OFFICE PUBLIC DE LA LANGUE BRETONNE

LA LANGUE BRETONNE A L'ERE DU NUMERIQUE

Diagnostic et stratégie de développement

Novembre 2020

Langue et innovation numérique

Table des matières

❖ Introduction

Depuis l'apparition de l'outil informatique dans nos vies, la langue bretonne a su s'adapter. D'un point de vue terminologique tout d'abord, dès les années 1980, pour pouvoir parler, jouer, échanger, travailler en langue bretonne sans difficulté. De très nombreux néologismes ont ainsi vu le jour au fil des années et sont rentrés aujourd'hui dans le language commun : *urzhiataer, meziant, pladenn galet, memor, logodenn, klikañ, postel, enrollañ, restr, Kenrouedad*...

Dans un deuxième temps c'est la traduction qui a occupé la communauté brittophone afin de proposer des outils aux utilisateurs. De nombreux logiciels libres ont ainsi été traduits en breton et sont mis à jour régulièrement depuis une vingtaine d'années.

La présence de la langue sur Internet afin de proposer de l'information en ligne au sens large, a également fait l'objet d'un gros travail, comme le montre le projet d'encyclopédie libre Wikipedia en langue bretonne lancé en 2004.

Enfin, des outils internes spécifiques ont été développé : un traducteur automatique breton-français, des dictionnaires et bases de données linguistiques, des correcteurs orthographiques et grammaticaux.

La plupart de ces avancées ont été le fait d'initiatives individuelles et bénévoles qu'il convient ici de saluer particulièrement. Cela montre que la communauté est vivante et s'adapte en continu aux challenges que représentent les évolutions technologiques incessantes et de plus en plus rapides dans nos vies.

Cependant, force est de noter que ces évolutions positives se sont faites sans coordination et ne peuvent toucher nécessairement qu'un secteur marginal du numérique (le domaine du libre). L'investissement personnel et bénévole, bien que nécessaire, ne sera jamais suffisant à lui seul pour combler les manques importants qui subsistent et lever les obstacles à l'utilisation naturelle de la langue.

C'est pourquoi il a semblé indispensable à l'OPLB de mener une réflexion approfondie et globale sur ce domaine afin de pouvoir proposer à l'ensembles des acteurs, qu'ils soient institutionnels, professionnels ou tout simplement geek, une stratégie adaptée et progressive.

Ce rapport a donc souhaité dans une première partie dresser l'état des lieux de la langue bretonne dans le monde numérique et des nouvelles technologies. Il s'agit de prendre en compte ce qui a été réalisé et de pointer les manques les

plus criants. Cet état des lieux systématique nous permet d'esquisser les contours de ce qui pourrait être une stratégie pluriannuelle pour développer la présence de la langue, son utilisation et son adaptabilité aux outils numériques sur les 10 prochaines années.

L'objectif que nous nous fixons est de faire de la langue bretonne une langue moderne, parfaitement intégrée dans l'environnement numérique et adaptée aux usages de ses locuteurs. Cet objectif est également une réponse aux nombreuses sollicitations de la part des britophones de pouvoir utiliser leurs dispositifs numériques en breton dans leur vie quotidienne.

Nous souhaitons également, grâce à ce rapport, que les technologies de la langue, aujourd'hui trop peu développées à l'égard du breton, prennent davantage en compte les particularités de cette langue et que soient lancés des projets de recherche dans cette direction en développant des partenariats avec les universités et les différents instituts de recherche.

❖ Partie 1. La langue bretonne dans notre environnement numérique, état des lieux

L'un des enjeux importants et probablement le plus prégnant pour la langue bretonne est d'augmenter sa visibilité dans notre environnement numérique. De notre smartphone à notre ordinateur, nous avons répertorié les différentes façons de voir et d'utiliser le breton dans notre vie de tous les jours afin, dans un premier temps, de recenser l'existant pour ensuite pouvoir s'attacher à combler les manques.

1 Utiliser ses appareils numériques en langue bretonne

1.1 Sur son ordinateur

Il existe de nombreux logiciels libres traduits en langue bretonne, la plupart du temps par des associations comme An Drouizig, par l'Office public de la langue bretonne ou encore par des bénévoles qui travaillent en dehors de toute structure sans malheureusement toujours bien maîtriser la langue, les concepts qu'ils traduisent ou les terminologies adéquates.

L'offre en logiciel libre de droits traduits en breton couvre la plupart des usages actuels d'un ordinateur :

- Une suite bureautique (**LibreOffice**) ainsi que des correcteurs d'orthographe et de grammaire (disponibles pour Microsoft Office, Chrome, Firefox, Thunderbird…) :
  - Le correcteur **Hunspell**, sous licence libre, corrige l'orthographe (https://github.com/Drouizig/hunspell-br).
  - Le correcteur **An Drouizig,** disponible sur Microsoft Office, corrige l'orthographe et la grammaire (http://www.drouizig.org/index.php/br/binviou-br/an-drouizig-difazier/172-difazier-an-drouizig-pajenn-nevez).
  - **LanguageTools** corrige l'orthographe et quelques éléments de grammaire (https://languagetool.org/br/).
- Des logiciels pour la navigation sur le web (**Firefox**) et l'échange de courriels (**Thunderbird**)
- Des logiciels multimédia (**VLC** pour la vidéo, **Clementine** pour la musique)

- Des logiciels d'édition graphique (**Inkscape**, **Gimp**, **Tuxpaint**)

En revanche, si des traductions de logiciels propriétaires ont existé par le passé (dans les années 1990/2000), comme Opera, il n'en existe aujourd'hui plus aucune.

Du côté des systèmes d'exploitation, il existe des versions bretonnes pour certaines interfaces Linux comme **Gnome**, parfois incomplètes. Nous pouvons également citer une distribution Linux bilingue (français-breton), à destination des écoles bilingues : https://theosept.fr/spip.php?article272.

L'OPLB a également mené à bien une première collaboration avec Microsoft permettant d'intégrer la langue bretonne dans la liste des langues prise en compte par cette entreprise. La locale Windows qui permet de présenter la date et le calendrier en langue bretonne a également été traduite et est aujourd'hui fournie par défaut sur tous les appareils.

Cependant, de manière générale, les systèmes et logiciels grand public les plus populaires et les plus utilisés, souvent privés, ne sont pas disponibles en breton voire incompatibles avec l'utilisation du breton.

La langue bretonne est davantage présente dans le monde du Libre, qui dispose d'une moindre popularité auprès du public.

1.2 Sur son smartphone

La situation est encore plus dégradée en ce qui concerne les smartphones. L'offre d'applications permettant une utilisation de son téléphone en breton est extrêmement réduite. Parmi les applications indispensables (client mail, navigateur internet, messagerie, appareil photo, GPS…) il n'existe de version bretonne que pour :

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

On peut néanmoins noter la disponibilité de l'autocorrection et la prédiction de mot en breton sur le clavier virtuel **Microsoft SwiftKey**.

Du côté des systèmes d'exploitation pour smartphone, les principaux produits du marché ne sont pas disponibles en breton (iOS pour Apple et Android pour Google) même s'il existe une version alternative libre d'Android, Lineage OS, ouverte à la traduction en breton (https://crowdin.com/project/lineageos). Des

projets libres comme Firefox OS et Ubuntu Phone ont été traduit en breton, mais comme ils sont maintenant abandonnés, le travail de bénévoles a été perdu.

1.3 Les jeux vidéo en breton

En ce qui concerne les jeux vidéo, la situation est la même sur ordinateur ou sur smartphone. Quelques titres ont été traduits, certains propriétaires et d'autres sous licence libre.

On peut mentionner le jeux populaire **Minecraft** par exemple (https://www.minecraft.net/fr-fr), ou d'autres jeux comme **FreeCol**, sous licence libre.

Du côté des smartphones on peut trouver une poignée de jeux proposant une version bretonne, comme **Steredenn** (http://steredenn.pixelnest.io/) ou **Flipon**.

1.4 Les paramètres régionaux en langue bretonne

La langue bretonne dispose d'un **code ISO**, qui est une norme internationale de codification des langues, permettant leur utilisation dans un environnement informatique.

Les paramètres régionaux, ou locales en anglais, sont un ensemble d'informations spécifiques à une langue ou à une région, utilisées par les systèmes d'exploitation (d'ordinateur ou de smartphone) pour afficher diverses données, par exemple :

- Les noms des langues, des territoires, des pays ;
- Les jours de la semaine, les mois, le format de l'heure ;
- Les caractères utilisés pour écrire la langue, et l'ordre dans lequel classer les mots ;

L'OPLB a d'abord mené une collaboration avec Microsoft pour rendre disponible les paramètres linguistiques du breton sur Windows. A l'heure actuelle, l'OPLB alimente le **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, le breton a atteint le niveau de couverture *Moderate*++, le dernier pallier avant d'atteindre le dernier niveau de couverture *Modern*.

Cependant, ces paramètres régionaux n'ont de réelle utilité que s'ils sont implémentés par les différents éditeurs de logiciels. Par exemple, certains moteurs de recherche ou réseaux sociaux ne considèrent toujours pas « c'h » comme une lettre, et découpent les mots au niveau de l'apostrophe.

2 Surfer sur Internet

2.1 Les moteurs de recherche

La première chose dont on a besoin pour utiliser Internet, c'est d'un moteur de recherche qui prenne en charge la langue dans laquelle on navigue. Pour le breton, cette première étape présente déjà des obstacles, même si l'on ne part pas de zéro :

- **Google** a bénéficié et bénéficie encore partiellement aujourd'hui de la traduction de son interface en breton. Son algoritme ne semble cependant pas optimisé pour cette langue, et il est difficile de savoir si l'alphabet breton est bien pris en charge (notamment la lettre *c'h*).
- **Qwant** quand a lui est entièrement traduit en breton pour ce qui est de son interface, mais ne prend pas en charge le breton comme langue de recherche.
- **DuckDuckGo** présente également une interface presque entièrement traduite, mais son algoritme de recherche ne prend pas en compte non plus le breton.

2.2 Les réseaux sociaux

Même si les principaux réseaux sociaux ne disposent pas d'interface officielle en langue bretonne, ils peuvent compter sur un nombre d'utilisateurs croissant qui s'y expriment en breton.

- Sur **Twitter**, on comptabilise 713 utilisateurs qui tweetent en breton[1] (au moins un tweet) (17 000 pour le basque, 15 000 pour le gallois ou l'irlandais, des langues où l'enseignement bilingue est généralisé depuis longtemps, 2 000 pour le gaêlique écossais, 100 pour l'occitan). Au total 180 824 tweets avaient été écrits en breton le 27 octobre 2020. L'interface de Twitter n'est pourtant toujours pas disponible en langue bretonne.
- **Facebook** est disponible en breton. Il s'agit d'une traduction participative de qualité variable. Le groupe « Facebook e brezhoneg », lui, compte plus de 10 000 membres.

Mastodon, réseau semblable à Twitter mais libre de droits, fait l'objet d'un projet de traduction participative en breton (https://crowdin.com/project/mastodon).

[1] Source : http://indigenoustweets.com/br/, consulté le 27 octobre 2020

2.3 Les contenus en langue bretonne

Les contenus en langue bretonne sur internet se trouvent principalement sur des sites peu fréquentés, qui sont spécialisés dans la langue et la culture bretonne, ou sur des sites institutionnels.

- Il existe une version en langue bretonne de **Wikipédia** depuis 2004. Elle contient près de 70 000 articles[2] (86 000 en occitan, 132 000 en gallois ou 367 000 en Basque contre 2 millions en français, et 6 millions en anglais). Sur les 57 000 utilisateurs inscrits seulement 85 sont régulièrement actifs dans l'écriture et la correction d'articles. Wikipédia est le plus grand site de contenus en langue bretonne. Malheureusement la qualité et l'exhaustivité des articles demanderaient à être améliorées.
- De nombreux sites ou blogs existent en breton, pour l'essentiel dédiés à la langue ou la culture bretonne.
- On trouve également des sites de médias en ligne tels **Brezhoweb** pour la TV, **Radio Breizh** pour la radio.
- Enfin, il existe quelques sites institutionnels traduits en breton. De manière générale ils sont peu nombreux et la langue n'y est pas toujours bien mise en valeur : les sites ne sont pas entièrement traduits ou les mises à jour ne sont pas effectuées.

3 Apprendre la langue à l'aide d'outils numériques

Les outils numériques pour apprendre la langue bretonne sont rares. La langue n'est présente sur aucun grand nom du domaine (Duolinguo, Memrise, Babbel…). Il existe quelques applications spécialisées.

- L'application **Kwizh** (iOS et Android) pour apprendre des mots de vocabulaire en breton.
- L'application de la méthode **Assimil** (iOS et Android), payante, qui reprend en fait la version papier.
- L'application et le site **Edu-Breizh** qui propose des cours en ligne payants mais qui n'a pas rencontré son public.
- **Kervarker** (http://www.kervarker.org/fr/lessons_01_toc.html) où l'on trouve quelques leçons et plusieurs autres sites internet, la plupart du temps anciens et non entretenus.

[2] https://meta.wikimedia.org/wiki/List_of_Wikipedias consulté le 26/11/2020

La langue bretonne a pris beaucoup de retard sur ce dossier stratégique. La mise en place d'une plateforme d'auto-apprentissage en ligne attractive, de matériaux pédagogiques ludiques fait cruellement défaut.

❖ **Partie 2.** Inventaires des ressources linguistiques et des outils numériques en langue bretonne

Moins visibles, les ressources et outils linguistiques n'en sont pas moins essentiels. Ils sont destinés avant tout aux linguistes et aux informaticiens. Ils permettent le développement de nouvelles applications, de nouveaux services et de nouvelles technologies en langue bretonne. Pour établir cet inventaire, nous avons choisi une classification inspirée de celle utilisée par les Livres blancs de META-NET[3]. Tout d'abord les ressources linguistiques, qui comprennent les matériaux nécessaires à l'élaboration d'outils de traitement de la langue. Nous avons répertorié uniquement les ressources disponibles sur Internet. Ensuite nous avons listé les outils et technologies de traitement de la langue développés à ce jour pour le breton.

Des tableaux présentent la situation actuelle des ressources et des outils disponibles pour la langue bretonne. Ils sont également inspirés des tableaux utilisés dans les rapports META-NET. Pour chaque type de ressource ou d'outils, 6 indicateurs sont présentés, dont la couleur varie suivant que l'outil est satisfaisant (vert) ou insatisfaisant (rouge). Les 6 indicateurs sont :

- **Quantité** : Les ressources ou les outils existent-ils pour le breton ? En existe-t-il plusieurs ?

  Il n'existe pas de ressource ou d'outil de ce type pour le breton.

  Ces ressources ou outils existent en quantité insuffisante.

  La quantité de ressources ou d'outils de ce type est satisfaisante.

- **Disponibilité** : Les ressources ou les outils sont-ils disponibles en ligne ? Sont-ils libres de droits ? Gratuits ? Compatibles avec plusieurs systèmes d'exploitation ?

[3] META-NET est une organisation européenne qui rassemble des acteurs des technologies de la langue. Elle a écrit plusieurs "Livres blancs" rendant compte de l'état des langues officielles de l'Union Européenne et de plusieurs langues non-officielles de l'UE dans le domaine des technologies de la langue.

Les ressources ou les outils ne sont pas disponibles en ligne, ne sont pas consultables ou ne sont pas téléchargeables.

Les ressources ou outils sont disponibles mais ne sont pas sous licence libre, ne fonctionnent pas sur tous les systèmes d'exploitation ou sont payants.

Les ressources et outils de ce type sont disponible en ligne gratuitement, téléchargeables et libre de droits.

- **Qualité** : Les ressources ou outils sont-ils de bonne qualité ? Ont-ils été évalués ? Sont-ils performants ?

Les ressources et outils sont de mauvaise qualité, leurs performances n'ont pas été évaluées.

Les ressources et outils peuvent être de bonne qualité mais leurs performances n'ont pas été évaluées.

Les ressources et outils sont de bonne qualité, leurs performances ont fait l'objet d'évaluations.

- **Couverture** : Les ressources ou outils assurent-ils une bonne couverture des usages pour lesquels ils ont été conçu (genre, domaine, langues prises en charge...) ?

Les ressources ou outils ont une couverture très faible ou couvrent des cas très spécifiques.

Ces ressources ou outils offrent une bonne couverture mais restent spécialisés à certains domaines.

Les ressources ou outils offrent une très bonne couverture et couvrent des domaines variés.

- **Durabilité** : Les ressources ou les outils sont-ils toujours maintenus ? Sont-ils documentés ? Correspondent-ils aux normes ?

Utilisation de formats propriétaires, peu ou pas de mises à jour, pas de documentation.

Les outils et ressources n'utilisent pas les standards à jour, peu de mise à jour, peu de documentation.

Les ressources et les outils utilisent les dernières normes et sont très bien documentés.

- **Adaptabilité** : Les ressources ou les outils peuvent-ils être adaptés à de nouvelles tâches, de nouveaux domaines ? Avec quelle facilité ?

  Il est impossible d'utiliser les ressources ou les outils pour effectuer d'autres tâches que celles pour lesquelles ils ont été conçus, ou de les transposer à d'autres domaines.

  Les ressources ou les outils sont pertiellement adaptables à d'autres tâches ou domaines, ou difficiles à adapter.

  Les ressources ou les outils peuvent être facilement et entièrement adaptés à de nouvelles tâches ou de nouveaux domaines.

1 Les ressources linguistiques

Dans le but de concevoir des outils linguistiques il est nécessaire en premier lieu de collecter un ensemble de données de différentes natures. Ces données seront ensuite utilisées pour créer, tester ou alimenter différents algorithmes, logiciels et systèmes. Nous pouvons diviser ces ressources en plusieurs catégories : les ressources lexicales qui concernent les mots et leurs caractéristiques, les ressources grammaticales qui offrent des représentations informatiques du fonctionnement de la langue, et les corpus qui donnent des contextes d'utilisation de la langue.

1.1 Ressources lexicales

Les **dictionnaires**, **lexiques** et **bases de données terminologiques** sont les principales ressources disponibles pour la langue bretonne.

Ils sont utilisés de la même façon qu'un dictionnaire papier par la majorité des utilisateurs, mais ils sont aussi utilisés par des logiciels ou algorithmes pour effectuer des opérations de traitement de la langue : auto-correction, traduction

automatique… à condition qu'ils soient disponibles dans un format numérique et normalisé (TEI, XML…).

Malgré le fait qu'elles soient les plus répandues, les ressources lexicales bretonnes sont vieillissantes et leur qualité parfois moyenne, à l'exception de celles maintenues par l'OPLB, OpenStreetMap.bzh ou encore Wikeriadur.

Les **dictionnaires monolingues** proposent en général des informations diverses : définition, prononciation, exemples, catégorie grammaticale, étymologie, entre autres. Dans les **dictionnaires bilingues**, chaque mot breton correspond à au moins un mot dans au moins une autre langue.

	Quantité	Disponibilité	Qualité	Couverture	Durabilité	Adaptabilité
Dictionnaires monolingues						
Dictionnaires bilingues						
Dictionnaires multilingues						
Bases terminologiques						
Bases toponymiques						

On appelle **base de données terminologique** une base dans laquelle on retrouve des termes spécifiques à certains domaines, qu'on ne retrouve pas habituellement dans les dictionnaires courants.

Comme le montre le tableau ci-dessus, les ressources lexicales pour le breton présentent toutes un **problème d'adaptabilité** : il est difficile, parfois impossible, de les utiliser dans d'autres projets que celui pour lequel elles ont été créées.

Les dictionnaires bilingues existants présentent des **problèmes de qualité** : les informations qui y sont présentées ne sont parfois pas vérifiées ni sourcées, elles sont parfois incomplètes (absence de prononciation, d'indication de niveau de langue, de contexte d'emploi) ; **de durabilité** : certains sont peu documentés et peu mis à jour ; et **d'adaptabilité** : les dictionnaires les plus complets sont disponibles uniquement sous forme PDF. La plupart des dictionnaires sont le fruit du travail d'une seule personne (bénévole) et non le fruit d'une équipe. Si la personne qui gère le projet connait des difficultés, voire même décède, le site

peut tout simplement disparaître. La disponibilité et le maintien des données sur le temps long est une question très importante.

- Sur Internet, on peut trouver deux **dictionnaires monolingues** :

Meurgorf http://meurgorf.brezhoneg.bzh

Dictionnaire historique de la langue bretonne développé par l'OPLB. Il contenait 55 747 entrées le 27 octobre 2020 et propose des définitions, formes dérivées, exemples historiques…

Disponible en ligne, exemples historiques téléchargeables

Wikeriadur https://br.wiktionary.org/wiki/Wikeriadur:Degemer

Version bretonne de Wiktionary, un dictionnaire libre de droits, participatif. Il est incomplet et sa qualité varie suivant les entrées.

Disponible en ligne, il ne peut pas être téléchargé

- On peut également trouver trois **bases terminologiques** :

TermOfis https://br.wiktionary.org/wiki/Wikeriadur:Degemer

Base terminologique de l'OPLB. Base officielle pour la langue bretonne, elle contenait 73 456 termes au 23 novembre 2020.

Disponible en ligne, elle ne peut pas être téléchargée

Brezhoneg21 http://www.brezhoneg21.com/geriadurGB.php

Dictionnaire des sciences et techniques. Il a été conçu pour les écoles Diwan, sur les domaines touchant à l'éducation.

Disponible en ligne, il peut être téléchargé en PDF

Preder https://preder.net/r/geriadur/geriadur.php?locale=fr

Base de données conçue à partir des dictionnaires publiés par Preder.

Disponible en ligne, elle ne peut pas être téléchargée

- Il existe deux **bases toponymiques** :

KerOfis http://www.brezhoneg.bzh/91-kerofis.htm

Développée par l'OPLB, la base fournit les formes normalisées des noms de lieux, communes, rues, etc.

Disponible en ligne, les données peuvent être téléchargées

OpenStreetMap https://www.openstreetmap.bzh/fr/

Version en breton du site de cartographie en ligne libre de droits OpenStreetMap qui reprend une partie des données de KerOfis.

Disponible en ligne, code disponible sur GitHub

- On peut trouver plusieurs **dictionnaires bilingues**, de qualité variable :

Glosbe https://fr.glosbe.com/br/fr

Dictionnaire participatif multilingue dans lequel chacun peut ajouter des mots. On y trouve des mémoires de traduction compilées sur Internet, cependant souvent mal alignées.

Disponible en ligne, il ne peut pas être téléchargé

Geriadur Cornillet-Ménard http://www.brezhoneg.org/sites/default/files/documents/ger.bzh-gal.cornillet-2020.pdf

Dictionnaire bilingue breton-français, construit à partir de plusieurs dictionnaires papier (Geriadur An Here, Dictionnaire français-breton par Martial Ménard…) par Gérard Cornillet.

Téléchargeable au format PDF

Favereau http://www.grandterrier.net/dicobzh/

Ancienne version du dictionnaire Favereau, mis en ligne. L'interface datée complique l'utilisation du dictionnaire.

Disponible en ligne

Favereau https://geriadurbrasfavereau.monsite-orange.fr/index.html

Version plus récente du dictionnaire Favereau

Téléchargeable au format PDF

Devri http://devri.bzh

Dictionnaire diachronique breton-français.

Disponible en ligne, ne peut pas être téléchargé

FreeLang https://www.freelang.com/enligne/breton.php

Dictionnaire créé par Tomaz Jacquet, contient 37 800 entrées.

Disponible en ligne, sous Windows ou sous Android, téléchargeable

FreeDict https://freedict.org/downloads/

Dictionnaire également créé par Tomaz Jacquet, 27 034 mots dans le sens breton-français et 36 017 dans le sens français-breton.

Téléchargeable sous formes normalisées (TEI, ...)

Apertium https://github.com/apertium/apertium-br-fr

Dictionnaire inclus dans le programme Apertium, basé sur le dictionnaire de Tomaz Jacquet et complété par Fulup Jakez.

Téléchargeable sur le GitHub d'Apertium

Tyers, Francis M. 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.

Licence Creative Commons CC BY-SA

1.2 Corpus

Un corpus est une large collection de textes, enregistrements ou vidéos, qui représente un échantillon d'utilisation de la langue. Les corpus sont transcrits

sous forme numérique, classés et soumis à différents traitements avant d'être utilisés pour développer des outils linguistiques.

Les corpus sont très importants pour le traitement automatique des langues. Ils représentent la manière dont une langue est utilisée en réalité. Quand il est assez complet pour couvrir la plupart des cas d'utilisation de la langue on parle de **corpus de référence**. Il existe plusieurs types de corpus : les corpus monolingues qui présentent le fonctionnement de la langue, et les corpus **bilingues** ou **multilingues** qui sont utiles pour comparer deux langues entre elles.

	Quantité	Disponibilité	Qualité	Couverture	Durabilité	Adaptabilité
Corpus monolingues						
Corpus monolingues arborés						
Corpus bilingues parallèles						
Corpus de parole						
Corpus multimédia						

Comme on peut le voir dans le tableau ci-dessus, il n'existe pas suffisamment de corpus bretons pour mener à bien le développement d'outils avancés de traitement de la langue bretonne.

Par leur nombre et leur taille réduits, les corpus monolingues offrent une couverture très incomplète de la langue bretonne.

Bien que peu nombreux, les corpus de parole sont ceux qui présente la meilleure qualité : ils sont accessibles et téléchargeables, sous licence libre de droits, ils sont bien documentés et dans des formats correspondants aux standards en matière de construction de corpus (divisés en plusieurs parties dédiées à l'entrainement, l'évaluation ; emploi de formats standards comme MP3 ou WMA…).

1.2.1 Corpus textuels monolingues

Les corpus monolingues sont utilisés pour construire des modèles statistiques par exemples, ou pour tester des algorithmes. On peut diviser ces corpus en

deux catégories : les **corpus spécialisés**, qui rassemblent des textes présentant un trait particulier de la langue : l'époque, le domaine, le niveau de langue… ; et les **corpus de référence** représentent la langue dans son ensemble.

1.2.1.1 Corpus spécialisés

Il n'existe qu'un seul corpus spécialisé du breton, alors qu'il s'agit pourtant d'une ressource de base. Ces corpus peuvent être lemmatisés[4] ou annotés morpho-syntaxiquement[5].

Leipzig Corpora https://wortschatz.uni-leipzig.de/en/download

Corpus Web, 100 000 phrases extraites du Wikipédia breton, ce qui donne environ 2M de mots. Domaine encyclopédique/internet.

Disponible en ligne, téléchargeable

D. Goldhahn, T. Eckart & U. Quasthoff: Building Large Monolingual Dictionaries at the Leipzig Corpora Collection: From 100 to 200 Languages. *In: Proceedings of the 8th International Language Resources and Evaluation (LREC'12),* 2012

Licence Creative Commons CC BY

1.2.1.2 Corpus arborés

Un **corpus arboré**, ou **corpus syntaxique**, est un corpus de référence qui, en plus des informations grammaticales, contient des informations sur les relations syntaxiques entre les mots. Ce type de corpus est très utile pour la recherche linguistique ou informatique, mais il requiert une somme de travail très importante. Il existe un corpus arboré du breton, de petite taille.

Breton Treebank https://github.com/UniversalDependencies/UD_Breton-KEB

[4] La lemmatisation consiste à remplacer les mots par leur forme non fléchie, par exemple l'infinitif pour les verbes, le masculin singulier pour les noms et adjectifs, etc.

[5] L'annotation morpho-syntaxique consiste à étiqueter les mots avec leur classe grammaticale, ou leur partie-du-discours par exemple.

Petit corpus de 888 phrases (ce qui fait environ 10 000 mots) annoté (catégorie grammaticale et arbre syntaxique). Extrait de livres de grammaire, Wikipédia, Bremaik, de textes de l'OPLB et de chansons.

Disponible en ligne et téléchargeable

Francis M. Tyers and Vinit Ravishankar, *A prototype dependency treebank for Breton*, Actes de la 25e conférence sur le Traitement Automatique des Langues Naturelles (TALN), 2018

Licence Creative Commons CC BY-SA

ARBRES http://arbres.iker.cnrs.fr.

Grammaire descriptive du breton présentant de nombreux exemple annotés au niveau de la catégorie grammaticale et de la structure syntaxique.

Disponible en ligne

Jouitteau, Mélanie. (éd.). 2009-2023. ARBRES, wikigrammaire des dialectes du breton et centre de ressources pour son étude linguistique formelle, IKER, CNRS.

Licence Creative Commons CC BY-NC-SA.

1.2.2 Corpus parallèle bilingue

Un corpus parallèle est un corpus bilingue (ou multilingue) où sont mis en relation les segments (phrases, propositions…) et leur traduction. Ils sont souvent constitués à partir de mémoires de traduction. On utilise ce genre de corpus pour développer des traducteurs automatiques, des logiciels d'aide à la traduction, etc.

OPUS http://opus.nlpl.eu

Corpus breton-français, 400 000 phrases, 3M de mots. Un tiers du corpus provient de mémoires de traduction de l'OPLB (http://opus.nlpl.eu/OfisPublik.php), le reste de traduction de logiciels ou de sous-titres. Il contient des

erreurs d'alignement.

Téléchargeable dans plusieurs formats

Jörg Tiedemann, 2012, Parallel Data, Tools and Interfaces in OPUS. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC'2012)*

Données libres de l'OPLB

http://www.brezhoneg.bzh/211-roadennou-frank-a-wiriou.htm

Corpus de 4 500 phrases de mémoires de traduction de l'OPLB.

Téléchargeable au format texte

Tatoeba

https://tatoeba.org/eng/downloads

7 000 phrases multilingues du breton vers une autre langue, 5600 du breton vers le français et 3700 du français vers le breton.

Disponible en ligne, téléchargeable

Licence Creative Commons CC-BY

1.2.3 Corpus de parole

Pour construire des systèmes de **synthèse de la parole** ou de **reconnaissance de la parole**, il est nécessaire de disposer de données d'entrainement. Un **corpus de parole** est un corpus constitué d'enregistrements accompagnés de leurs transcriptions. Il peut être monolingue ou multilingue. Seul le projet Common Voice a été conçu spécialement pour la reconnaissance vocale.

Mozilla Common Voice

https://voice.mozilla.org/fr/datasets

Corpus de parole participatif. 12 heures enregistrées pour le breton.

Téléchargeable sur le site Common Voice

Ardila, R., Branson, M., Davis, K., Henretty, M., Kohler, M., Meyer, J., Morais, R., Saunders, L., Tyers, F. M. and Weber, G. (2020) "Common Voice: A Massively-Multilingual Speech

	Corpus". *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020).* pp. 4211—4215 Licence Creative Commons CC-0
Banque Sonore des dialectes bretons	http://banque.sonore.breton.free.fr/ Ensemble d'enregistrements effectués auprès de locuteurs traditionnels du breton. Il comprend 8 115 fichiers audio pour une durée totale de 18 heures d'enregistrements. *Disponible en ligne, peut être téléchargé* Licence Creative Commons CC BY-NC-ND
Dictionnaires bretons parlants	http://dico.parlant.breton.free.fr/index.htm Dictionnaire multimedia comprenant des enregistrements effectués auprès de locuteurs traditionnels du breton. Présente une dizaine de points de collecte au 4 novembre 2020. *Disponible en ligne*
Corpus de Dastum	http://www.dastumedia.bzh/dyn/portal/index.seam?page=listalo&firstResult=0&fonds=&fi=LANG&va=BR&aloId=53433&cid=4161 Ensemble de corpus de langue bretonne présents dans la base de données de Dastum. Certains éléments ne sont pas transcrits et ne sont pas destinés à des traitements informatiques. *Disponible en ligne et peut être-téléchargé, après inscription gratuite* Utilisation soumise à autorisation de la part de Dastum

1.3 Grammaires et modèles de langue

Une langue peut être représentée sous forme numérique de deux façons, avec une grammaire formelle, qui décrit l'ensemble des règles grammaticales de la langue sous forme numérique, ou grâce à un modèle de langue, qui représente de manière statistique la langue.

	Quantité	Disponibilité	Qualité	Couverture	Durabilité	Adaptabilité
Grammaire formelle						
Modèle de langue						

Ces deux représentations sont utilisées pour reconnaitre la langue, pour vérifier qu'une phrase fait partie de la langue ou encore pour faire de la correction automatique.

Il existe en breton des grammaires incorporées dans d'autres outils :

- Dans Apertium, pour faire de la désambiguïsation
- Dans LanguageTools pour l'autocorrection

Ces grammaires sont cependant limitées et difficilement exploitables en dehors des projets pour lesquels elles ont été conçues.

1.4 Ressources sémantiques et bases de connaissances

Cette ressource est de haut niveau, elle est utilisée pour des traitements sémantiques (désambiguïsation, analyse de sentiments…).

Il existe plusieurs types de ressources sémantiques :

- Les **Word Nets** sont des bases de données qui regroupent les mots par « *synsets* » (*synonym sets*), des groupes de mots interchangeables.
- Une **ontologie** est une base de données qui contient des relations entre les mots : hyperonymie, hyponymie, synonymie, antonymie, méronymie…
- Un **thésaurus** est une base de données qui regroupe les mots par champ lexical.

Il n'existe pas de Word Net, ontologie ou thésaurus en breton.

2 Technologies de la langue

2.1 Analyse syntaxique, outils de base

Pour effectuer des traitements sur un texte, on commence généralement par utiliser des outils dits de pré-traitement : le **segmenteur** détache les mots les uns des autres et des signes de ponctuation, **l'étiqueteur morpho-syntaxique**

marque la **catégorie grammaticale** des mots et le **lemmatiseur** remplace les mots par leur **lemme** (infinitif du verbe, masculin singulier du nom…)

	Quantité	Disponibilité	Qualité	Couverture	Durabilité	Adaptabilité
Segmenteur						
Lemmatiseur						
Etiqueteur morphosyntaxique						

- **Lemmatiseur, segmenteur**

Il n'existe pas de segmenteur prenant en compte les spécificités du breton et de son orthographe standard (le peurunvan) : prise en compte de « c'h » comme un seul caractère, prise en compte de l'élision des particules verbales…

Un lemmatiseur est utilisé dans le projet Apertium, sans qu'il soit possible de l'extraire pour d'autres usages.

- **Etiqueteur morphosyntaxique**

Les correcteurs orthographiques comme LanguageTools ou encore le traducteur Apertium utilisent des étiqueteurs morphosyntaxiques intégrés. Cependant il est difficile de les utiliser pour un autre travail, et ils n'ont pas été évalués.

POS tagger Apertium	https://beta.apertium.org/index.fra.html?choice=bre#analyzation Etiqueteur morphosyntaxique intégré dans le traducteur automatique Apertium. *Disponible en ligne, téléchargeable sur le GitHub d'Apertium*

Sheikh, Z.M.A.W. and Sánchez-Martínez, F. (2009) "A trigram part-of-speech tagger for the Apertium free/open-source machine translation platform". In *Proceedings of the First International Workshop on Free/Open-Source Rule-Based Machine Translation,* p. 67-74, Alacant, Spain.

Licence Creative Commons CC BY-SA

2.2 Analyse syntaxique

Un analyseur syntaxique permet de représenter la structure de la phrase. Un analyseur syntaxique de surface reconnait les syntagmes[6] nominaux, verbaux, etc. alors qu'un analyseur syntaxique profond peut construire l'arbre syntaxique complet de la phrase et indiquer le rôle de chaque mot.

Il n'existe pas d'analyseur syntaxique pour le breton.

2.3 Analyse sémantique

Les analyseurs sémantiques sont utilisés pour représenter et désambiguïser le sens d'une phrase. Ils prennent le plus souvent la forme d'un arbre sémantique.

Il n'en existe pas pour le breton.

2.4 Extraction d'informations

Un système d'extraction d'informations permet de trouver les informations importantes dans un ensemble de données ou de les résumer. On l'utilise aussi pour catégoriser des textes ou pour repérer des comportements dangereux sur les réseaux sociaux.

Il n'existe pas de tel système pour le breton.

2.5 Génération de textes

Les systèmes de génération automatique de texte mettent sous forme lisible par les humains des informations ou des données. Ils sont par exemple utilisés pour des rapports météo, des moteurs de recherche, etc.

Il n'existe pas de tel système en breton.

[6] Un syntagme est un groupe de mots organisés autour d'un noyau. Ce noyau peut-être un verbe, un nom, etc.

2.6 Reconnaissance de l'écriture

Ce domaine est très utile pour la transcription sous forme numérique de texte anciens, manuscrits ou imprimés. Il n'y a pas encore eu de travaux sur ce sujet en breton.

2.7 Traduction automatique

Il existe aujourd'hui trois types de traducteurs automatiques. Les premiers fonctionnent à base de règles de traduction et de dictionnaires bilingues. Ils nécessitent moins de ressources. Les deux autres types de traducteurs, dits statistiques ou neuronaux, ont besoin d'une grande quantité de textes bilingues alignés pour s'entrainer et apprendre.

2.7.1 Traducteurs automatiques de base

Apertium

https://www.apertium.org/index.fra.html?dir=bre-fra#translation

https://github.com/apertium/apertium-br-fr

Traducteur automatique construit à l'aide de règles et dictionnaires. Toujours en développement.

Disponible en ligne, code téléchargeable sur GitHub

Tyers, F. M. 2010. 'Rule-based Breton to French machine translation', in *Proceedings of the 14th Annual Conference of the European Association of Machine Translation*, 174-181

Licence Creative Commons CC BY-SA

2.7.2 Traducteur automatique avancé (statistique, neuronal, hybride)

Il n'en existe pas encore puisqu'il n'y a pas ou pas suffisamment de ressources nécessaires à leur élaboration (des corpus bilingues).

2.8 Traitement de la parole

2.8.1 Synthèse de la parole

L'élaboration de la synthèse de la parole du breton est en cours. Le projet a été lancé par l'OPLB et la réalisation technique a été confiée à l'IRISA. La livraison est prévue courant 2021.

2.8.2 Reconnaissance de la parole

Il n'existe aucun outil de reconnaissance de la parole en breton, mais des travaux préliminaires ont déjà été entamés par Francis Tyers à l'Université de l'Indiana (États-Unis), à partir des données de Common Voice.

3 Conclusion

De nombreux outils et ressources sont disponibles pour le breton, sur internet. Cependant beaucoup d'entre eux sont vieillissants et ne sont plus mis à jour. En général ils n'ont pas fait l'objet d'évaluation. Entre ceux qui ont été abandonnés et ceux dont l'existence reste inconnue du plus grand nombre, il est parfois difficile de trouver l'outil que l'on recherche. Dans ce diagnostic, nous n'avons listé que les outils qui étaient encore en état de fonctionner.

Du point de vue des ressources linguistiques, nous disposons d'une bonne base de départ mais il devient urgent de travailler tant sur l'augmentation de la quantité de ressources disponibles que sur leur qualité et leur diversification.

❖ Partie 3. Préconisations

A la suite du diagnostic fait précédemment, nous pouvons établir une liste de préconisations à l'ensemble des acteurs engagés dans les technologies de la langue pour le breton. Ces différentes actions seront à répartir entre les différents acteurs suivant leurs compétences et leurs moyens d'action.

1 Augmenter sensiblement la visibilité du breton dans le monde numérique

1.1 Les contenus en langue bretonne sur Internet

La visibilité du breton passe d'abord par la disponibilité de contenus utiles et en plus grande quantité sur les principaux sites d'Internet. Parmi eux, l'un des plus consultés est Wikipédia.

Il parait également nécessaire que les structures qui proposent des sites bilingues français-breton mettent mieux en avant l'utilisation du breton et facilitent la navigation dans un environnement entièrement en langue bretonne.

- **Sensibiliser le public brittophone** à l'utilisation de Wikipédia en breton
- Recenser les articles les plus importants de Wikipédia et **publier des traductions de qualité** basées sur leur version anglaise ou française.
- Encourager la traduction de **sites culturels** (cinéma, musique, sport…) ou **d'actualité** (Presse régionale, portails d'informations, météo, horaire des marées…) qui sont les plus consultés par le public.
- **Former et coordonner** les différents acteurs qui prennent part à la traduction de contenu et de sites Internet.
- Veiller à la **qualité de la langue** utilisée (orthographe, syntaxe, terminologies adaptées)
- Publier un guide des **bonnes pratiques de traduction** pour aider à mettre en place et à mettre en valeur la langue bretonne sur les sites bilingues français/breton.

1.2 L'utilisation du breton sur les médias sociaux

Même si l'utilisation du breton sur les réseaux sociaux est de plus en plus fréquente, elle reste cependant trop rare, même pour des utilisateurs qui parlent couramment la langue. La première raison qui pourrait expliquer une si faible utilisation du breton est l'absence d'interface bretonne pour la plupart des sites

et applications, disponibles en français et en anglais. L'environnement dans lequel évoluent les utilisateurs n'est pas fait pour les inciter à utiliser d'autres langues.

D'autres hypothèses qui peuvent être avancées sont l'isolement des locuteurs entre eux ainsi que l'insécurité linguistique des locuteurs quant à leurs compétences linguistiques. Sur la plupart des réseaux, il n'existe pas de communauté brittophone ni de moyen de rencontrer d'autres brittophones, ni même d'identifier des brittophones.

- **Publier un lexique des médias sociaux** pour encourager les utilisateurs à employer la langue sur internet.
- **Créer un évènement** autour des médias sociaux pour permettre aux utilisateurs de se rencontrer et de publier du contenu en langue bretonne.
- **Participer à la traduction** des plateformes de réseaux sociaux quand c'est possible ou inciter à leur traduction.

1.3 Des outils pour l'apprentissage et la mise en valeur du breton

L'une des plus fortes demandes de la part du grand public est le développement d'applications et d'une plateforme web permettant l'apprentissage et le perfectionnement en langue bretonne. C'est une demande qui vient à la fois du public non-brittophone que du public déjà brittophone. L'enquête TMO-Région de 2018 a montré que 30 % des Bretons qui ne parlent pas la langue souhaiterait la connaître.

Pour être efficaces, les outils d'apprentissage du breton doivent intégrer une réelle mise en valeur de la langue et utiliser les dernières innovations technologiques, aussi bien pour ce qui est du contenu que de la façon de la présenter.

- Mettre en place une plateforme d'autoapprentissage publique en ligne
- **Participer à la traduction** de logiciels pédagogiques utilisés dans les écoles (Traitement de texte, système d'exploitation bilingues…).
- **Mener des actions en faveur de la présence du breton** sur DuoLinguo ou d'autres application d'initiation ou apprentissages de nouvelles langues.
- **Développer et proposer du contenu pédagogique moderne** accessible à tous pour apprendre ou appronfondir ses connaissances sur la langue bretonne.

1.4 La traduction d'applications

A l'heure où les dispositifs numériques occupent une place centrale dans nos vies, que ce soit au travail, à la maison ou lors de nos sorties, il est essentiel de les prendre en compte dans la constitution d'un environnement brittophone.

Il est aujourd'hui difficile voire impossible d'activer son GPS en breton, d'utiliser son smartphone en breton, de naviguer sur son ordinateur en breton ou d'utiliser des bornes multimédia en breton, dans les gares ou les banques.

- Entrer en contact avec Microsoft pour la **traduction de leur suite logicielle** bureautique Office
- Entrer en contact avec les développeurs des **principaux systèmes d'exploitation** pour ordinateur et pour smartphone, afin d'étudier la faisabilité de versions bretonnes.
- Participer à la **traduction collaborative en ligne** d'applications majeures
- Proposer des solutions pour **utiliser son téléphone en breton** : clavier, messagerie, navigation, GPS, application…
- **Sensibiliser les grandes entreprises** afin qu'elles proposent leurs services numériques en langue bretonne

2 Encourager l'innovation numérique au service de la langue bretonne

2.1 Mettre en réseau les différents acteurs

La Bretagne dispose sur son territoire de nombreux laboratoires et centres de recherche en linguistique et en informatique, dont plusieurs spécialisés en traitement automatique de la langue : les équipes de l'IRISA à Rennes et Lannion, le LS2N à Nantes…

Cependant les acteurs travaillant sur la langue bretonne dans le domaine numérique sont aujourd'hui peu nombreux. On peut compter parmi eux l'OPLB, qui met à disposition sur son site plusieurs outils numériques qu'il a développé et qui a participé à la traduction de plusieurs logiciels, ainsi que des associations comme An Drouizig qui propose des correcteurs orthographiques et des traductions de logiciels, ou encore OpenStreetMap.bzh qui propose un système de cartographie numérique entièrement en breton.

Une première initiative a permis de mettre en relation ces différents acteurs via le réseau *Bed Niverel* en 2017. A l'époque avaient pris part à la rencontre l'OPLB, An Drouizig, OpenStreetMap.bzh, Roued, EduBreizh et des informaticiens

d'Arkea. Même si cette rencontre n'a pas donné lieu à de nouvelles coopérations autours de projets numériques, elle reste néanmoins une base intéressante pour la constitution d'un réseau plus actif, en langue bretonne.

Nous proposons qu'une future rencontre sur un modèle similaire soit organisée afin de faire le point sur les projets des différentes structures, leur avancement. Aussi, chacune de ces structures dispose de ressources linguistiques qui pourraient être mises en commun. L'OPLB pourrait avoir un rôle de coordination des projets et des acteurs.

Cette rencontre permettrait d'envisager les bases d'une coordination plus reserrée entre les acteurs pour la mise en place de nouvelles ressources et de nouveaux outils, en se basant sur le plan d'action détaillé en partie 3 de ce dossier.

- **Organiser des rencontres** entre les acteurs du numérique brittophone, en reprenant par exemple l'initiative Bed Niverel.
- Coordonner la mise en place du **plan d'action pluriannuel**.
- **Construire des liens** avec le monde universitaire et de la recherche.

2.2 Publier des données de qualité

Les données sont la matière première dans les processus de développement de technologies de traitement des langues naturelles. Elles sont utilisées à la fois pour alimenter des algorithmes et des intelligences artificielles, pour étudier des phénomènes linguistiques ou pour tester les performances d'un programme de traitement de la langue.

Les données peuvent être de plusieurs nature : sémantiques, grammaticales, textuelles, audio, vidéo, multimodales… Elles sont organisées différemment et parfois enrichies d'informations supplémentaires (annotations, métadonnées).

Aujourd'hui la quantité et la qualité des données numérisées disponibles en langue bretonne posent problème. Il n'existe à ce jour pas de corpus de référence pour le breton, constitué de textes annotés lexicalement et syntaxiquement, qui puisse servir de base de référence pour les linguistes et les informaticiens. De plus, il n'existe aucun corpus dont la qualité ait été évaluée.

L'OPLB dispose cependant d'une assez grande quantité de donnée de diverses natures : des corpus de textes bilingues, un dictionnaire historique, une base de données terminologiques ainsi qu'une base de données toponymiques. Néanmoins ces données sont à l'heure actuelle parfois difficilement accessibles et non formatées pour une utilisation informatisée.

Pour remédier à la dispersion des données, l'OPLB est en train de metttre au point un portail web des technologies numériques pour la langue bretonne. Il accueillera les données et les outils de l'OPLB mais a aussi vocation à accueillir les outils et ressources linguistiques numériques de ceux qui le souhaitent.

- Faire appel à des **spécialistes en linguistique de corpus** pour créer des corpus de haute qualité exploitables de manière informatique.
- Rendre les données présentes dans les bases Meurgorf, TermOfis et KerOfis **accessibles via une API.**
- Evaluer la **qualité des données** disponibles pour le breton.

2.3 Financer l'innovation en breton

L'un des leviers sur lesquels il est possible d'agir pour encourager le développement de nouvelles technologies de la langue pour le breton est le financement de la recherche.

- **Encourager la mise en place d'une bourse** publique spécifique pour les étudiants ou les doctorants qui mènent un projet liant numérique et langue bretonne, ou permettant la création de nouvelles ressources linguistiques.

2.4 Promouvoir les licences libres de droit

Les différentes applications et données développées pour le breton doivent être publiées avec des licences claires qui définissent leurs conditions d'utilisation et de modification. Dans l'optique de leur utilisation par le plus grand nombre et qu'elles servent de base à des développements plus poussés par des entreprises, les licences libres de droits avec droit d'utilisation commerciales sont les plus intéressantes.

- Elever l'ensemble des outils et ressources de l'OPLB à des **licences libres de droits** ouvrant le droit à la modification pour usage non-commercial ou commercial.
- Veiller à ce que les projets financés par l'argent public soient **librement accessibles au plus grand nombre** et restent libres de droits.

❖ Partie 4. Plan d'action

L'amélioration de la place de la langue bretonne dans le monde numérique et le développement de nouvelles applications grand public passent aussi par l'amélioration ou le développement de nouveaux outils et ressources à destination des développeurs, chercheurs et linguistes.

1 Ressources linguistiques

Le tableau ci-dessous présente l'état actuel des ressources linguistiques disponibles pour le breton, et leur état souhaité dans les 9 prochaines années, par période de 3 ans.

La couleur **verte** signifie que les ressources disponibles sont satisfaisantes, la couleur **orange** signifie que les ressources doivent être améliorées ou enrichies, ou qu'elles sont en cours de développement ; la couleur **rouge** signifie que les ressources sont manquantes ou insuffisantes, ou que leur développement n'est pas encore lancé.

Ressources		2020	2021-2023	2024-2026	2027-2029
Corpus textuels monolingues	Web				
	Langue				
	Arboré				
Corpus textuels bilingues					
Corpus de parole					
Grammaires					
Modèles de langue					
Lexiques					
Dictionnaires					
Bases terminologiques					
Bases toponymiques					
Bases sémantiques					

1.1 Corpus

Le **corpus web** actuel issu de Wikipédia contient 100 000 phrases, il peut être augmenté mais il se heurte à la quantité d'articles disponibles en breton sur le site.

La constitution d'un **corpus de langue général** est une priorité. L'objectif est d'atteindre un corpus de **20 millions de mots** courant 2024-2026. Pour alimenter le corpus, on peut envisager de collecter des textes auprès des maisons d'éditions, des associations ou parmi les exemples historiques de la base Meurgorf. Les différents textes de ce corpus devront être catégorisés suivant leur genre, la qualité de la langue et la date de publication. Pour assurer la bonne qualité et la pertinence de ce corpus, il devra être contrôlé par des linguistes.

La constitution d'un **corpus de référence** prend plus de temps, car il doit être annoté à la main. Il reste néanmoins très important pour le développement d'outils linguistiques performants, de traduction automatique ou de reconnaissance de la parole par exemple. À l'horizon 2027-2029 nous pouvons fixer l'objectif de **30 000 à 50 000 mots** (actuellement près de 10 000 mots)

Le **corpus bilingue français-breton** de l'OPLB contient actuellement environ 1 million de mots. L'objectif pour 2021-2023 est de le porter à **5 millions de mots**, puis de **15 à 20 millions de mots** d'ici 2027-2029. Ce corpus est actuellement constitué d'une partie des mémoires de traduction de l'OPLB. En collectant toutes les mémoires de traduction disponibles, il est possible d'augmenter son volume rapidement. Il faudra cependant veiller à la diversité du corpus en termes de type de texte.

Le **corpus de parole** Common Voice doit encore croître sensiblement pour permettre les premiers travaux sur la reconnaissance vocale. Il devra être alimenté par une grande diversité de locuteurs. Le premier objectif est de **diversifier le plus possible les locuteurs** et d'aboutir à **plusieurs dizaines d'heures d'enregistrement** (actuellement 8 heures). Pour être en mesure de construire un système de reconnaissance vocale, le corpus devra **atteindre les 200 heures** à l'horizon 2024-2026.

Un corpus de parole standardisé pour la synthèse vocale devrait être disponible en 2021, suite au travail de l'OPLB-IRISA sur le projet de synthèse vocale du breton.

1.2 Grammaires et modèles de langue

Des éléments de grammaires à base de règles ont déjà été élaborés notamment dans le projet Apertium. Cependant leurs performances n'ont pas été évaluées et il faudrait les extraire et les améliorer pour aboutir à une grammaire efficace et générale.

En ce qui concerne les modèles de langue, le préalable est la constitution d'un corpus monolingue de taille suffisante. Le travail doit donc d'abord s'orienter vers l'élaboration de ce type de corpus.

1.3 Dictionnaires et ressources lexicales

Les dictionnaires, lexiques et bases de données terminologiques devront continuer à être enrichis de nouvelles entrées (mots et unités de traduction). Les bases terminologiques devront également s'élargir à de nouvelles langues, l'anglais étant indispensable pour le développement numérique. L'OPLB pourrait travailler à l'élaboration d'un **dictionnaire bilingue/multilingue de référence dès 2023-2025**.

Les bases toponymiques doivent continuer leur enrichissement également. Une base de données anthroponymique pourrait être créée également, regroupant prénoms, noms et personnages historiques.

Le plus gros chantier en termes de ressources lexicales est la constitution de bases sémantiques. Pour le breton on peut imaginer s'appuyer sur le dictionnaire historique de Meurgorf en ajoutant des informations sémantiques pour chaque entrée. Les termes devront aussi être regroupés par ensembles de synonymes. A partir de ces travaux, il sera possible de commencer la constitution d'une ontologie, un système permettant de classer les mots de la langue bretonne suivant leur sens.

2 Technologies de la langue

Comme pour les ressources linguistiques, le tableau suivant présente l'état actuel de la disponibilité d'outils linguistiques pour le breton, et leur état souhaité dans les 9 prochaines années, par période de 3 ans.

La couleur **verte** signifie que les outils disponibles sont satisfaisants, la couleur **orange** signifie que les outils doivent être améliorés ou enrichis, ou qu'ils sont en cours de développement ; la couleur **rouge** signifie que les outils sont manquants ou insuffisants, ou que leur développement n'est pas encore lancé.

Technologies de la langue	2020	2021-2023	2024-2026	2027-2029
Analyse morpho-syntaxique				
Analyse syntaxique				
Analyse sémantique				
Traduction automatique				
Synthèse de la parole				
Reconnaissance de la parole				
Extraction d'informations				
Génération de textes				

2.1 Outils d'analyse

Les outils d'analyse morpho-syntaxique (étiqueteurs grammaticaux) doivent être développés et rendus disponibles de manière **autonome dès 2021-2023**, dans l'objectif d'aboutir à **un outil performant en 2024-2026**. Ils devront également faire l'objet d'une évaluation de leurs performances.

A partir de là, il pourra être développé un outil d'analyse syntaxique de surface. L'élaboration d'un outil d'analyse syntaxique profonde, qui peut être construite à base de données statistiques et d'intelligence artificielle devra attendre la disponibilité de corpus arborés prévu pour 2027-2029.

Si les outils d'analyse sémantique ne sont pas une priorité, il peut néanmoins être intéressant de travailler sur des premières versions **dès 2024**, au fur et à mesure que les bases de données sémantiques s'étoffent.

2.2 Traduction automatique

Le logiciel de traduction Apertium pourra être amélioré et développé **dès 2021-2023**. Il devra également être effectué une évaluation des performances

d’Apertium pour la paire breton-français. **A partir de 2024** il sera possible de se baser sur le corpus bilingue français-breton pour développer un système de traduction automatique à base statistique.

2.3 Traitement de la parole

Le premier projet de synthèse de la parole verra le jour **en 2021**, suite au travail commandé par l’OPLB à l’IRISA. La livraison des corpus et des algorithmes pourra également servir de base de travail à d’autres projets.

Les outils de reconnaissance de la parole sont conditionnés à la disponibilité d’un corpus de parole suffisamment grand, du type de Common Voice. Les premiers développements pourraient donc commencer **dès 2024-2026**.

3 Mise en œuvre du plan d'action

3.1 Ressources matérielles, humaines et financières

Le tableau suivant permet d'évaluer la quantité et le type de ressources nécessaires à réalisation de chaque ressource ou outil. Il est présenté à titre indicatif, et son contenu peut varier suivant les structures en charge du développement de certains outils ou ressources, le recours au bénévolat, l'utilisation de l'existant, les budgets dégagés, etc.

	Ressources matérielles	**Ressources humaines**	**Ressources financières**
Constitution de corpus écrits	Des ensembles de textes du domaine public ou dont les droits ont été achetés ou légués. Des mémoires de traductions.	Des techniciens ayant été formées ou possédant des connaissances à la fois sur la langue bretonne et la linguistique de corpus. Un travail de constitution de corpus peut être mené par une équipe de deux ou trois personnes.	L'aquisition de droits pour des textes peut s'élever à plusieurs dizaines de milliers d'euros. La constitution d'un corpus nécessite relativement peu de temps, mais dépend du besoin de transcription de texte en version numérique ou dans l'orthographe standard. Le financement nécessaire équivaut à celui d'un poste de 1 à 3 mois.
Annotation de corpus	Des corpus	Des techniciens maitrisant la	L'annotation de corpus est un

	textuels. Un logiciel d'annotation de texte.	langue bretonne.	travail chronophage et donc coûteux. Il faut compter environ 150 à 200 heures pour l'annotation d'un corpus de 50 000 mots.
Enregistrement de corpus de parole	Un studio d'enregistrement ou Une plateforme d'enregistrement en ligne.	Des locuteurs, qui peuvent être des bénévoles. Un ingénieur capable de paramétrer l'enregistrement et de traiter les données obtenues.	Pour une heure d'enregistrement, il faut compter environ 1h à 1h30 de traitement dans le cas d'un corpus de synthèse de la parole, mais beaucoup moins dans le cas d'un corpus destiné à la reconnaissance vocale.
Ecriture de grammaires et conception de modèle de langue	Ressources grammaticales. Corpus de textes conséquent.	Un spécialiste de la grammaire bretonne, capable de la décrire sous forme de grammaire formelle. Un spécialiste du traitement des langues, capable de transposer ces règles sous forme numérique ou de générer un	Le développement de grammaires formelles ou de modèles de langue peuvent prendre environ 3 à 6 mois (ou plus s'ils ne sont pas développés par des spécialistes, difficiles à trouver pour le breton).

		modèle de langue.	
Elaboration d'un dictionnaire bilingue	Ressources lexicales.	Un développeur informatique Une équipe spécialisée pour alimenter le dictionnaire.	Le développement de la structure d'un dictionnaire peut faire l'objet d'un poste de travail pour une durée de 1 à 3 mois. Le développement de bases sémantiques peut faire l'objet d'un travail de thèse ou équivalent. Les dictionnaires et bases sémantiques doivent ensuite faire l'objet d'une alimentation et d'une maintenance régulière sur des années.
Des outils d'analyse de texte	Des corpus de textes annotés. Des grammaires et modèles de langue.	Chercheurs ou développeurs spécialisés dans le traitement automatique des langues.	Le développement d'outils d'annotation automatique grammatical ou syntaxique peut faire l'objet d'un poste sur environ 6 mois.
Système de traduction automatique	Corpus bilingue parall-le français-	Chercheurs ou développeurs spécialisés dans le	Un tel système peut prendre plusieurs années

avancé	breton. Ordinateur disposant de ressources suffisantes pour faire tourner des systèmes d'intelligence artificielle.	traitement automatique des langues.	(1 à 3 ans) pour une première version, et mobiliser une équipe de recherche ou de développement. On peut compter 150 000 à 200 000 euros par année de développement.
Reconnaissance de la parole	Corpus de parole pour la reconnaissance de la parole Ordinateur disposant de ressources suffisantes pour faire tourner des systèmes d'intelligence artificielle	Chercheurs ou développeurs spécialisés dans le traitement automatique des langues et en traitement du signal audio	Un tel système peut prendre plusieurs années (2 à 3 ans) pour une première version, et mobiliser une équipe de recherche ou de développement. On peut compter 150 000 à 200 000 euros par année de développement.

3.2 Acteurs concernés

Ce plan d'action repose sur les ressources et outils existants aujourd'hui en langue bretonne. Ils offrent un bon point de départ pour les développements futurs.

Le bon déroulement de cette stratégie demande un effort de coordination entre les différents acteurs du numérique en langue bretonne, de la part de ceux qui sont déjà impliqués comme de celle de ceux qui sont amenés à s'impliquer à l'avenir.

Il demande en outre un investissement en personnes, temps et crédits de la part des universités et instituts de recherche qui aujourd'hui ne travaillent pas suffisamment dans le domaine du traitement automatique de la langue bretonne.

Nous appelons dès à présent l'ensemble des acteurs développant des compétences et souhaitant participer à un effort commun pour traduire dans les faits cette stratégie à entrer en contact avec l'OPLB.

Enfin la mise en place de ce plan repose également sur un investissement important des pouvoirs publics bretons en vue de développer la place de la langue dans le numérique.