d'accord que si l'élément en breton en montre (cf. *allé*). Parfois, le glosage au plus près de la composition bretonne crée en français des approximations (cf. *sur-avant*). La troisième ligne visible du tableau fournit la traduction globale de la phrase en français standard. Pour obtenir une telle visualisation, alignement des gloses compris, le code wiki est comme ci-dessous (abstraction faite de la balise diu superscript de mutation).

```
0 {| class="prettytable"
1 |(1)|| N'int || ket || aet || war-raok.
2 |-
```

```
3 ||| [[ne]][[1]] [[COP|sont]] || [[ket|pas]] || [[mont|allé]] || [[war-raok|sur-avant]]
4 |-
5 |||colspan="10" |'Ils n'ont pas progressé.'
6 |-
7 ||||||||colspan="10" |*Dialecte*, source référencée de la donnée
8 |}
```

Dans le code, les colonnes (||) de la première ligne fournissent un premier découpage de la donnée bretonne. Cette ligne comporte la ponctuation. Le découpage y est inégal, souvent prosodique car les éléments marqués d'une apostrophe ou d'un tiret n'y sont qu'exceptionnellement séparés. Il découpe aussi parfois des blocs de constituants syntaxiques. La seconde ligne visible pour l'utilisateur est la ligne 3. C'est la ligne de gloses qui fournit une tokenisation plus fine et la lemmatisation. Avec l'exemple de la négation et de sa copule, on voit que le découpage en double crochets dessine alors les sous-parties du découpage de la première ligne. Les tokens atomiques sont séparés, les clitiques y sont ainsi séparés de leur hôte.

Pour que les gloses soient cliquables pour les utilisateurs, le script wiki nécessite que chaque traduction mot-à-mot, la glose, soit associée à une adresse d'article dans la grammaire. Dans la syntaxe wiki, ce script est ordonné comme suit: [[adresse du lien| glose]]. C'est ainsi que grâce à un script [[mont|allé]], l'utilisatrice qui clique sur la glose *allé*, visible pour elle juste sous le mot breton *aet*, ouvre la page du site dédiée au verbe *mont* 'aller'. Ce script, pour un format UD, fournit le lemma. Ce lemma est associé à la traduction française du token aligné en colonne avec lui. Dans le cas de la préposition composée *war-raok* /sur-avant/ 'en avant', un seul lemma lui est associé.

```
0 {| class="prettytable"
1 |(1)|| mot 1' mot 2 || mot 3 || mot 4 || mot 5-mot 6.
2 |-
```

```
3 ||| [[lemma breton 1|français pour lemma 1]][[mutation déclenchée]] [[lemma 2|français pour lemma 2]] || [[lemma 3|français pour lemma 3]] || [[lemma 4|français pour lemma 4]] || [[lemma 5|approximation française pour lemma 5]]
4 |-
5 |||colspan="10" |'Traduction de la phrase en français.'
6 |-
7 ||||||||colspan="10" |*Dialecte*, source référencée de la donnée
8 |}
```

Le lemma breton est donné sous sa forme non-dérivée, ce qui signifie dans cette langue celtique que le lemma est donné au singulier pour un nom comptable mais au pluriel pour un nom collectif. Pour la flexion verbale, le lemma donné est, par convention, la forme infinitive dans la wikigrammaire comme dans UD. Il y a une petite divergence avec le format UD pour les formes qui ont des racines supplétives au comparatif de supériorité comme *gwell* ou *gwelloc'h* 'mieux', ou *gwazh* ou *gwashoc'h* 'pire'. UD recommande de leur assigner le lemma non-comparatif ce qui donnerait *gwelloc'h* 'mieux' > [[mat|bien]].[[-oc'h|plus]] et *gwasoc'h* 'pire' > [[fall|mal]].[[-oc'h|plus]], alors que la wikigrammaire a prévu de dédier un article à chaque racine irrégulière, ce qui est géré pour l'instant par des redirections ([[-oc'h|mieux]]). Ce pourrait être régularisé assez facilement.

UD requiert que les lemmas soient fournis sous la forme de surface canonique, ce qui pose le problème des formes ambiguës, concrètement en breton les verbes infinitifs et les noms déverbaux, ainsi que les noms différenciés par leur genre en situation (*pal, ar pal* 'le but', *pal, ar bal* 'la pelle' ou *taol, an taol* 'le coup', *taol, an daol* 'la table'). Dans le dictionnaire en ligne Menard & Bihan (2016-), ces ambiguïtés sont prises résolues par un système de spécifieurs numériques assez régulier (*pal.1*, *pal.2*) mais le format UD recommande de privilégier les formes de surface comme lemmas. UD propose de classer ces homonymes dans la colonne MISC dans l'attribut optionnel LId (LId=can-1). Le désambiguïsateur morphologique de Tyers & Howell (2021) semble pouvoir se charger des homophones. Ce dernier pourrait

peut-être être solidifié par la liste des pages de désambiguïsation qui liste dans la wikigrammaire les suffixes pouvant être ambigus.

Les mots fusionnés sont un ensemble de plusieurs mots syntaxiques qui apparaissent en breton comme un mot opaque. Ils sont traités en ligne de glose comme des tokens distincts reliés par un point. Ainsi, la préposition *e* devant un article défini est notée *en* en ligne 1 et glosée : [[P.e|dans]].[[art|le]]. La plupart des prépositions peuvent recevoir un pronom objet incorporé - on les appelle les prépositions fléchies. La préposition fléchie *ennon* 'en moi' est glosée [[P.e|dans]].[[pronom incorporé|moi]]. La préposition *ganin* 'avec moi' est glosée [[gant|avec]].[[pronom incorporé|moi]], ce qui permet de récupérer deux formes différentes de pronom incorporé 1SG: *-in* et *-on*, et d'associer chacune avec la préposition qui la déclenche.

En ligne de glose, le découpage en tokens descend au niveau morphologique dans la mesure où le permet son lectorat prioritairement humain. Celui-ci a témoigné régulièrement d'une difficulté d'accès à des formes trop décomposées, ou à des abréviations linguistiques pourtant communes de type 3SG, 3PL. Ces abréviations sont la plupart du temps évitées dans ARBRES, et la dérivation morphologique est inégalement prise en charge dans les gloses de la wikigrammaire. Lorsqu'un seul suffixe est repérable, le découpage donne directement le suffixe en question dans la glose, mais lorsque plusieurs suffixes forment une finale complexe, le lemma donné est directement cette finale complexe. Le nom *distresadur* 'transformation' est glosé [[di-, dis-|trans]].[[tres|form]].[[-adur|ation]]. C’est dans la page de la finale complexe *-adur* que la finale est décomposée dans ses différents suffixes. Le système de catégorisation de pages permet de générer automatiquement la liste des finales complexes et la liste des suffixes répertoriés dans le site.

La dérivation flexionnelle est prise en charge pour les pluriels des noms. Pour les pluriels simples, le morphème pluriel final apparaît séparé d'un point. Ainsi, le nom pluriel *krouadurioù* 'enfants' est glosé [[krouadur|enfant]].[[-ioù (PL.)|s]]. En breton, les pluriels dits "pluriels internes" ont la propriété de modifier leur racine. Le nom pluriel *bugale* 'enfants' est glosé [[bugel|enfant]].

[[pluriel interne|s]], avec le lemma qui est la forme de surface au singulier, et le pluriel qui renvoie l'utilisateur à la page sur les pluriels internes. Les morphèmes porte-manteaux de la flexion verbale, les traits de conjugaison, ne sont pas non plus donnés en glose. Ces traits de flexion verbale sont calculables par la traduction française associée, qui, elle, est donnée fléchie dans les gloses. Les traits UD (UD features) sont donc récupérables dans la mesure où la morphologie verbale française est assez riche. La matrice de traits "Mood=Ind|Number=Plur|Person=3|Tense=Pres|VerbForm=Fin" du verbe breton *int* 'sont' peut-être récupérée par la glose en français *sont*. Cette carence dans la glose de ARBRES pour la flexion verbale pourrait en principe aussi être supplée par les données de DVB, *displeger verboù* développé par Per Morvan.

Un cas difficile et intéressant est posé par la tempête de variation morphologique (et syntaxique) dans le verbe et auxiliaire 'avoir'. En (2), ce verbe précédé de la négation *ne* sous sa forme proclitique est orthographié de manière discontinue, *o dez*. Il comporte les traits du sujet interprété sur sa gauche avec un pronom 3PL *o* sous une forme qui semble oblique, puis d'une initiale /d-/ typique des personnes 3 (au singulier comme au pluriel; *en de(v)ez* 3SGM, *he de(v)ez* 3SGF, *o de(v)ez* 3PL). La racine marque la trace de la forme dite d'habitude, qui n'est pas interprétée ou produite dans toutes les variétés sur ce verbe. On pourrait, dans le même contexte syntaxique, trouver *n'o deus ket* en breton standard, la notion d'habitude étant convoyée par un présent à lecture générique. Ceci implique qu'un glosage précis nécessite d'être en mesure de vérifier pour chaque variété si le morphème comprend réellement ces traits, en syntaxe comme en sémantique. Enfin, la finale pourrait être, selon les analyses, une racine dénuée à sa droite de morphème d'accord, un accord 3SG réalisé avec un élément qui n'est pas le sujet, ou encore un morphème d'accord par défaut qui ne fait qu'emprunter la morphologie 3SG et qui apparaît lorsque le sujet est exprimé ailleurs (se reporter aux analyses formelles du système d'accord).

(2)	Ha	forzh	boued	n'o dez		ket...
	et	beaucoup	nourriture	ne 3PL	3.a	pas

'Et ils n'ont pas beaucoup de nourriture.'

Vannetais, Herrieu (1994:90)

Ce problème n'est pas facilement écartable car certains dialectes centraux ont, de toute façon, pour une sous-partie du paradigme, un morphème d'accord à droite du composé (*memp* 'nous avons'), dialectes dans lesquels peuvent exister en plus des règles d'accord différentes (*ni meump* /1PL 1.racine.1PL/ vs. *ni neus* /1PL 3.racine.3SG/ ou /1PL 3.racine.Ø/, 'nous avons'). Les gloses dans la wikigrammaire reflètent la diversité des données au plus près de ce qu'on en comprend scientifiquement, et cela peut être un frein à la conversion automatique. Les buts d'un traitement automatique peuvent nécessiter de faire abstraction de la variation et de se contenter de stocker les formes diverses en lien avec leur traduction française.

Les traits de tous les types de pronoms sont récupérables en glose. Le pronom fort indépendant (pfi) 1SG *me* 'moi' est noté en glose [[pfi|moi]], Le pronom fort indépendant 2SG *te* 'toi' est noté en glose [[pfi|toi]], etc. De même, le déterminant possessif (POSS) *ma* 'mon, ma', qui déclenche une mutation mixte (codée 2 en superscript), est glosé [[POSS|mon]]^{[[2]]} dans la plupart de ses occurrences. Cependant, comme le site documente la variation dialectale, les occurrences du cornouaillais de Locronan documentées dans la grammaire, où ce possessif déclenche une lénition (codée 1 en superscript), sont glosées [[POSS|mon]]^{[[1]]}.

La morphologie flexionnelle n'impacte qu'exceptionnellement les adjectifs bretons par suffixation (*mezvez* 'saoule', glosé [[mezv|saoul]].[[-ez (F.)|e]]). Cependant, la qualité, présence ou absence de mutation sur l'adjectif renseigne sur les traits du nom qu'il modifie. En ligne de glose, la traduction de l'adjectif en français révèle les traits obligatoirement interprétables: *an hini vrav* la belle' est glosé :

[[art|un]] [[hini|celui]]^{[[1]]}[[brav|belle]].

Cet exemple permet aussi de noter que les rares éléments qui n'ont pas d'équivalent en français comme la tête nominale

sémantiquement générique *hini* sont traduits en glose par une approximation qui a été jugée commode par le lectorat humain.

On a vu que la ligne de gloses comprend, balisées en superscript (<(/)sup>), les mutations morphosyntaxiques associées à chaque élément qui les déclenche. on marque par le chiffre 1 pour la lénition, 2 pour la spirantisation, 3 pour la mutation durcissante, 4 pour la léniprovection et 5 pour la mutation réduite. Les consonnes épenthétiques du breton sont marquées +C en superscript dans la glose. Il arrive que le découpage morphologique d'un mot breton nécessite de mentionner une consonne épenthétique dans la glose en français. Elle est alors écrite, et non-cliquable puisque ne correspondant à rien en breton (*kozhni* 'vieillesse' est glosé [[kozh|vieil]].l.[[-ni, -oni|esse]]).

Le format UD comporte en tout 17 étiquettes de parties du discours (POS tags). Le code de la wikigrammaire ne fournit qu'exceptionnellement la catégorie grammaticale des éléments directement en glose. Les 5 formes du verbe 'être' et la variation dialectale de leur distribution ont nécessité dans la grammaire un glosage hybride, parfois morphologique (*eo*, *a* zo, *emañ*, *ez eus*, *vez*), parfois syntaxique (COP renvoie à l'article sur l'emploi syntaxique de la copule) ou même sémantique (le signe E en adresse renvoie à l'article sur la copule existentielle). La catégorie des éléments est cependant toujours récupérable automatiquement par les catégorisations de pages (*eo* => auxiliaire, car l'article de la wikigrammaire intitulé *eo* est catégorisé dans le site comme une page concernant un auxiliaire. Tous les éléments sont ainsi catégorisés via la page qui leur est dédiée, par exemple les adjectifs, mais aussi avec une granularité plus fine dans la mesure où ils ont un comportement grammaticalement distinguable, les adjectifs de couleur (voir la liste des catégories).

Ci-dessous, j'inventorie les catégorie UD et je détaille pour chacune les équivalences sur la wikigrammaire, en ajoutant une estimation des nombres de membres de chaque catégorie fin 2021. Ces chiffres vont progresser à l'avenir, surtout pour les catégories lexicales, au fur et à mesure que des exemples nouveaux alimenteront la grammaire.

- ADJ = adjectif. Ils sont listés dans la wikigrammaire dans la (238 membres), auxquels on ajoute les numéraux ordinaux, les participes (une partie sont mentionnés en glose par la dérivation du suffixe *-et*).
- ADP = adposition (préposition et postposition). Ils sont listés dans la wikigrammaire dans la (158 membres) et dans la (11 membres)
- ADV = adverbe. Ils sont listés dans la wikigrammaire dans la (219 membres)
- AUX = auxiliaire. Ils sont listés dans la wikigrammaire dans la (18 membres)
- CCONJ = conjonctions de coordination. Ils sont listés dans la wikigrammaire dans la liste des conjonctions (12 membres)
- DET = déterminants. Les déterminants sont encore à catégoriser dans le corps de la wikigrammaire, qui comprend cependant la liste des quantifieurs (56 membres). Il faut rajouter les deux articles, défini *an, al, ar* et indéfini *un, ul ur*, les déterminants possessifs et le complémenteur *peseurt*. Attention, les pages thématiques de la grammaire ont été catégorisées sous le titre "articles", en opposition aux "fiches" de linguistique formelle.
- NOUN = nom. Ils sont listés dans la wikigrammaire dans la (799 membres)
- VERB = verbe. Ils sont listés dans la wikigrammaire dans la (354 membres), auxquels on peut ajouter la liste des modaux (sauf peut-être *dav, ret* et *arabat* qui ont plutôt une distribution adjectivale), et retrancher les verbes légers *-a, -at* et *-aat* qui ont une distribution suffixale.
- SCONJ = conjonction de subordination. Dans la wikigrammaire, ils sont compris dans les complémenteurs.
- PART = particule. La particule préverbale (*rannig*) est signalée en glose par la lettre R, suivie lorsque le dialecte le permet de la mutation associée à cette particule. Attention, UD classe les particules Q des questions polaires, de 'est-ce que', dans les particules, qui sont dans la wikigrammaire des complémenteurs.

- NUM = numéral (numéraux cardinaux, car les ordinaux sont classés avec les adjectifs).
- INTJ = interjection. Certaines sont signalées directement en gloses, d'autres ont chacun une page dédiée qui est catégorisée comme interjection (liste des interjections).
- PRON = pronom. Les pronoms ne sont pas identifiés individuellement dans les gloses. Seul le type du pronom y est spécifié (pronom fort indépendant, pronom écho, pronom incorporé, etc.).
- PROPN = nom propre. Quelques noms propres sont mentionnés comme tels en glose, mais cette pratique est récente sur le site. Il est plus sûr de passer par les recensements déjà établis par d'autres programmes (Tyers 2008 les avait extraits de Wikipedia), ou de s'appuyer sur la majuscule en graphie pour les récupérer.
- PUNCT = ponctuation. Cette information est présente en graphie en ligne 1, et devrait avoir un parallèle dans la traduction française.
- SYM = symbole. Il s'agit de symboles écrits ne sont pas codés à ce jour dans la wikigrammaire.
- X = autre. Cette notation n'a pas été nécessaire.

En dehors du système d'annotation des données, le site a nécessité pour son développement interne des outils et listes qui pourraient directement alimenter les entraineurs d'algorithme, comme:

- la liste des pages de désambiguïsation qui liste les suffixes pouvant être ambigus
- la liste des finales de mots qui liste les ensembles de suffixes existants, et les décompose
- le liste des redirections de pages, qui gèrent les différences d'orthographe ou de dialecte. L'exploitation de cette dernière liste nécessiterait cependant de nettoyer les redirections concernant les ouvrages de recherche et les abréviations.
- des inventaires trilingues par catégories grammaticales: inventaire des noms, inventaire des adjectifs, inventaire

des adverbes, inventaire des prépositions (très partiel), inventaire des verbes modaux, inventaire des verbes lexicaux, et par sous-catégories, inventaire des verbes inaccusatifs, inventaire des verbes inergatifs.

- trois glossaires (en anglais, breton et français) de plus de 250 termes de grammaire descriptive et formelle, liés chacun à des définitions illustrées par des faits du breton.

4. Conclusion et pistes de discussion

4.1. Prospective et repérage des besoins

Si on synthétise les prospectives dessinées par les différents secteurs, des pôles de demandes émergent assez nettement.

Pour le domaine de l'écrit, Annie Foret (laboratoire LOUSTIC, Rennes I) a mené un repérage des besoins de développement des outils du TAL pour le breton en 2017-2018 (Foret 2018a,b). L'enquête a consisté initialement en huit entretiens libre/semi-orienté d'1h30 d'enseignant.e.s et d'apprenant.e.s, complété par 61 réponses à un questionnaire en ligne comprenant une suggestion ouverte, deux questions sur le profil des répondants (niveau et usage professionnel du breton) et deux autres questions listant des outils développables en demandant lesquels étaient les plus urgents:

- système de lecture augmentée sur écran/tablette avec des livres enrichis de bulles d'information intégrées
- correcteur orthographique / grammatical
- système d'aide à la recherche/exploration d'information
- plateforme de discussion (exemple : échange de recettes ou autre sujet)
- analyseur (aux niveaux morphologique, syntaxique)
- système de détection d'ambiguïtés pour le breton
- dictionnaire des synonymes et expressions/proverbes
- lien entre un dictionnaire et un réseau sémantique

Mekacher (2018) analyse les résultats des questionnaires: il y a unanimité sur le manque de ressources sonores pour l'apprentissage d'une accentuation correcte et une souplesse cross-dialectale. Les locuteurs souhaitent un correcteur orthographique et grammatical intégré aux outils bureautiques, et sont enthousiastes à l'idée d'un système de lecture augmentée. Les résultats doivent être pondérés car il y a peu de répondants, et la liste proposée dans le questionnaire peine à prendre en compte le manque de familiarité des brittophones, enseignant.e.s ou non, avec des outils que justement, ils utilisent peu, d'autant que certains de ces outils sont des outils de développement d'outils numériques. Foret (2016) a exploré un système d'enrichissement de textes qui fournit des synonymes à partir de *Wordnet* et de la base Apertium. Erwan Hupel de l'Université Rennes II a déposé en 2020 une demande de financement pour une thèse sur ce sujet de l'enrichissement de textes par synonymes.

Dans le domaine de la parole orale, la synthèse de la voix de l'IRISA a été livrée à l'OPLB en septembre 2021. Sa diffusion reste un chantier ouvert. Entre autres, un besoin identifié de longue date est celle d'un système GPS capable de prononcer les noms de lieux en Bretagne. En son absence, ce sont les brittophones natifs qui pour utiliser un GPS apprennent dans les faits à interpréter des formes produites par des synthèses de la voix opérant sur d'autres langues. La communication à distance entre jeunes brittophones privilégie les sms, or la dictée des sms en français est possible, efficace et rapide alors qu'envoyer un sms en breton demande de taper le message, voire de stopper l'autocorrection à chaque mot, interprété comme du français. Développer cet outil demanderait de progresser sur la reconnaissance vocale, sachant que c'est un défi conséquent: si la synthèse de la voix a pu se concentrer sur le breton standard, la reconnaissance vocale nécessite de pouvoir traiter une source multidialectale. En ce qui concerne les conditions de réalisabilité de ce gros chantier de la reconnaissance vocale, et étant donné les techniques actuelles, Damien Lolive (c.p. 10.2021) estime que la reconnaissance de la voix nécessiterait un corpus d'un millier de locuteurs différents ne parlant pas plus de trois minutes, si l'audio est transcrit et que le son est propre (pas de

chevauchements, environnement calme). Cela représenterait 50h en tout. Kerbrat (2021a,b) estime, lui, que le corpus devrait atteindre les 200 heures. Kerbrat (2021a,b) mentionne par ailleurs des essais effectués par Francis Tyers avec les données de *Common Voice*. La prosodie de phrase est mal prise en charge dans la synthèse de la voix bretonne actuelle, mais l'un des coordinateurs de la création de la synthèse de la voix travaille de longue date sur la synthèse de la prosodie (cf. Lolive 2008, 2017). Il reste par ailleurs à faire l'étude formelle de la prosodie des phrases en breton, pour systématiser le lien avec la structure syntaxique et avec la structure informationnelle des phrases (signal de focalisation de l'information nouvelle, de signal de l'information donnée, du topique de phrase, etc.).

Les projets qui comportent la création d'une plate-forme pérenne hébergeant les différents corpus sont récurrents, mais peinent à trouver un financement. Le projet *Tal-Breizh (chaînes de traitement et ressources linguistiques pour le breton)* porté en 2015-2017 par Annie Foret (Rennes 1, IRISA) et Ronan Le Coadic (Rennes 2, CRBC) n'a pas été retenu par la Maison de Science de l'Homme de Bretagne. Foret & al. (2015) ont présenté le projet d'une plate-forme ouverte abritant les ressources disponibles pour le breton. Mélanie Jouitteau et Reun Bideault ont présenté en 2018 à la DGLFLF (Délégation Générale à la Langue Française et aux Langues de France) un projet de plate-forme numérique pouvant articuler les données enrichies de la wikigrammaire ARBRES avec des données de dépôt libre, dont chaque collecteur pourrait rester indépendamment propriétaire, afin de pouvoir proposer un hébergement pérenne, dans une banque cross-interrogeable et sous forme réutilisable. L’idée était de fonder une interopérabilité entre ARBRES et les différents sites de collecte individuels et collectifs, et d’offrir un hébergement pérenne pouvant accueillir et inciter de futurs projets émergents. Tyers & Howell (2021) mentionnent aussi en prospective la mise à disposition de la banque d'arbres UD dans une interface de corpus cherchable destinée aux linguistes.

Enfin, en ce qui concerne les sites webs de contenu en langue bretonne de manière générale, l'adaptation ‘responsive web design’ est récemment devenue indispensable à leur lecture sur écran réduit. Les terminaux de consultation d'internet sont de taille de

plus en plus petite, ce qui a obligé les services web à s'adapter rapidement. Le smartphone est maintenant le premier terminal web utilisé, avec une démocratisation rapide. Il touche presque toutes les couches sociales et tous les âges en sont largement équipés. Les applications dédiées pour ces terminaux sont normalisées pour offrir une lecture facile et ciblée. Le passage au responsif reste à faire pour la plupart des contenus web en breton. Ces travaux sont prévus sur la grammaire ARBRES en 2022. Enfin, une traduction automatique d'applications déjà adaptées serait envisageable si un balisage adapté est mis en place.

4.2. Sociolinguistique et observation des usages

La recherche sociolinguistique est attentive aux usages émergents, à la façon dont les locuteurs des langues minorisées s'emparent des outils numériques, et dont cela peut transformer l'acte de parole dans ces langues. Ce champ universitaire, dont nous ne pouvons rendre compte ici, est d'une vitalité revigorante (Baxter 2009, Moal 2017:76 et les références incluses, Blanchard 2014, 2015, Hicks 2017, Davies-Deacon 2020, Dauneau 2019, rapports réguliers approfondis de l'Observatoire de l'OPLB, etc.). Ce champ sociolinguistique est en dialogue avec les rapports commandés par des suprastructures (DGLFLF, appareils d'État, Europe), à qui il fournit des retours d'analyse proches du terrain. Nous recommandons cependant que ces structures relativisent la portée des études sociolinguistiques en termes de prospective car par définition même, les études de sociologie sont intéressées uniquement par l'impact sur la société des outils numériques qui sont déjà finalisés et largement distribuées. Par nature, les études sociolinguistiques sont de prospective limitée puisqu'elles étudient la façon dont les utilisateurs s'accommodent ou se saisissent de l'existant. Elles sont donc justifiées à ignorer parfaitement les savoirs de la recherche universitaire fondamentale, les potentiels et les acteurs de développement, les structures de formation essentielles. La mésinterprétation des états de lieux sociolinguistiques comme synthèse des réalisations du TAL en sa globalité et comme base d'analyse servant à son développement pose un problème réel, non pas pour la sociolinguistique qui ne fait que tenir adéquatement son rôle, mais pour le développement TAL

puisque cela impacte la visibilité de ses réalisations et potentiels, et donc ses ressources pour les appels à projets.

Le champ sociolinguistique universitaire pourrait par ailleurs se saisir des données de type nouveau que les nouveaux outils numériques fournissent. Si les questionnaires en ligne sont apparus dans les pratiques, les études à ce jour ignorent complètement l'existence des outils d'analyse de fréquentation et d'usage des outils en ligne. Jouitteau (2009-) a un outil *google analytics* associé qui lui fournit une vision assez détaillée des usages de son lectorat (volume d'utilisateurs, pages d'entrée, durée de consultation, flux d'utilisateurs de page à page, trouvabilité par les moteurs de recherche, présence de site le mettant en lien et générant du trafic, etc.). L'étude en est assez ludique. On peut deviner quelle année la wikigrammaire est utilisée par les cours de breton à Moscou, ou quand sont les examens universitaires de linguistique au Québec car les fiches de linguistique formelle rédigées en français montrent alors un pic de connections. Sur les quatre dernières années, l'ouvrage a été ouvert par 130 000 utilisateurs qui ont visionné 165 468 pages. Parmi les utilisateurs, 285 sont revenus plus de 5 fois et 579 plus de trois fois. La durée moyenne des sessions dépasse légèrement 2 minutes. Ci-dessous, vous pouvez voir la synthèse *google analytics* sur cinq ans (2017-2021) de la provenance de source de connection des utilisateurs par pays et par régions de l'État français.

Pays	Acquisition			Comportement		
	Utilisateurs	Nouveaux utilisateurs	Sessions	Taux de rebond	Pages/session	Durée moyenne des sessions
	129 924 % du total: 100,00 % (129 924)	133 045 % du total: 100,09 % (132 929)	160 064 % du total: 100,00 % (160 064)	79,03 % Valeur moy. pour la vue: 79,03 % (0,00 %)	1,66 Valeur moy. pour la vue: 1,66 (0,00 %)	00:02:01 Valeur moy. pour la vue: 00:02:01 (0,00 %)
1. France	**68 860** (51,99 %)	69 565 (52,29 %)	87 614 (54,74 %)	76,04 %	1,86	00:02:27
2. Morocco	**9 395** (7,09 %)	9 365 (7,04 %)	10 440 (6,52 %)	85,00 %	1,31	00:01:19
3. Canada	**8 828** (6,67 %)	8 855 (6,66 %)	9 991 (6,24 %)	83,77 %	1,33	00:01:15
4. Algeria	**6 415** (4,84 %)	6 455 (4,85 %)	7 305 (4,56 %)	86,32 %	1,23	00:01:30
5. Belgium	**5 005** (3,78 %)	4 998 (3,76 %)	5 567 (3,48 %)	87,64 %	1,31	00:01:00
6. Switzerland	**2 552** (1,93 %)	2 550 (1,92 %)	2 881 (1,80 %)	84,59 %	1,40	00:01:24
7. Tunisia	**2 312** (1,75 %)	2 321 (1,74 %)	2 568 (1,60 %)	85,36 %	1,25	00:01:03
8. United States	**2 055** (1,55 %)	2 042 (1,53 %)	2 298 (1,44 %)	80,50 %	1,98	00:02:03
9. Cameroon	**1 622** (1,22 %)	1 631 (1,23 %)	1 971 (1,23 %)	77,22 %	1,47	00:02:24
10. Spain	**1 359** (1,03 %)	1 350 (1,01 %)	1 613 (1,01 %)	77,93 %	1,76	00:01:43

Figure 1 : provenance de source de connection des utilisateurs par pays (2017-2021)

Région	Acquisition			Comportement		
	Utilisateurs	Nouveaux utilisateurs	Sessions	Taux de rebond	Pages/session	Durée moyenne des sessions
	68 860 % du total: 53,00 % (129 924)	69 565 % du total: 52,33 % (132 929)	87 614 % du total: 54,74 % (160 064)	76,04 % Valeur moy. pour la vue: 79,03 % (-3,78 %)	1,86 Valeur moy. pour la vue: 1,66 (12,20 %)	00:02:27 Valeur moy. pour la vue: 00:02:01 (21,30 %)
1. Ile-de-France	**23 696** (33,59 %)	23 396 (33,63 %)	28 169 (32,15 %)	79,38 %	1,66	00:02:09
2. Brittany	**13 166** (18,66 %)	13 019 (18,71 %)	20 385 (23,27 %)	63,38 %	2,65	00:04:04
3. Auvergne-Rhone-Alpes	**5 861** (8,31 %)	5 789 (8,32 %)	6 771 (7,73 %)	82,41 %	1,51	00:01:43
4. Occitanie	**4 101** (5,81 %)	4 016 (5,77 %)	4 500 (5,14 %)	83,56 %	1,41	00:01:26
5. Pays de la Loire	**4 002** (5,67 %)	3 931 (5,65 %)	5 062 (5,78 %)	70,80 %	2,14	00:03:08
6. Nouvelle-Aquitaine	**3 871** (5,49 %)	3 790 (5,45 %)	4 446 (5,07 %)	81,71 %	1,74	00:02:01
7. Hauts-de-France	**3 500** (4,96 %)	3 435 (4,94 %)	3 968 (4,53 %)	82,11 %	1,44	00:01:20
8. Grand Est	**3 441** (4,88 %)	3 414 (4,91 %)	3 784 (4,32 %)	84,22 %	1,41	00:01:25
9. Provence-Alpes-Cote d'Azur	**3 340** (4,73 %)	3 287 (4,73 %)	3 700 (4,22 %)	83,30 %	1,40	00:01:22
10. Normandy	**1 936** (2,74 %)	1 911 (2,75 %)	2 145 (2,45 %)	81,68 %	1,44	00:01:19

Figure 2 : provenance de source de connection des utilisateurs par régions (2017-2021)

On voit se dessiner un lectorat dans les zones traditionnelles de pratique de la langue et dans les lieux d'immigration des brittophones, ainsi qu'un lectorat plus largement dans les pays les plus riches de la francophonie. Il existe en effet un lectorat distinct du lectorat brittophone ou d'apprenants, dont l'intérêt premier est la linguistique formelle plutôt que la description du breton. En novembre 2021, les requêtes web qui ont le plus amené sur ARBRES sont les mots clef: structure syntaxique (59), morphème libre et lié (20), construction syntaxique (19), verbe factif (17), verbes factifs 16), complémenteur (14), grammaire bretonne (13), morphème zéro (12), déictique[sic] spatiaux (11) et verbe ditransitif (10). Ce lectorat apprend les notions de base de linguistique formelle en français à travers des exemples du breton.

La synthèse *google analytics* ci-dessous montre la prédominance du moteur de recherche *google* dans les sources de connection sur un flux d'utilisateurs de page à page, ainsi que la présence timide d'un lectorat fidélisé, à connections directes. Le groupe le plus important est ensuite celui des sources de connection non-retrouvables.

Figure 3 : flux des utilisateurs connectés à partir de l'État français, sur 2021

M. Jouitteau fournit annuellement à la structure d'évaluation du CNRS une synthèse détaillée du développement de ARBRES à partir de ces données d'utilisation, enrichies et éclairées par des séances régulières de « surf accompagné » où des utilisateurs de ARBRES montrent leurs usages, mésusages ou incompréhensions. Les outils d'analyse automatisés des usages ont révolutionné le développement des outils en ligne, et il serait étonnant que les autres développeurs, ou même les rédacteurs de blogs culturels en breton, n'aient pas des outils similaires qui renseigneraient la sociolinguistique des usages des brittophones 2.0.

4.3. Pragmatique d'un projet de science ouverte

Un projet comme la wikigrammaire nécessite un investissement sur le long terme rendu possible par l'existence de contrats de

recherche sur des périodes longues comme le statut de fonctionnariat au CNRS. Dans ce cadre qui le rend possible, les conséquences pragmatiques d'un tel choix de recherche sont aussi évaluables.

ARBRES est encore souvent amalgamé avec *wikipédia* et appréhendé par ses utilisateurs comme une ressource sans auteur réel, et les bibliographies scientifiques sont encore centrées sur le circuit papier, même alors qu'elles sont directement copiées à partir de formats numériques. Les travaux de science ouverte comme les travaux numériques en général sont encore généralement peu cités, et dans des formats des plus créatifs, irrécupérables automatiquement. Ces facteurs rendraient dangereuse une évaluation exclusive de la recherche par les mentions correctes en bibliographies universitaires.

L'attitude des entités évaluatrices s'est cependant nettement améliorée depuis les débuts de la wikigrammaire en 2009. La Déclaration de San Francisco sur l'évaluation de la recherche (Dora) en 2012 et le Manifeste de Leiden en 2015 ont livré une analyse critique des pratiques d'évaluation de la science restreintes à la citométries de revues et d'articles. Ils ont avancé des recommandations en matière d'utilisation d'indicateurs scientométriques pour l'évaluation des scientifiques (Pourret 2021). Le CNRS s'est engagé fermement ces dernières années en soutien de la science ouverte, et les critères bougent. Il y a une dizaine d'années, dans les évaluations annuelles du CNRS, ARBRES a reçu occasionnellement, au milieu de soutiens déclarés de collègues qui en cernaient le potentiel, des évaluations bonhommes se souciant de ce que l'autrice ne « perde pas trop de temps » sur son site, appréhendé comme un blog de loisirs. Certains avis évaluaient son volume de publications en laissant ouvertement de côté le développement de la base de données. Ce type de retour a disparu et il est chaque année plus facile d'insérer une wikigrammaire dans les critères d'évaluation proposés.

De façon assez ironique mais sans doute transitoire, ce sont les initiatives de diffusion numérique des produits scientifiques qui s'appuient le plus sur la science ouverte qui manquent à valoriser les réalisations numériques qui répondent le mieux à ses critères.

En France, les plates-formes de visibilisation comme HAL valorisent les articles papier rendus disponibles à l'intérieur de la plate-forme, et non pas les banques de données, articles en ligne ou sites de recherche disponibles en ligne sans le recours de HAL. L'interface *google scholar* les rend entièrement invisibles, comme c'est d'ailleurs le cas pour tout système numérique qui requiert un ISSN ou ISBN pour reconnaître un travail.

L'aide au transfert des savoirs vers la société est officiellement souhaitée, mais les conditions de réalisabilité peinent encore à être assurées largement. De nombreux laboratoires de linguistique n'ont pas d'ingénieurs de recherche en informatique, et le développement numérique dépend *in fine* des capacités informatiques en propre de chercheurs en sciences humaines. En pratique, un chercheur doit trouver régulièrement et au coup par coup des ressources pour la gestion des mises à niveau du logiciel. L'accessibilité d'un ouvrage numérique dépend principalement des algorithmes de *google*, avec les risques que cela comporte. Dans le tableau ci-dessous construit par *google analytics*, on voit nettement les effets du changement d'algorithme du moteur de recherche de *google* en janvier 2020, qui a déclassé volontairement les productions numériques non-adaptées à la consultation sur smartphones et tablettes. L'adaptation à la visualisation sur smartphones et tablettes est appelée « passage au responsif ». Il n'est pas effectué sur ARBRES.

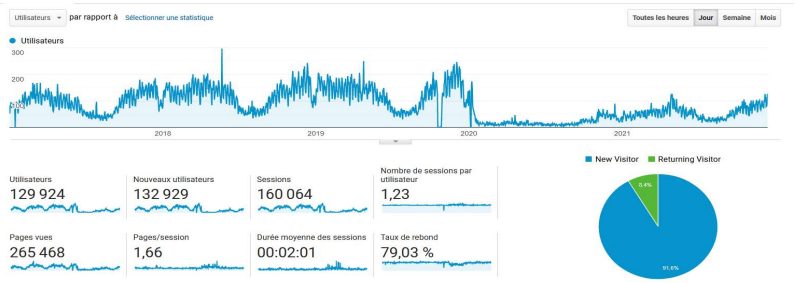

Figure 4 : synthèse google analytics sur 5 ans du nombre d'utilisateurs de la wikigrammaire ARBRES

La moyenne de 130 utilisateurs humains par jour sur 2019 est descendue brutalement en 2020 à 25 par jour, c'est-à-dire une division par plus de cinq. Le site a été globalement laissé aux utilisateurs déjà fidélisés : entre 2019 et 2020, la moyenne de temps d'une session est passée de 1:50 à 3:15 minutes. Sur la courbe des fréquentations, on repère la chute nette de janvier 2020, et une période de mise hors ligne en 2019, qui est significative. Le laboratoire IKER (CNRS) a financé sur fonds propres la mise à niveau du logiciel wiki et le passage au responsif, confié à une entreprise privée en Bretagne. L'administration a déclenché le paiement avant clôture des travaux, et le passage au responsif n'a finalement jamais été effectué par l'entreprise, qui a même laissé un temps le site hors ligne (la chute exceptionnelle à zéro utilisateurs fin 2019). En bibliothèque papier, dans de telles situations, l'ouvrage serait considéré TDE (Tombé Derrière les Étagères), et des professionnels bibliothécaires et documentalistes s'attaqueraient au problème. Pour le circuit numérique, nous ne repérons pas l'équivalent de ces professionnels chargés de s'assurer de l'accessibilité des ouvrages pour les publics concernés. Ce problème est considérable et touche tous les grands sites numériques scientifiques financés par projet qui sont déposés à clôture dans les grandes infrastructures de type HumaNum. A notre connaissance, après financement de création, les usages et accessibilités web ne sont plus évalués et un changement d'algorithme google peut drastiquement diviser leur lectorat sans que ce soit même repéré. A noter que dans le cas de ARBRES et des autres projets de science ouverte en développement constant, cette question de l'accessibilité ne pourrait de toute façon pas être déléguée car leur dépôt dans les grandes infrastructures n'est pour l'instant pas une option.

Nous pensons avoir concouru ici à documenter et appuyer la conclusion que pour évaluer chercheuses et chercheurs, les critères quantitatifs de publications et de citation, et les critères qualitatifs de gradation des revues et des langues d'expression de la recherche (anglais vs. langues des communautés de locuteurs) devraient être

enrichis par une appréhension de la réalité des transferts de savoirs faits vers la société, en facilitant la citabilité des ouvrages numériques libres et en visibilisant les pratiques de science ouverte. Soutenir pragmatiquement le transfert des savoirs vers la société permettrait en sus à plus de chercheurs de se lancer avec confiance dans des projets d'envergure en science ouverte, à même d'alimenter les traitements automatiques des langues.

Boîte noire

Dans le cadre des recherches pour cet article, Mélanie Jouitteau a contacté et interviewé plusieurs collègues. Annie Foret a été interviewée via le logiciel libre *Jitsi* en décembre 2021. Le résumé de la rencontre lui a été communiqué en ligne. Damien Lolive a été contacté le 15 octobre, puis recontacté le 4 novembre, avec deux collègues hommes en copie du message, dont Reun Bideault, développeur web qui opère régulièrement les mises à jour de la wikigrammaire ARBRES, et qui était alors consultant pour l'article. Ce dernier a accepté d'organiser l'interview de Monsieur Lolive le 24 novembre dans des locaux prêtés par le centre de formation *Kelenn*, et de rejoindre l'article en signature. Le résumé du développement de la synthèse de la voix a été communiqué ensuite à Damien Lolive, et corrigée par ce dernier et Gaëlle Vidal après vérification de quelques points auprès des locuteurs dont la voix a servi à la constitution du corpus oral. Dewi Kerbrat, auteur du rapport 2021 de l'OPLB, a été contacté le 14 octobre, puis le 4 novembre avec deux collègues hommes et les deux co-directeurs de l'OPLB en copie, puis le 6 novembre via Facebook. Il a confirmé par email le 8 novembre que son rapport sur les outils numériques en breton n'incluait pas la wikigrammaire ARBRES, et laissé ouverte la possibilité qu'elle le soit. Recontacté le 9 novembre avec une explication du potentiel pour le TAL de la base de données de ARBRES, et averti que le rapport négligeait des ressources universitaires, il n'a pas donné suite.

Bibliographie

AUBRY, Yves (2000). *Synthèse vocale en breton*, ms. De mémoire de maîtrise, IUP MIME Le Mans, TES/ENSSAT.

AUBRY, Yves (2004). *Logiciel du traitement de la parole et d'aide à l'enseignement et à l'apprentissage de la prosodie: application au breton*, travaux de D.R.T. d'ingénierie, Université du Maine.

BAXTER, R.N. (2009). 'New technologies and terminological pressure in lesser-used languages. The Breton Wikipedia, from terminology consumer to potential terminology provider', *Language Problems*

and Language Planning 33:1, John Benjamins: Amsterdam/Philadelphia, 60-80.

BLANCHARD, Jean-François (2014). « Pratiques langagières et processus dialogique d'identification pour une langue minorée. Le web en langue bretonne », Gaël Hily (dir.), *Expression de l'identité dans le monde celtique*, Rennes : TIR. 9-34.

BLANCHARD, Jean-François (2015). *Pratiques langagières et processus dialogiques d'identication sur les réseaux socionumériques. Le cas de la langue bretonne*, ms. thèse. Université Rennes 2. texte.

CHEVEAU, Loïc & Pierre-Yves KERSULEC (2012-évolutif). *Dictionnaires bretons parlants*.

CORNILLET, Gérard (2017). *Geriadur Brezhoneg-Galleg*, (version corrigée en 2020, texte).

AN DROUIZIG (2021). *Site de ressources numériques et association de création de ressources*, [accédé le 20.11.2021].

DAUNEAU, Goulven (2019). *Brezhoneg, Niverel, Deskadurezh : hiziv ha warc'hoazh*, ms. de mémoire de master, U. Rennes II. texte.

DAVIES-DEACON, Merryn (2020). *New speaker language and identity: Practices and perceptions around Breton as a regional language of France*, ms. de thèse.

DELOOF, Jan (2008-2010). *Bretons-Nederlands Woordenboek*, interface web par Kevin Donnelly.

DESSEIGNE, Adrien, Loïc CHEVEAU & Pierre-Yves KERSULEC (2013-2018). *Banque Sonore des Dialectes Bretons, projet de documentation multimédia en ligne*, site.

DONNELLY, Kevin (2010). 'Jan Deloof Breton-Dutch Dictionary', blog *Me, Myself, Why? Free software and languages, not necessarily in that order*, texte, [consulté le 13.12.2021].

LE DÛ, Jean (2001). *Nouvel Atlas Linguistique de Basse-Bretagne*, vol. I et II, Centre de Recherche Bretonne et Celtique, Université de Bretagne Occidentale, Brest.

FAVEREAU, Francis (1993). *Dictionnaire du breton contemporain / Geriadur ar brezhoneg a-vremañ*. Morlaix: Skol Vreizh, moulte rééditions, version en ligne.

FAVEREAU, Francis (2015). *Geriadurig ar brezhoneg a-vremañ / Dictionnaire compact du breton contemporain*, Morlaix: Skol Vreizh.

FAVEREAU, Francis. 2016-évolutif. *Grand dictionnaire bilingue breton-français, français-breton*, texte.

FAVEREAU, IRISA & TES (1999). *Ar geriadur a gomz brezhoneg a-vremañ*, Morlaix : Skol Vreizh. CD-ROM.

FORET, Annie, Valérie BELLYNCK & Christian BOITET (2015). 'Akenou-Breizh, un projet de plate-forme valorisant des ressources et outils informatiques et linguistiques pour le breton', présentation à la conférence *TALARE (Traitement Automatique des Langues Régionales de France et d'Europe)*, texte.

FORET, Annie (2016). « Enrichissement de données en breton avec Wordnet », Poibeau, Thierry, Teresa Lynn, Delyth Prys & John Judge (éds.), *Proceedings of the Second Celtic Language Technology Workshop* (CLTW 2016), 55-61. texte.

FORET, Annie (2018). « Breton-français et numérique, projet LangNum-br-fr (phase conception) ». *Conférence Langues et numérique 2018*, Juillet 2018, Paris, France. texte ou texte.

FORET, Annie (2018b). 'Logiciels et ressources pour le breton', document du projet LangNum-br-fr, ms. 12p.

GUILLOU, A. (2000). *Correcteur de prosodie pour la langue bretonne*, ms. de rapport de projet.

HERRIEU, Loeiz (1994). *Kammdro an ankoù*, Gourhelen, Ronan Huon embanner: Al Liamm.

HICKS, Davyth (2017). « Breton – a digital language ? », *The Digital Language Diversity Project*, Erasmus +. texte.

IRISA (2001). *Rapport d'activité 2001. Projet CORDIAL. Communication multimodale personne-machine à composantes orales : méthodes et modèles*, texte.

AN INTANV, Pascal (1994). *War hent fonetikadur ar Brezhoneg / Sur les chemins de la phonétisation du breton*, ms. de mémoire de maîtrise, Université de Rennes II.

JOUITTEAU, Mélanie (2013b). « La linguistique comme science ouverte; Une expérience de recherche citoyenne à carnets ouverts sur la grammaire du breton », *Lapurdum* XVI, Charles Videgain (dir.), 93-115, texte.

JOUITTEAU, Mélanie (éd.) (2009-2021). « Traitement automatique du langage - Breton », *ARBRES, wikigrammaire des dialectes du breton et centre de ressources pour son étude linguistique formelle*, IKER, CNRS, URL. - génèse du présent article et mises à jour.

KREIZENN AR GERIAOUIÑ (2016). *Geriaoueg yezhadur*, Brezhoneg 21 (éd.), texte.

KERBRAT, Dewi (2021a). *Ar brezhoneg en oadvezh an niverel, diagnostik ha strategiezh diorren*, ms. de rapport pour l'OPLB.

KERBRAT, Dewi (2021b). *La langue bretonne à l'ère du numérique, diagnostic et stratégie de développement*, ms. de rapport pour l'OPLB.

LEIXA, Jérémy, Valérie MAPELLI & Khalid CHOUKRI (2014). *Inventaire des ressources linguistiques de langues de France*, Organisme ELDA, ms. de rapport pour la DGLFLF.

LOLIVE, Damien (2008). *Transformation de l'intonation : application à la synthèse de la parole et à la transformation de voix. Intelligence artificielle [cs.AI]*, ms. de thèse de l'Université Rennes I. texte.

LOLIVE, Damien (2017). *Vers plus de contrôle pour la synthèse de parole expressive. Intelligence artificielle [cs.AI]*, ms. de HDR, Université de Rennes I.

MARTINET, Pierre (2021). *Contributions à l'enrichissement automatisé de langues peu dotées. Cas du breton et des grammaires formelles*, ms. de rapport de stage (6 mois), laboratoire SemLIS (IRISA), Rennes I. texte.

MENARD, Martial et Hervé LE BIHAN (2016-évolutif). *Devri: Le dictionnaire diachronique du breton*, Université Rennes II & Kuzul ar Brezhoneg, en ligne.

MENARD, Martial et Iwan KADORED (dir.) (2001). *Geriadur Brezhoneg*, Embannadurioù An Here.

MEKACHER, Echraf (2018). *Projet LangNum-br-fr*, ms. du laboratoire LOUSTIC, U. Rennes I. texte.

MINOCHA, Akshay et Francis TYERS (2014). « Subsegmental language detection in Celtic language text », *Proceedings of the First Celtic Language Technology Workshop* CLTW1, 76-80, texte.

MOAL, Stefan (2017). *Médiation, transmission, création. La revernacularisation du breton au 21e siècle*, ms. de HDR.

MOCQUARD, Guillaume (1999). *Correcteur de prosodie*, ms. de rapport de stage IFSIC, TES/IRISA, ENSSAT.

MOCQUARD, Guillaume (2001). *Korpus prosodiezh*, ms. de mémoire de maîtrise, Université de Rennes II.

Ar MOGN, Olier (2015). « Langue bretonne et nouvelles technologies : une vitalité à soutenir », présentation à *Technologies pour les Langues Régionales de France*, Meudon. vidéo.

MORVAN, Pierre (2019). *Ha difaziañ a ra LanguageTool ar c'hemmadurioù? Peseurt hentenn sevel evit gellet gouzout peseurt barregezh a zo gant an difazier LanguageTool war ar c'hemmadurioù?*, ms. de mémoire de maîtrise, Université Rennes II.

OPLB (2021a). *Termofis*, dictionnaire terminologique, en ligne.

OPLB. (2021b). *Kerofis*, base de données toponymique, en ligne.

OPLB. (2021c). *Corpus de phrases en breton, ou en français*, accessible.

Petit, M (2003). *Correcteur orthographique de langue bretonne*, ms. rapport de projet, ENSSAT, 1-37.

POIBEAU, Thierry (2014). 'Processing Mutations in Breton with Finite-State Transducers', *Proceedings of the First Celtic Language Technology Workshop*, Dublin, Ireland. texte.

TYERS, Francis Morton (2008). 'Extracting bilingual word pairs from wikipedia', *Proceedings of the SALTMIL Workshop at the Language Resources and Evaluation Conference*, LREC2008, 19–22.

TYERS, Francis Morton (2009). 'Rule-based augmentation of training data for breton–french statistical machine translation', *Proceedings of the 13th Conference of the European Association for Machine Translation*, 213–218. texte.

TYERS, Francis Morton (2007-2009). *Breton morphological analysis*, http://xixona.dlsi.ua.es/~fran/breton/index.php, GNU-GPL.

TYERS, Francis Morton (2010a). 'Rule-based Breton to French machine translation', *Proceedings of the 14th Annual Conference of the European Association of Machine Translation*, 174-181. texte et poster.

TYERS, Francis Morton (2010b). « An treiñ emgefreek diazezet war reolennoù evit treiñ ar brezhoneg e galleg », *Hor Yezh* 262, 27–39. [traduction par Thierry Fohanno]

TYERS, Francis Morton (2015). *Rule-based augmentation of training data in breton–french statistical machine translation*, rapport.

TYERS, Francis Morton et Vinit RAVISHANKAR (2018). « A prototype dependency treebank for Breton », *Actes de la conférence Traitement Automatique de la Langue Naturelle*, TALN 2018, 197-204. texte.

TYERS, Francis Morton et Nicholas HOWELL (2021). « Morphological analysis and disambiguation for Breton », *Language Resources and Evaluation*, 431-473. preview.

POURRET, Olivier (2021). « Comment la science ouverte peut faire évoluer les méthodes d'évaluation de la recherche », *The conversation*, [4 novembre 2021, accédé le 06 novembre].

YEKEL, Tangi, Riwal GEORGELIN et Juluan AR C'HOZH (2015-2021). *Brezhoneg Bro-Vear*, Blog de l'association *Hent don*.

Mélanie Jouitteau
IKER, CNRS, UMR 5478,
Université de Pau et des Pays de l'Adour
Université Bordeaux Montaigne
melanie.jouitteau@iker.cnrs.fr
Reun Bideault
développeur Web Indépendant

OUTILS NUMÉRIQUES ET TRAITEMENT AUTOMATIQUE DU BRETON

lepoleethik2@gmail.com