Lapurdum
Revue d'études basques

24 | 2023
Numéro XXIV

Guide de survie des langues minorisées à l'heure de l'intelligence artificielle : Appel aux communautés parlantes

Mélanie Jouitteau

Édition électronique
URL : https://journals.openedition.org/lapurdum/4500
DOI : 10.4000/127u1
ISSN : 1965-0655

Éditeur
IKER

Édition imprimée
Date de publication : 14 décembre 2023
Pagination : 199-217
ISBN : 978-2-9590077-1-2
ISSN : 1273-3830

Référence électronique
Mélanie Jouitteau, « Guide de survie des langues minorisées à l'heure de l'intelligence artificielle : Appel aux communautés parlantes », *Lapurdum* [En ligne], 24 | 2023, mis en ligne le 01 juillet 2024, consulté le 18 octobre 2024. URL : http://journals.openedition.org/lapurdum/4500 ; DOI : https://doi.org/10.4000/127u1

Guide de survie des langues minorisées à l'heure de l'intelligence artificielle : Appel aux communautés parlantes

MÉLANIE JOUITTEAU
IKER, CNRS (Université de Pau et des Pays de l'Adour, Université Bordeaux Montaigne)[1]
melanie.jouitteau@iker.cnrs.fr

Résumé :

Cet article s'adresse à la fois aux chercheurs en sciences humaines et aux structures de politique linguistique des langues minorisées. Il comprend un résumé compréhensible de la chaîne de développement du numérique pour le Traitement Automatique des Langues (TAL), avec une attention particulière pour ces langues qui n'ont pas des corpus aussi vastes que le français ou l'anglais. Dans le contexte de la révolution de l'Intelligence Artificielle (IA), dont j'explique les enjeux pour ces les langues à corpus restreint, je liste les recommandations indispensables au développement de ressources pérennes à même d'assurer un développement durable, avec une liste des pièges et mécompréhensions les plus courantes. L'article prend le breton comme exemple illustratif d'une langue à corpus restreint démontrant des ressources et applications numériques émergentes. Ses conclusions s'appliquent largement à toutes

1 La partie de cet article qui concerne l'intelligence artificielle a une version antérieure publiée en breton sur le site du *Peuple Breton/Pobl Vreizh*. La conception de cet article a bénéficié grandement d'échanges avec les acteurs des ANRs DIVITAL et AUTOGRAMM, l'équipe de création de ressources de la plate-forme DORANUM, ainsi que Pierre-Alexis Michaud (Lacito), et les développeurs indépendants Reun Bideault et Gweltaz Duval-Guennoc. Je tiens ici à les en remercier chaleureusement. Il va sans dire que ces collègues ne sont en rien responsables des éventuelles imprécisions ou erreurs qui pourraient persister. N'hésitez pas à contacter l'auteur à leur propos.

les langues à corpus restreint dont le développement numérique est amorcé, mais encore insuffisant pour profiter directement des applications d'IA développées actuellement pour les langues à corpus massif.

1. Introduction

Nos générations scientifiques sont à la charnière historique du déploiement du numérique. Les chercheurs seniors ont produit leurs travaux de thèse hors numérique. En l'espace d'un quart de siècle seulement, nous avons vu naître les corpus numériques et leurs nouveaux usages (Casenave 2003). Ces corpus ont nourri les larges modèles de langues de l'intelligence artificielle. Début 2023, il est certain que cette révolution va toucher encore plus profondément le cœur de nos modes de communication, et de nos métiers. L'intégration des technologies numériques dans les échanges entre humains n'en est qu'à ses balbutiements. Les traducteurs multilingues (Sayers 2021) et les outils reposant sur l'intelligence artificielle vont révolutionner les pratiques, offrir des possibilités inédites pour les locuteurs des langues répandues. Pour les langues peu développées en numérique, cela va aussi ouvrir quelques perspectives inédites, mais aussi des dangers brutaux. Les langues qui ne pourront pas suivre cette mutation digitale subiront de lourdes pertes de pratiques, potentiellement fatales pour des langues déjà socialement fragilisées. L'accessibilité par et pour ces usages émergents est un enjeu considérable pour les langues à corpus restreint, de l'ordre de la survie numérique.

L'article est composé comme suit. La seconde section est une synthèse de ce qu'est la chaîne de développement des applications numériques. Le développement du TAL est un flux. Il suit un ordre particulier de trois étapes sans la compréhension duquel les communautés linguistiques se condamnent à épuiser en vain leurs précieuses énergies et ressources. La troisième section synthétise le contenu de la révolution de l'intelligence artificielle à date du début 2023 pour expliciter son impact pour le développement numérique des petites langues. La quatrième section présente brièvement les pièges les plus courants rencontrés par nos communautés pour le développement TAL. Je fournis un résumé synthétique des recommandations pour le développement des corpus, socle indispensable pour le développement, et les pratiques vertueuses de science ouverte qui en permettent l'exploitation durable et efficace.

2. Les trois étapes du développement TAL

Le développement TAL est construit en trois étapes, dont l'ordre global ne peut pas être inversé :

(i) ressources
(ii) outils informatiques
(iii) applications

Les **ressources** sont les données produites par la communauté parlante, elles sont constituées de la réalité des productions linguistiques que l'on peut capturer sous forme

numérique. Ce sont les corpus de données brutes ou enrichies. Mieux ces corpus seront organisés, plus largement ils seront accessibles, et plus il sera aisé d'en développer des outils informatiques. Le développement des ressources dépend largement de l'organisation de la communauté linguistique.

Les **outils informatiques** sont développés à partir des corpus par des informaticien·ne·s. Ces outils informatiques sont des robots simples ou complexes. Ce sont les pièces détachées du développement numérique. Ces pièces détachées vont permettre de construire des applications. Mieux ces pièces détachées vont fonctionner et plus il sera aisé de développer des applications. Ils peuvent aussi servir à rebours à augmenter ou améliorer les corpus.

Les **applications**, finalement, constituent tout ce qui est utilisable directement par le public. Les applications qui fonctionnent bien au niveau technique sont évaluables selon leur utilisabilité, qui dépend de leur distribuabilité technique (disponible sur smartphone ou pas ? utilisable sans connaissances informatiques ou pas ?), et de la rencontre sociale avec les usages des locuteurs. Dans le cas des langues à corpus restreint, les usages des applications sont la plupart du temps largement installés dans une langue majoritaire pour qui le développement TAL est plus avancé.

2.1. Les applications du TAL

Voyons tout d'abord ce que recouvrent les applications en traitement automatique des langues. Il permet en 2023 un vaste éventail de tâches :

Analyse de texte.

Elle permet la reconnaissance de mots, de phrases, limites de phrases et de sections, la reconnaissance des caractéristiques morphologiques des mots, des rôles syntaxiques et sémantiques, des relations qui lient les constituants d'un texte.

Traduction automatique.

Elle permet la traduction d'une langue naturelle vers une autre (sur un smartphone, ou pour traduire des fichiers de sous-titres de films).

Extraction d'informations.

Le système extrait des informations appropriées à partir de documents non structurés (comme Internet).

Génération automatique de texte.

Le système résume ou reformule un texte, synthétise ou pose des questions appropriées sur un texte (génération de résumés scientifiques ou médiatique, d'articles sur wikipedia)

Le traitement automatique de la parole permet aux humains de communiquer avec des appareils électroniques par directement par la voix. Le traitement automatique de la parole comporte deux domaines distincts, techniquement différents :

La synthèse de la voix, où l'ordinateur ou le smartphone lit oralement un texte écrit (la voix du GPS, d'une compagnie de transports). Éventuellement, le système peut reproduire une voix humaine spécifique.

La reconnaissance automatique de la parole, où l'ordinateur ou le smartphone prend la langue en dictée. Éventuellement, le système reconnaît le locuteur individuel. Cette application permet de dicter un SMS, ou de générer la transcription intégrale d'un derush de film.

Interaction Homme-Machine, les chatbots.

Le système permet de converser avec des ordinateurs à partir de texte, parole, gestes ou expressions faciales (*Siri, Alexa* qui interagissent avec les humains, ou encore ChatGPT et les systèmes qui intègrent cette technologie).

Les tâches des différentes applications que l'on vient de voir peuvent être mises en relation, et intégrées dans une même application : l'intégration de la reconnaissance vocale et de la traduction automatique permettent de faire un commentaire de match sportif en direct dans une langue minorisée tout en la sous-titrant en live dans une langue majoritaire. Chacune de ces applications nécessite pour son développement efficace l'existence d'une boite à outils informatiques adaptés à la langue, que nous voyons maintenant.

2.2. Kit de base des outils informatiques nécessaires

Mettons tout d'abord de côté le CLDR (*Common Locale Data Repository*) d'Unicode, le lexique nécessaire aux applications qui regroupe l'ensemble des paramètres régionaux à destination des applications informatiques. Ce lexique est indispensable à la traduction d'applications qui existent déjà dans d'autres langues. A part ce lexique, les outils nécessaires à la construction des applications sont des robots spécialisés sur une tâche particulière :

Reconnaisseur de caractères : permet de reconnaître à partir d'une image ou d'un pdf les caractères graphiques d'une langue donnée. Ce robot est nécessaire pour que des documents scannés deviennent cherchables par mot-clef, ou traitables numériquement. C'est un accélérateur de constitution de ressources puisqu'il fournit du corpus numériquement traitable.

Identifieur de langue : identifie la langue dans un texte ou un ensemble de textes qui peut être multilingue. Il permet d'en reconnaître des occurrences au milieu d'autres données non-pertinentes. C'est très important pour constituer des corpus numériques à partir de données massives en ligne. C'est donc aussi un accélérateur de constitution de corpus

traitables puisque cette « épuisette » permet de remonter du texte laissé épars en ligne.

Tokeniseur : découpe un corpus en unités minimales. Si un mot comporte un espace, comme les formes du verbe 'avoir' en breton de type *en deus*, le tokeniseur sait reconnaître qu'il s'agit d'un seul mot.

Lemmatiseur : regroupe les différentes formes de mots sous une seule qui correspond à une entrée de dictionnaire. Ce sont toutes ses formes conjuguées, ou en langues celtiques ses formes mutées.

Taggeur : assigne automatiquement une catégorie grammaticale à chaque élément.

Moteur de règles : ensemble de règles prédéfinies, paramétrisées pour une langue donnée. Il nécessaire dès les premiers correcteurs grammaticaux.

Concordancier multilingue : permet de traiter des corpus parallèles

Analyseur morphologique : C'est un parseur qui analyse les mots dans le but d'obtenir les relations coexistant entre ses sous-parties.

Analyseur syntaxique : C'est un parseur qui analyse une chaîne de mots dans le but d'obtenir les relations coexistant entre ces mots.

Réseaux de neurones (*deep learning*) : architectures complexes de réseaux de circuits modélisés sur le fonctionnement électrique d'un neurone. Ces « neurones » artificiels sont ordonnés en de nombreuses couches de successives, une technique utilisée pour modéliser des structures sous-jacentes de données, surtout à partir de corpus très larges.

Le kit de base des ressources langagières pour les développeurs informatiques du TAL est parfois abrégé en BLARK (*Basic LAnguage Resource Kit*, Krauwer 2003). Un BLARK complet de ces outils permet à une équipe informatique de développer de bonnes applications et, avantage non-négligeable, sans nécessiter que cette équipe parle la langue. Quelques consultations ponctuelles avec une personne parlant elle-même la langue peuvent suffire. L'expertise langue est en fait déployée en amont, dans la construction des corpus. Chacun de ces outils informatiques, chacune de ces pièces détachées d'applications ne peuvent en effet être développés qu'à partir de corpus traitables dans la langue qui ont été rendus largement disponibles et facilement accessibles aux développeurs informatiques.

Heureusement, et au moins pour l'écrit, un même corpus peut servir au développement de différents outils.

2.3. Les types de ressources de corpus nécessaires

La ressource de base du développement TAL est l'ensemble des **corpus** disponible

dans la langue en question. Il existe une méthodologie de diagnostic précis des ressources TAL pour une langue donnée (Ceberio et al. 2018). Cette méthodologie est utile pour développer un plan de développement circonstancié, et dans les cas de recours à des financements externes. Cependant, pour les langues peu dotées comme le sont les langues minorisées de l'État français, vous pouvez sans vous tromper considérer que le développement nécessite des corpus : des corpus bruts, des corpus parallèles et des corpus annotés, tels qu'on va les définir ci-dessous.

On distingue tout d'abord les corpus de **données brutes**. Ce sont les données monolingues, au kilomètre, directement telles que produites par les locuteurs. Il peut s'agir d'enregistrements de collectages, d'interviews, de bande-son de films, archives de radios locales, enregistrements de cours et conférences, etc., et pour l'écrit d'articles de journaux, romans et nouvelles, comptes-rendus publics de réunions, décisions collégiales, transcription de conférence, manifestes, versions corrigées de copies scolaires, textes de chanson, wikipedia dans la langue, articles de recherche dans la langue, etc.

Leur traitement obtient des **corpus enrichis**, qui comportent des informations supplémentaires nécessaires à leur traitement pour développer des outils. Avec l'évolution des technologies, de moins en moins d'enrichissement reste nécessaire au processus, mais le démarrage des outils en nécessite toujours. Un corpus écrit peut être enrichi par sa traduction dans une autre langue, préférablement une langue à corpus large comme l'anglais ou le français. Un corpus oral peut être transcrit, ce qui est une forme de traduction dans une autre modalité. On parle alors de **corpus parallèle**, puisqu'il organise une correspondance entre un corpus monolingue et sa traduction dans une autre langue ou modalité. Il peut s'agir de versions corrigées de traductions universitaires, d'archives de bureaux de traductions, de traduction de films doublés, d'éditions bilingues dans la mesure d'ouverture des droits, de traductions d'articles de wikipedia, de transcription de collectages, etc.

La modalité de la donnée brute du corpus en fait un corpus **oral, écrit ou gestuel**. Un **corpus multimodal** est un corpus parallèle qui met en relation plusieurs modalités ; qui documente la gestualité du langage oral, ou qui transcrit une langue signée.

Un bon corpus de départ est produit par des locuteurs natifs dont le dialecte est spécifié. Il représente des styles d'expression différents (journalistique, littéraire, familier, narratif et d'échanges, etc.). En ce qui concerne l'écrit, on peut commencer par un premier écrit de 10 000 mots, et en préparer le traitement en désambiguïsant les points (ceux qui marquent la fin d'une phrase vs. ceux d'un acronyme par ex.), et les espaces (y a-t-il des mots qui comprennent un espace ou l'espace est-il toujours une frontière de mots ?). Pour améliorer les performances de plusieurs outils de traitement de la phrase, on peut estimer avoir besoin d'un bon corpus de taille assez importante (un million de mots). En ce qui concerne l'oral, la première ressource peut être un enregistrement brut, monolingue, dont les sons parasites éventuels sont nettoyés, sans chevauchements de la parole, et dont le format d'enregistrement

est standard.[2]

Les **corpus richement annotés** peuvent servir à une amélioration équivalente tout en étant de taille beaucoup plus modeste (mille phrases). Ce sont les corpus qui représentent le plus de valeur ajoutée dans leur mise en forme. Ils sont enrichis d'annotations sous un format largement établie à l'international comme le format CoNLL (*Conference on Natural Language Learning*), avec des critères eux-aussi largement établis à l'international (comme l'annotation *Universal Dependencies* (UD, cf. De Marneffe & al. 2021). Ces formats comprennent des gloses (des traductions mot-à-mot, en plus de la traduction globale), des informations morphologiques et catégorielles, et jusqu'à des informations grammaticales et sémantiques sur la structure de la phrase, les relations des éléments de la phrase entre eux.

Osborne & Gerdes (2019) fournissent une synthèse claire et critique des critères UD, dont on peut discuter avec eux de la nécessaire évolution, mais l'élément crucial est que si vous choisissez d'autres modèles d'annotation que UD, leur traduction en UD doit pouvoir être réalisé, ce qui est techniquement réalisable avec le formalisme SUD (Gerdes & al. 2019) qui est par ailleurs plus confortable pour les linguistes habitués aux structures en constituants. SUD, contrairement à UD, reconnaît les têtes fonctionnelles comme régissant le syntagme qu'elles dominent. L'outil *Grewmatch* développé par Bruno Guillaume (Sémagramme, LORIA/INRIA) prend en charge l'automation de la traduction d'une annotation en une autre (SUD > UD et UD > SUD). Dans l'État français, le réseau *AUTOGRAMM* qui les rassemble, financé par une ANR et dirigé par Sylvain Kahane (Modyco, Paris Nanterre), est spécifiquement orienté sur la constitution de corpus richement annotés de diverses langues à corpus restreint.

Ces corpus les plus enrichis d'informations sont particulièrement utiles pour la construction d'un analyseur syntaxique, et pour l'apprentissage automatique dit « supervisé ». Contrairement aux corpus bruts ou juste traduits, la construction de ces corpus richement annotés demande une petite expertise grammaticale et surtout un un temps considérable de codage. Cependant, les talistes ont développé des nouveaux **outils facilitateurs** qui réduisent la charge de travail que représente l'annotation des corpus et fournissent des environnements de travail relativement accessibles sans formation informatique (cf. *Arboratorgrew*, Guibon & al. 2020). Le coût d'entrée, les efforts nécessaires à des non-informaticiens pour maîtriser ces outils, s'il n'est pas négligeable, reste raisonnable. L'équipe de *AUTOGRAMM* améliore régulièrement ces outils, et est spécialisée sur la création de corpus UD/SUD pour les langues à corpus restreint.

Finalement, il est important de noter que les applications pédagogiques indispensables à une communauté linguistique pour ses apprenant·e·s et son cursus universitaire, comme les

2 Pour des recommandations pour la constitution de corpus oraux en particulier, se reporter à Baude et al. (2006), ainsi qu'aux modalités de leur dépôt sur la plate-forme *Cocoon* de Huma-Num.

dictionnaires et grammaires de la langue, peuvent être pensés dès leur conception comme des bases de données de corpus parallèles enrichis. Selon leur conception et les droits qui leurs sont attachés, ils peuvent fournir automatiquement les trois quarts des données nécessaires à l'établissement d'un corpus *Universal Dependencies*. Le réseau *AUTOGRAMM* travaille ainsi depuis le printemps 2022 à extraire les annotations d'une wikigrammaire des dialectes du breton en un Corpus UD/SUD (sur l'écriture d'une wikigrammaire à ces fins, voir Jouitteau & Bideault 2023, Jouitteau 2013) .

3. Ce que change l'Intelligence Artificielle

ChatGPT (*Generative Pre-trained Transformer* «transformateur pré-entrainé génératif») est une forme très connue de chatbot basé sur l'IA. Des langues minorisées apparaissent dans ses sets d'entraînement. Dès sa version 3, ChatGPT est capable de générer du texte qu'un humain peut reconnaître comme du breton. Du breton tout-à-fait agrammatical, étonnant, bancal, mais de façon reconnaissable, du breton. Ce fait découle de l'étendue de ce qui a été disponible en breton dans la base de données de ChatGPT datée de 2021, c'est-à-dire le contenu rédigé en breton sur internet à cette date, qui est de taille assez respectable en regard de la taille de communauté linguistique estimée à moins de 200.000 locuteurs, de moyenne d'âge assez élevée. Le logiciel ChatGPT3 a construit un modèle statistique de quel mot vient généralement après un autre en breton à partir des données disponibles sur internet lors de la constitution de sa base de données. Pour mesurer la masse de données disponibles sur internet pour une langue donnée, on peut utiliser un logiciel « renifleur » comme OSCAR (Ortiz Suárez et al. 2020). Ce programme délivre régulièrement des packs de données monolingues massives sur une langue donnée, telles que collectées en ligne. La qualité de ces données dépend évidemment de la qualité du robot reconnaisseur de langue, l'« épuisette » chargée de reconnaître la langue en question et de l'identifier parmi des masses de données diverses. Pour le breton, la qualité du reconnaisseur de langue est moyenne et la base de données que OSCAR remonte surtout du breton, mais contient aussi sporadiquement des phrases de français, thaï ou portugais.

3.1. Une conception émergente de la syntaxe

ChatGPT début 2023 ne « parle » pas de langue, ni n'en « comprend » aucune dans le sens humain de ces termes. Le terme « intelligence artificielle » ne devrait pas induire l'idée que nous rencontrons là l'intelligence d'un être nouveau. Ce que produisent les robots conversationnels est notre reflet. Plus précisément, un calcul statistique sur les données massives que nous avons disponibles sous format numérique. Prêter à ces modèles des intentions, bienveillantes ou malveillantes, revient à rater le test du miroir que nous considérons par ailleurs comme un test d'intelligence à travers les espèces. C'est bien notre image, là, dans le reflet déformé. Cela ne veut pas dire que la technologie sous-jacente est impuissante à saisir le langage humain.

ChatGPT 3,5 a accès à un modèle du sens individuel de chaque mot. Ce sens est construit à partir des autres mots qui lui sont statistiquement associés. Pour lui, « comprendre »

un mot signifie fondamentalement savoir avec quel autre mot il est associé statistiquement. Pour les humains, le sens d'une phrase découle d'abord de sa structure syntaxique, qui met en relation les sens particuliers des mots de la phrase. Quand ils reçoivent la suite linéaire des mots de leur langue, les humains interprètent une structure qui hiérarchise et ordonne les relations entre ces mots, crée du sens en organisant leurs relations entre eux. La sémantique de la phrase est ensuite un calcul sur cette structure, et non pas sur des mots individuels ou leur alignement. C'est la syntaxe qui nous permet de comprendre la différence radicale de sens entre la phrase *Le chat mange la souris* et la phrase *La souris mange le chat*. C'est la syntaxe qui nous permet de comprendre le sens identique de *La souris mange le chat* et *Le chat est mangé par la souris*. C'est encore la syntaxe qui nous permet de calculer sur quoi porte une négation (*Elle n'a pas dit qu'elle viendrait* vs. *Elle a dit qu'elle ne viendrait pas*), ou encore une question (*Qui tu dis qu'elle a pensé que la caméra a filmé* ?). La révolution dont ChatGPT est le signe se situe à cet endroit, dans la capacité émergente à appréhender la structure syntaxique, car les modèles semblent avoir des représentations hiérarchiques (Manning et al. 2020).

Trois avancées majeures ont révolutionné l'avenir numérique de toutes les langues, y compris les langues à corpus restreint. Voyons-les une par une.

3.2. Une conception émergente de la syntaxe

La grande avancée de l'intelligence artificielle qu'a illustrée ChatGPT vient de ce qu'on appelle les *transformateurs*. Ils viennent de recherches menées sur la traduction automatique (pour une bonne approche synthétique, voir Huang 2023, et plus technique, Worlfram 2023).

Les anciens modèles, *Recurrent Neural Networks* (RNNs) et *Long Short-Term Memory* (LSTM) analysaient les phrases de manière strictement linéaire, à partir d›une représentation vectorielle de mots qui prenait des mots en séquence et rendait des mots les uns après les autres. Traiter une séquence d'entrée de mots signifiait les introduire dans le modèle dans l'ordre. L'ordre des mots d'une phrase était donc intégré de manière uniquement implicite, et la phrase ne pouvait être appréhendée en son ensemble. La syntaxe nous glissait donc entre les doigts comme un poisson dans l'eau, mais deux innovations majeures ont sorti les modèles de l'ornière.

Les transformateurs ont d'abord commencé à donner une forme mathématique à l'ordre des mots. Ils comprennent maintenant explicitement un **encodage positionnel**. L'ordre des mots devient un objet mathématique. En pratique, à chaque mot d'une phrase est associé à un chiffre (*Je suis contente*. = (*Je* 1) (*suis* 2) (*contente* 3) (. 4), ou pour sa traduction en breton *Laouen on-me* = (*Laouen* 1) (*on* 2) (*-me* 3) (. 4)). Il devient donc possible de calculer sur ces ordonnancements en passant d'une langue à une autre. L'opération de traduction devient exprimable comme une opération de calcul (pour cette phrase, français 1, 2, 3, 4 = breton 3, 2, 1, 4). Il devient donc possible de faire des calculs sur l'ordre des mots dans une phrase, c'est-à-dire sur sa syntaxe. Les linguistes voient ici que le concept de matrice de phrase émerge… et que le modèle ignore encore le reste de la structure syntaxique. Mais la seconde innovation majeure y pallie en partie, en aidant l'algorithme à appréhender l'intégralité de la phrase, et

peser la pertinence de ses sous-parties, en considérant en avant et en arrière de la linéarité du message les mots qui l'entourent immédiatement. Elle est appelée l'**attention à plusieurs têtes** (*multi-headed attention*). Le modèle est conçu pour pouvoir regarder l'intégralité de la séquence d'entrée, en avant ou en arrière (> *attention*) différentes parties de la séquence (> à *plusieurs têtes*), afin d'évaluer une pertinence pour l'output. Ces deux innovations techniques pour la traduction, alliées à la puissance de calcul des algorithmes développés dans le domaine graphique et utilisés eux depuis une décade (GPU, Graphics Processing Units), ont réalisé un bond qualitatif dans les résultats. La partie de la structure syntaxique ici saisie est encore en débat (voir Gauthier et al. 2020 pour une batterie de tests sur les clivées, l'enchâssement, les cataphores, les items de polarité négative, les dépendances à lacunes, la subordination, l'accord, etc.), mais nous n'avons jamais été aussi près de la voir émerger dans les modèles mathématiques.

Ce sont ces deux innovations techniques pour la traduction, alliées à la nouvelle puissance de calcul des algorithmes développés dans le domaine graphique et utilisés eux depuis une décade (GPU, *Graphics Processing Units*), ont permis le bond qualitatif qu'on observe dans les résultats. Mais cela ne constitue que la première de trois révolutions.

3.3. Une alimentation généraliste

La seconde révolution concerne l'alimentation des modèles, qui est devenue plus autonome, et généraliste. Plus question d'étiqueter extensivement à la main des données d'entraînement qui restent forcément de taille modeste. Il y a dix ans, il fallait dire aux programmes de langage que tel mot est un nom, ce que c'est qu'un nom, et l'instruire en plus du sens de ce nom en le codant d'une manière que ce programme puisse manipuler. Un travail de titan ! Les transformateurs se nourrissent maintenant aisément de corpus parallèle (texte source et sa traduction), ou même de corpus bruts, monolingues. Ils commencent donc à pouvoir apprendre à partir de matière simple, non-transformée. On sait même comment les faire s'entraîner tous seuls à déterminer les catégories grammaticales des mots. L'algorithme peut s'entraîner sans humain, en se divisant en deux. Le premier soumet un texte avec des trous aléatoires au second, et le corrige. Ces exercices à trous sont fondamentalement des tests de commutation qui permettent au programme de faire correspondre par exemple les noms dans une certaine position et les sujets en général. C'est ce qui donne, si appliqué sur un corpus assez large, une bonne apparence à ses phrases. Il ne reste plus alors plus aux développeurs qu'à leur fournir du corpus – beaucoup de corpus brut, au kilomètre. Idéalement, le contenu d'internet.

ChatGPT est bluffant pour sa capacité à produire tout seul des phrases, ou à raconter la fin d'une histoire. Cette propriété découle tout simplement de ces exercices à trous. Quand l'algorithme est devenu efficace à remplir des petits trous… on élargit les trous. A terme, le programme arrivera à « restituer » une phrase qui n'était ... qu'un grand trou. En « restituant » un texte qui n'a jamais été, le programme produit du texte nouveau selon les indices de contexte qu'on lui a laissé. Dans cet exercice de création de phrases originales et nouvelles, l'algorithme s'appuie sur un modèle statistique lourd pour essayer de « restituer » du contenu

qui n'a jamais été présent. Les intelligences artificielles qui créent des images fonctionnent de façon similaire - en dégradant une image au lieu de trouer un texte. C'est très puissant car nous avons vu qu'un transformateur est fondamentalement un outil développé pour la traduction. Cette traduction est aussi applicable dans une même langue d'un style à l'autre, en changeant le niveau de langue, le style littéraire, la densité du texte, etc. Ceci obtient un potentiel génératif réalisé dans l'outil de conversation GPT. Qui peut ainsi résumer un texte, le réécrire « à la manière de », ou le mettre en poème.

3.4. Propriétés émergentes

Voyons pour finir la plus étonnante de ces révolutions. La capacité générative de phrases est en effet un type de résultat nouveau, mais il découle directement de l'exercice pour lequel ce programme a été écrit. Or, depuis le début des années 20, les modèles montrent des **propriétés nouvelles qui n'ont pas été codées par des humains.** GPT2, avec 1,5 milliard de paramètres, date de 2019. Il pouvait générer, s'il recevait des commandes bien précises, des textes courts étonnamment réalistes et dotés d'une relative cohérence interne (je laisse ici de côté la véracité ou non des contenus). GPT3 est 100 fois plus grand, avec 175 milliards de paramètres. Il date de 2020 et marque une avancée nette. Il peut écrire dans des langues à corpus large comme l'anglais ou le français des essais cohérents (encore une fois, de façon interne) presque impossibles à distinguer de l'écriture humaine. Ce qui est nouveau est que cette version 3 suit des instructions humaines simples sans que le modèle ait été conçu explicitement pour cela. Cet outil linguistique dispense donc ses utilisateurs d'apprendre des langages informatiques, ce qui est très pratique et décuple son utilisabilité. Mais le plus important est que cette capacité a émergé seule, sans être elle-même codée spécifiquement. C'est la preuve par la pratique que les larges modèles de langues apprennent des comportements entièrement nouveaux sur simples présentation de nouvelles données d'entraînement. C'est un saut qualitatif dans le développement : plus de la même chose a obtenu une chose qualitativement différente.

Personne ne peut actuellement prédire ce que cette révolution naissante amènera spécifiquement dans les usages et nos métiers, mais il est d'ores et déjà évident que nous considérons ici une mutation rapide, large et profonde pour tous les usages linguistiques qui transitent par des langues à large corpus numérique. L'enjeu est de savoir comment rattraper ce retard pour des langues qui n'ont pas à ce jour de corpus numérisés aussi massifs, enjeu vital pour ces langues. La technologie ouvre des perspectives nouvelles. Si une communauté linguistique arrive au stade où une intelligence artificielle peut elle-même générer du corpus grammatical dans sa langue, la technologie peut maintenant artificiellement augmenter le socle de son propre développement.

4. Les pièges à éviter à l'échelle de la communauté

4.1. Ne voir que les applications

On ne fait pas pousser une plante en lui tirant dessus. Ce que réclament les communautés, ce sont les applications, mais se concentrer d'abord sur elles ralentit le

développement. Les corpus et les outils sur lesquels les applications sont construites doivent être priorisés, partagés, et rendus largement accessibles pour nourrir le développement global. Si la communauté peut financer un programme de recherche sur la synthèse de la voix, on veut avant tout que le corpus créé à cette fin soit distribuable, et serve aussi au développement de la reconnaissance vocale. Le bond de l'IA montre que les applications peuvent changer de nature rapidement. Les corpus nécessaires, eux, restent. Leur besoin ne fait qu'augmenter. Dans la chaîne de développement, plus on s'avance vers les applications, plus on doit prioriser les outils qui nourrissent les corpus en retour, les traitent, les augmentent, les organisent (reconnaissance de caractère, outils de transcription, etc.).

Dès Alegria & al. (2011), la communauté basque a prévenu les financeurs que l'industrie du TAL ne se comporte pas comme avec les grandes langues, et que l'aide au développement devait absolument suivre l'ordre ressources > outils > applications. Il est contre-productif de mettre en concurrence pour financement des développeurs de produits finis (applications) avec des développeurs des données servant au développement (corpus, outils informatiques).

Une façon originale de ne voir que les applications découle d'un souci compréhensible d'adaptation aux usages sociaux réels, et s'appuie sur les travaux de sociologie du langage qui étudient les représentations qu'ont les locuteurs des applications déjà disponibles (repérabilité, intégration et intégrabilité du numérique dans les usages). Par définition, la sociologie des usages linguistiques travaille sur les réceptions des applications existant pour le public, et non sur les ressources et outils du développement, largement inconnus du public. C'est nettement contre-productif pour établir un plan de développement des ressources fondamentales nécessaires aux développeurs informatiques. La sociologie reste utile, à l'intérieur d'un plan informé des ressources, pour définir des échelles d'urgence dans le développement des applications (par ex., un âge élevé des locuteurs peut suggérer des applications d'accès aux soins médicaux, mais si le personnel médical sur le terrain refuse dans les faits de traiter les locuteurs s'exprimant dans d'autres langues que le français, le gain sera pauvre).

4.2. Quand le mieux est l'ennemi du bien

Les experts de langues minorisées et les chercheurs savent bien l'importance pour l'apprentissage des langues d'écarter soigneusement certains corpus (textes avec des fautes d'apprenants, formes sur-savantes inutilisées, formes de code-switching ; passages d'une langue à une autre, formes sur-dialectalisées pour un apprentissage du standard, formes sur-standardisées pour des échanges ultra-locaux, etc.). Force est cependant de reconnaître que l'apprentissage le plus efficace, l'acquisition du langage par les enfants en immersion, s'accommode parfaitement de « fautes » dans les données. Loin de corriger les enfants, nous les félicitons même parfois, ou répétons nous-même les formes fausses. Un enfant qui dit « Veux tato ! » peut être corrigé par « Je veux un gâteau. », mais il n'est pas rare non plus qu'il rencontre des réactions adultes enthousiastes, voire reprenant la faute « C'est bien mamour ! Tiens, ton tato au tocolat ! », « On dit un tato, s'il te plaît ! », « C'est ton troisième tato, les tatos ça va bien ! ». L'enfant apprendra tout de même *gâteau* si l'input correct est par ailleurs

assez massif. C'est vrai pour un apprenant humain, et encore bien plus pour un apprenant algorithmique qui n'est pas limité par sa mémoire et est parfaitement adapté à traiter des corpus massifs. Plus la masse de données disponible sera grande, mieux les algorithmes sauront trier le « bruit » (fautes, formes typiques des apprenants, scories, typos, données intraitables, fautes nombreuses mais différentes d'un corpus à l'autre, etc.). La sur-correction peut donc être une perte brute de temps, alors que l'auto-correction vient de la massification du corpus indépendamment des fautes sporadiques.

La correction est aussi nettement contre-productive dans les cas où on nécessitera des corpus de « fautes ». Les applications de reconnaissance vocale doivent absolument pouvoir être « éduquées » dans la reconnaissance de toutes les variétés de langue, ce qui inclut les formes « fautives », et les prononciations d'apprenants et de non-natifs de la langue. Les outils pourraient autrement discriminer des populations entières pour leur accès à des services essentiels. Cela ne signifie pas que les formes fautives vont être irrémédiablement mélangées aux autres types de corpus, baissant globalement leur qualité. Il faut associer à chaque corpus des métadonnées claires pour pouvoir les associer ou les dissocier à un set d'entraînement donné.

Le dernier danger, répandu, est celui du corpus parfait. Un corpus parfait est le pire des corpus dans la mesure où il ne sera jamais partagé, et ne servira à rien ni personne. Tout corpus raisonnablement bon peut être augmenté et enrichi. Toute réalisation numérique en droits ouverts est utilisable et améliorable. Un corpus doit être partagé dès qu'on l'estime utilisable, même à la marge, et améliorable.

4.3. Repérabilité

Les développeurs TAL ne font pas nécessairement partie de la communauté linguistique. C'est tout-à-fait manifeste dans le cas du breton, où le développement TAL est porté par des développeurs universitaires, dont les acteurs clefs peuvent ne pas parler la langue en question[3]. Si ces développeurs ne trouvent pas sur internet les ressources de corpus établies pour la langue, ces ressources ne servent à rien. Leur repérabilité automatique, à l'échelle internationale, est donc essentielle. Celle-ci permet aussi que les pouvoirs publics ou autres structures de politiques linguistiques puissent intervenir sur des données repérées par contractualisation, curation, visibilisation, etc.

Les quelques lettres qui servent à identifier une langue comme *fr* pour le français, *de* pour l'allemand, ou *eng* pour l'anglais sont appelées un code ISO (639.3) assigné par SIL

3 Le premier corpus *Universal Dependencies* du breton a par exemple été développé par Fran Tyers (U. Bloomington, Indiana, États-Unis) et Vinit Ravinshankar (U. Oslo, Norvège), aucun n'étant locuteur du breton ni du français. Les traductions ont été corrigées à posteriori, cinq ans après (Tyers & Ravishankar 2018, Jouitteau 2023).

International. Il est nécessaire de mentionner et faire mentionner le code ISO sur les corpus existants. Il est aussi possible de regrouper les corpus sous un nom de domaine associé à ce code ISO (639.3). Des sites web marchands peuvent en effet constituer *de facto* des corpus unilingues ou multilingues. La demande de création de nom de domaine numérique associé à une langue se fait auprès de ICANN, organisme régulateur d'Internet. Les outils de politique linguistique peuvent ensuite faciliter, créer l'hébergement de corpus amples dans la langue sous ce nom de domaine. La repérabilité peut aussi être assurée en développant très tôt parmi les outils informatiques un bon reconnaisseur de langue.

Enfin, les corpus accessibles doivent être repérables par mot clef automatiquement sur internet, ou visible dans un entrepôt ouvert d'archivage numérique (HAL, *Human-num* pour l'écrit, ou *Cocoon* pour les corpus oraux). Assurez-vous que les données et métadonnées sont récupérables en accès libre, en utilisant un protocole standard. Mettez les ressources disponibles sur des hébergements pérennes. Offrez aux petites réalisations associatives indépendantes, les sites de collectage et de mise en lien des locuteurs, la protection pérenne des données de leurs réalisations.

4.4. Ouverture des droits et science ouverte

Il est essentiel de normaliser les pratiques de partage et d'ouverture à l'échelle de la communauté linguistique, qui permettront d'exploiter les corpus existants de manière à ce que le travail effectué ne soit pas indéfiniment à refaire par différents acteurs qui s'ignorent. Cette normalisation des ressources ouvertes est un but commun à atteindre. Elle doit se faire dans le respect de ce qui a été construit auparavant, en augmentant au mieux les compatibilités de partage lorsque le partage ouvert total n'est pas réalisable.

Le glanage automatique à but scientifique est légal en France depuis l'ordonnance du 24 novembre 2021, qui transpose la directive européenne 2019/790 sur le droit d'auteur et les droits voisins dans le marché unique numérique. Cette ordonnance introduit dans le droit français une exception aux règles du droit d'auteur, applicable à l'exploration et fouille automatisée de textes et de données (Text and Data Mining). « *On entend par fouille de textes et de données […] la mise en œuvre d'une technique d'analyse automatisée de textes et données sous forme numérique afin d'en dégager des informations, notamment des constantes, des tendances et des corrélations* ». Cette ouverture en droit rencontre parfaitement l'exploitabilité nouvelle des textes bruts, au kilomètre, et profite directement aux langues à corpus massif. Pour le développement des langues à corpus restreints, c'est certes une avancée dans la mesure où le milieu de la recherche peut accéder aux corpus en ligne, et les traiter pour leurs propres buts, individuels ou d'équipe, mais le passage au développement collectif reste entravé. Ces langues doivent compenser la masse restreinte de leurs données par un enrichissement plus extensif de corpus clefs de départ. Les corpus récupérés en ligne et traités pour leur utilisation informatique ne pourront pas à leur tour être distribués en ligne. Quelle que soit la valeur de la notation ajoutée, le corpus initial reste toujours sous copyright propriétaire, non-redistribuable. Les chercheurs ne peuvent pas distribuer ces corpus dont ils n'ont pas les droits, et ne peuvent pas ainsi valoriser ce qu'ils ajoutent à ce corpus, économiser ce travail

pour d'autres développeurs. A l'échelle du développement pour une communauté, le verrou essentiel reste donc le copyright propriétaire. Il est essentiel d'avoir accès à des corpus en droits ouverts, utilisables, enrichissables et librement redistribuables sous cette forme.

En pratique, cela demande une modification des pratiques culturelles pratiquées au XX°, y compris dans la recherche. Le développement TAL est profondément inadapté à la culture du corpus douanier, dont la valorisation était organisée autour de la capacité de son ayant droit à y restreindre accès. Les acteurs clef du développement TAL viennent massivement des cultures du logiciel libre. Le développement TAL est adapté, en conséquence, aux pratiques de science ouverte que ces cultures ont fondées (principes FAIR pour *Findable, Accessible, Interoperable, Reusable*). Ces principes s'appliquent aux données (tout objet digital, dont logiciel), aux métadonnées (informations sur ces objets digitaux), et aux infrastructures, en recherche appliquée comme fondamentale. Ce sont des principes d'économie en ce qu'ils assurent un écosystème collaboratif où des acteurs libres, potentiellement isolés, co-construisent une efficacité durable. Dans cet écosystème collaboratif, les acteurs qui cherchent à restreindre l'accès à des ressources existantes sont identifiés comme peu fiables et contre-productifs. Ils sont évités, et par conséquence les langues en question, par les développeurs.

Les infrastructures de politiques linguistiques doivent sensibiliser leur partenaires (plateformes pédagogiques, recueils de collectages, médias écrits et radios, etc.) à la nécessité de la mention des droits associés aux matériels disponibles dans la langue. Les corpus de son et de texte doivent être associés visiblement à une licence *Creative Commons* stipulant leurs droits attachés.[4] Les communautés parlantes doivent apprendre à mettre en valeur leurs corpus libres et leurs outils informatiques disponibles. Cette mise en valeur inclut de citer rigoureusement les travaux universitaires des développeurs et développeuses de corpus et d'outils informatiques, pour ne pas pénaliser professionnellement les personnes très qualifiées qui aident à développer nos langues, parfois sans les parler eux-mêmes. Pour développer ces pratiques, de nombreuses ressources sont disponibles en ligne, recensées sur la plate-forme DORANUM et dans Hadrossek et al. (2023).

Une communauté qui sait valoriser ses corpus en s'assurant qu'ils peuvent être augmentés et redistribués, est une communauté qui attire les développeurs.

4 En quelques clics sur la plate-forme de *Creative Commons*, vous pouvez choisir quelle licence est adaptée à votre corpus, L'essentiel étant ici qu'ils soit redistribuable en ligne sous une forme enrichie pour le traitement TAL. La licence la plus recommandée, dans la mesure où elle est possible pour les auteurs et leurs structures, est le CC-BY simple. C'est la licence qui assure l'accès le plus large. A noter que la mention des auteurs de la ressource par leurs utilisateurs reste dans tous les cas une obligation sans exceptions.

5. Conclusion

Les locuteurs des langues à corpus massifs ont accès à des applications difficiles à produire pour les langues à corpus restreint, et la révolution de l'IA accentue rapidement, durablement et profondément cette césure. L'exemple simple de la reconnaissance vocale des SMSs montre l'ampleur de l'enjeu. En breton, en l'absence de cette technologie dans les applications de téléphonie, les brittophones habitués à échanger entre eux en breton passent massivement au français chaque fois qu'ils peuvent échanger en SMSs dictables. Circonstance aggravante, cela touche en premier les plus jeunes générations. Avec l'IA, cet effet de délaissement de la langue sera répliqué sur des myriades de nouvelles applications installées au cœur de nos échanges, de nos métiers. Travailler dans une grande langue sera nettement plus productif, et donc préféré. Jouer dans une grande langue sera nettement plus attractif, et donc préféré. S'informer dans une grande langue sera nettement plus vérifiable, et donc préféré. Si les langues à corpus numérique restreint ne sont pas intégrables aux pratiques de communication émergentes, leurs usages entre humains réels vont diminuer drastiquement pour les générations futures. L'enjeu, pour ces langues aux usages déjà minorisés, est donc littéralement de survivre au XXI° siècle.

Pour avancer le développement numérique des langues à corpus restreint, et en considération de la chaîne de développement du TAL, le maître-mot est le corpus, qu'il soit brut, traduit et annoté. L'enjeu de développement de ces corpus en libre accès, et particulièrement de corpus enrichis d'annotations linguistiques, est plus urgent que jamais. Dans l'État français, des initiatives non-négligeables existent pour accélérer le développement TAL des langues minorisées. En plus du réseau *AUTOGRAMM* cité plus haut financé à hauteur de 525.000 euros pour 2022-2025, le réseau *DIVITAL* dirigé par Delphine Bernhard (U. Strasbourg) et financé par une ANR à hauteur de 413.631 euros pour 2021-2025 concerne spécifiquement le développement de ressources TAL pour l'alsacien, le poitevin saintongeais, l'occitan et le corse[5]. Il mobilise des communautés universitaires en interface avec leurs communautés linguistiques (Bernhard et al. 2021), et fournit des outils, expériences et savoirs-faire techniques et sociaux. Il est à souligner qu'au sein de ces deux réseaux universitaires pour le développement TAL de langues à corpus restreint, la première visée est la constitution de corpus UD/SUD, considérant que la plupart des outils puis applications en découlent ensuite. Élargir les langues concernées à la petite centaine de langues à corpus restreint de l'État Français (Jouitteau 2010), ou tout du moins celles dont le développement numérique est amorcé, demanderait une volonté politique nette et un effort financier assez modéré au vu de l'enjeu sociétal et patrimonial.

5 Ces sommes ne comprennent pas les salaires des chercheurs, enseignants chercheurs et ingénieurs et secrétaires mobilisés par ces projets, ni les contrats de thèse de doctorat associés qui peuvent être soutenus, comme à Nanterre, par une université. Ces sommes restent cependant très modestes au vu des investissements colossaux concernant la préparation du français pour l'IA.

Dans cette visée stratégique informée du développement TAL en trois étapes, trois facteurs clefs se dégagent pour la survie digitale des langues à corpus restreint. En premier lieu, la vitalité du réseau universitaire et de recherche à même de délivrer des corpus enrichis assez vastes pour le développement d'outils généralistes pour le développement des outils d'IA. En second lieu, la progression des pratiques de science ouverte et l'attitude des communautés de locuteurs quant à l'accessibilité globale des données (prolixité, copyright, accessibilité et repérabilité en ligne). En dernier lieu, la capacité des structures de politique linguistique à comprendre rapidement les enjeux des deux premières et à les soutenir efficacement.

Bibliographie et webliographie

Alegria, Iñaki, Xabier Artola, Arantza Díaz de Ilarraza, et Kepa Sarasola. 2011. Strategies to develop Language Technologies for Less-Resourced Languages based on the case of Basque, Poznan.

Baude, Olivier, Claire Blanche-Benveniste, Marie-France Calas, Paul Cappeau, Pascal Cordereix, et al. 2006. *Corpus oraux, guide des bonnes pratiques 2006*. CNRS Editions, Presses Universitaires Orléans.

Bernhard, Delphine, Anne-Laure Ligozat, Myriam Bras, Fanny Martin, Marianne Vergez-Couret, et al... 2021. Collecting and annotating corpora for three under-resourced languages of France: Methodological issues, *Language Documentation & Conservation, University of Hawai'i Press* 15, 316-357.

Casenave, Jean. 2003. A l'occasion de la mise en ligne sur Internet du manuscrit Lazarraga: éléments de réflexion sur la conversion numérique des corpus littéraires du domaine basque., Centre de recherche sur la langue et les textes basques IKER, UMR 5478 CNRS, *Lapurdum* VIII, 97-121.

Ceberio, Berger et al. 2018. *Digital Language Survival Kit. The DLDP Recommendations to Improve Digital Vitality*, The Digital Language Diversity Project.

De Marneffe, Marie-Catherine, Christopher D. Manning, Joakim Nivre, Daniel Zeman. 2021. 'Universal Dependencies', Computational Linguistics 47:2, 255–308.

DORANUM, plate-forme sur la science ouverte.

Gauthier, Jon, Jennifer Hu, Ethan Wilcox, Peng Qian & Roger Levy. 2020. SyntaxGym: An online platform for targeted evaluation of language models. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 70–76.

Gerdes, Kim, Bruno Guillaume, Sylvain Kahane & Guy Perrier. 2019. Pourquoi se tourner vers le SUD : L'importance de choisir un schéma d'annotation en dépendance surface-

syntaxique›, *Actes des Journées scientifiques « Linguistique informatique, formelle et de terrain »*, Orléans, France.

Guibon, Gaël, Martine Courtin, Kim Gerdes, Bruno Guillaume. 2020. When Collaborative Treebank Curation Meets Graph Grammars, *Proceedings of The 12th Language Resources and Evaluation Conference*, Marseille, France, European Language Resources Association, 5293–5302.

Hadrossek, Christine, Joanna Janik, Maurice Libes, Violaine Louvet, Marie-Claude Quidoz, Alain Rivet et Geneviève Romier. 2023. *Guide de bonnes pratiques sur la gestion des données de la recherche*.

Huang, Haomiao. 2023. The generative AI revolution has begun—how did we get here? A new class of incredibly powerful AI models has made recent breakthroughs possible, Ars Technica.

Jouitteau, Mélanie. 2009-2023. *Arbres, wikigrammaire des dialectes du breton et centre de ressources pour son étude linguistique formelle*, IKER, CNRS, http://arbres.iker.cnrs.fr, Licence Creative Commons BY-NC-SA.

Jouitteau, Mélanie. 2010. *Entrelangues – Site et portail d'information sur les langues minoritaires parlées dans l'État français.*

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

Jouitteau, Mélanie. 2023. Traitement automatique des langues - Breton, *ARBRES, wikigrammaire des dialectes du breton et centre de ressources pour son étude linguistique formelle*, IKER, CNRS, Licence Creative Commons BY-NC-SA.

Jouitteau, Mélanie et Reun Bideault. sous presse. Outils numériques et traitement automatique du breton, Annie Rialland, Michela Russo & Catherine Schnedecker (éds.), Société de Linguistique de Paris.

Krauwer, S. 2003. The Basic Language Resource Kit (BLARK) as the First Milestone for the Language Resources Roadma, p*Proceedings of the International Workshop "Speech and Computer"*, SPECOM 2003, Moscow, Russia.

Manning, Christopher D, Kevin Clark, John Hewitt, Urvashi Khandelwal et Omer Levy. 2020. Emergent linguistic structure in artificial neural networks trained by self-supervision. *Proceedings of the National Academy of Sciences* 117 (48), 30046–30054.

Sayers, D., R. Sousa-Silva, S. Höhn et al. 2021. *The Dawn of the Human-Machine Era: A forecast of new and emerging language technologies*. Report for EU COST Action CA19102 *Language In The Human-Machine Era*.

Osborne, Timothy & Kim Gerdes. 2019. The status of function words in dependency grammar: A critique of Universal Dependencies (UD), *Glossa: a journal of general linguistics* 4:1.

Ortiz Suárez, P. J., L. Romary, et B. Sagot. 2020. A monolingual approach to contextualized word embeddings for mid-resource languages. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 1703–1714.

Tyers, Francis M., et Vinit Ravishankar. 2018. 'A prototype dependency treebank for Breton', *Actes de la conférence Traitement Automatique de la Langue Naturelle*, TALN 2018, 197-204.

Wolfram, Stephen. 2023. 'What Is ChatGPT Doing … and Why Does It Work?', Stephen Worlfram Writings.