L'élaboration d'une banque de données textuelles du breton

Yves Le Berre, Jean Le Dû

La Bretagne Linguistique

9 | 1993
Varia

L'élaboration d'une banque de données textuelles du breton

The development of a Breton textual database

Yves Le Berre et Jean Le Dû

OpenEdition Journals

Édition électronique
URL : https://journals.openedition.org/lbl/5879
ISSN : 2727-9383

Éditeur
Université de Bretagne Occidentale – UBO

Édition imprimée
Date de publication : 1 juin 1993
Pagination : 99-104
ISSN : 1270-2412

Référence électronique
Yves Le Berre et Jean Le Dû, « L'élaboration d'une banque de données textuelles du breton », *La Bretagne Linguistique* [En ligne], 9 | 1993, mis en ligne le 02 janvier 2022, consulté le 15 janvier 2024.
URL : http://journals.openedition.org/lbl/5879 ; DOI : https://doi.org/10.4000/lbl.5879

Yves LE BERRE
Jean LE DÛ*

L'élaboration d'une banque de données textuelles du breton

Problèmes de représentation

Les dictionnaires de toutes les grandes langues possédant une longue tradition lexicographique ont en général pour référent la langue écrite, particulièrement la langue littéraire. Si certains ouvrages décrivent un aspect ou un autre de la langue parlée, ils précisent lequel dans leur titre et se présentent de ce fait comme des compléments spécialisés aux autres dictionnaires.

Bien que le breton possède une telle tradition lexicographique remontant au milieu du quinzième siècle – le *Catholicon* manuscrit date de 1564 –, ses grands corpus de mots mêlent les registres écrit et oral et, au sein de ce dernier, les entrées dialectales et les divers niveaux de langue – néologisme compris – sans presque jamais indiquer leurs limites d'emploi. Conséquence pratique : les copies d'étudiants, mais aussi bien des textes publiés par les revues littéraires, contiennent d'étonnantes hétérogénéités lexicales.

Certains travaux universitaires décrivant des parlers locaux ou des champs sémantiques précis, comme les thèses de Pierre Trépos (*Le*

* LE BERRE Y. et LE DÛ J., "L'élaboration d'une banque de données textuelles du breton", in *La Bretagne Linguistique*, n° 9, 1993, p. 99-104.

vocabulaire breton de la ferme), Lan ar Berr (*Ichtyonymie bretonne*), Jean-Marie Plonéis (*Description du parler de Berrien*), Elmar Ternes (*Description du parler de l'île de Groix*), Pierre Denis (*Description du parler de Douarnenez*), Jean Le Dû (*Description du parler de Plougrescant*), Jean-Yves Plourin (*Description du parler de Saint-Servais et Langonnet*), Francis Favereau (*Description du parler de Poullaouen*), Gary German (*Description du parler de Saint-Yvi*), H. Ll. Humphreys (*Description du parler de Bothoa*), et d'autres comme les ouvrages de Jules Gros sur la syntaxe contiennent au moins un chapitre, parfois des volumes entiers, consacrés au lexique courant, mais leurs dimensions leur interdisent toutefois l'exhaustivité. Et symétriquement aucun dictionnaire de texte ne vient offrir au lexicographe les matériaux qui lui permettraient de préciser le registre, le niveau et l'emploi de tel mot... et parfois d'éviter de fâcheuses erreurs.

Dès l'achat par le CRBC des premiers ordinateurs personnels mis à la disposition des chercheurs, vers 1986, nous avions pensé entamer un tel travail, tout en sachant que la fin de notre carrière nous permettrait seulement d'en jeter les premières bases. Mais comme nous ne possédions pas le logiciel susceptible de faire le gros œuvre automatiquement (reconnaissance des formes, classement alphabétique, comptage, regroupement des occurrences), Yves Le Berre a dû se contenter à l'époque d'une première expérience manuelle sur un très court extrait de la *Vie de sainte Nonne*. L'étude de ces trois strophes a paru dans les *Cahiers de Bretagne Occidentale* N° 6 (hommage à Yves Le Gallo) sous le titre : "Le Pur et l'impur".

Dans les années suivantes, un collègue celtisant de l'université de Berkeley nous a signalé le logiciel *Kwic-Magic*, conçu en Californie pour l'étude, semble-t-il, des langues amérindiennes. L'ayant acquis, nous avons pu commencer un travail plus sérieux. Nous avons d'abord saisi nous-mêmes le texte des *Trois poèmes moyen-bretons*, de *l'Eloge funèbre de Michel Morin*, de la *Vie de sainte Nonne* ; puis des secrétaires nous ont relevés, saisissant tous les textes moyen-bretons dont nous disposions, plus quelques autres des époques moderne et contemporaine : *Contes* de Luzel, *Livr el labourer* de Joachim Guillôme. Ce premier corpus est actuellement en cours de correction.

L'acquisition toute récente d'un scanner accompagné d'un logiciel de reconnaissance de caractères va nous permettre de réduire considérablement le temps de saisie et de correction des textes, et de reporter à leur traitement l'essentiel de nos efforts.

***

Le traitement se fait sur des textes non normalisés, la graphie d'origine étant scrupuleusement respectée. Les erreurs, même manifestes, de l'original sont conservées à l'identique. Une première tentative de normalisation nous a en effet démontré qu'on ne savait jamais où s'arrêter dans un tel processus.

Ces textes sont tous minutieusement corrigés. Tout d'abord, avant le traitement informatique. Ensuite, grâce au classement automatique des items par ordre alphabétique les erreurs de saisie qui pouvaient subsister sont facilement repérées.

Les entrées ont été maintenues à leur place alphabétique, mais sont renvoyées systématiquement à la forme 'normale' (par exemple *stler* n'est pas possible en breton : il s'agit bien évidemment de *scler* 'clair').

Les formes sont regroupées de façon de plus en plus étroite : pluriel renvoyé au singulier, verbes conjugués au radical etc. Mais la machine ne peut ranger à notre place les pluriels de noms (surtout s'ils sont anomaux) sous les singuliers, ni les divers avatars d'un paradigme verbal sous son radical. De même est-elle incapable de déduire d'une liste de contextes les divers sens et emplois d'un mot. Elle ne saurait non plus reconstituer sans erreurs les consonnes originales des mots mutés : comment en effet faire le départ entre un g- apparaissant à l'initiale absolue et un g- issu de la mutation d'un k- ? Le travail de rédaction proprement dite du dictionnaire de texte est un travail long et ardu mais passionnant. Encore plus long en breton, où le système des mutations consonantiques initiales demande des renvois innombrables. Ainsi, pour plusieurs textes, les entrées commençant par v- sont-elles toutes renvoyées soit à m- soit à b-, voire à gu-. Un z- peut être la lénition d'un d-, la mutation spirante d'un t-, la néo-lénition d'un s- ou dans certains textes la forme cardinale non mutée.

Chaque item est cité dans tous les contextes dans lesquels il apparaît. Ce contexte est modulable : dans les poèmes, il nous a semblé que la citation du vers entier suffisait.

Les textes sont considérés comme des entités. La traduction ne se fait qu'au vu du contexte, ce qui donne des résultats surprenants, particulièrement en moyen-breton.

***

Kwic-Magic est vraiment *quick* et vraiment *magic* : il "lit" tout texte mis en mémoire dans l'ordinateur, quelles que soient ses dimensions (Le *Mirouer de la mort* contient 3 602 vers ; la *Vie des quatre fils Aymon* se compose de plus de 10 000 alexandrins !), numérote et indexe ses lignes, compte ses mots, classe selon l'ordre alphabétique toutes les occurrences de chaque forme lexicale différente et regroupe sous chacune d'entre elles toutes les séquences du texte (vers, phrase...) dans lesquelles cette forme est présente. Ces opérations ne pourraient être réalisées sur fiches-papier qu'au prix de centaines d'heures de travail ; Kwic-Magic les effectue en quelques minutes. Il sait d'ailleurs faire quantité d'autres choses, mais nous n'en sommes pas encore au stade où nous pourrions utiliser la totalité de ses capacités.

A ce moment, deux voies s'ouvrent devant nous :

d'une part nous possédons sous la forme de fichier informatique et d'une sortie sur papier un dictionnaire automatique de notre texte.

d'autre part nous possédons une liste d'items, c'est-à-dire de formes lexicales différant chacune de toutes les autres. Nous pouvons (automatiquement et quasi instantanément, toujours grâce à Kwic-Magic) les classer selon leur longueur, selon le nombre croissant ou décroissant de leurs occurrences, selon l'ordre alphabétique direct (qui regroupe tous les préfixes identiques) ou inverse (qui regroupe tous les suffixes et désinences identiques). Nous pouvons encore "marquer" chaque item d'un code indiquant sa nature ou sa fonction grammaticale, son origine étymologique (celtique, latin, roman, germanique etc.), son rang dans le syntagme ou la proposition. Mieux, nous pouvons encore "croiser" ces listes pour établir des corrélations d'identité ou d'opposition entre ces diverses informations. Une sorte de spectre lexical du texte ainsi disséqué apparaît alors, qui l'identifie relativement à tout autre texte ayant subi le même traitement.

***

Une image de la langue se dessine de la sorte à travers le prisme d'un texte. En confrontant les résultats obtenus à partir de textes écrits à différentes époques, dans différents lieux et par des auteurs placés dans différentes situations socioculturelles, des traits communs se dégagent. L'ensemble de ces traits peut être considéré comme constituant un portrait de la langue qui échappe dans une large mesure aux variations chronologiques, spatiales et sociales.

Nous avons présenté une communication au Congrès international d'études celtiques de Paris de 1991 intitulée 'Celtique, latin, roman : approche lexicale du Mirouer de la Mort'. Ce texte, paru en 1992 dans *Etudes Celtiques*, constitue le premier résultat tangible d'une entreprise de longue haleine portant sur plusieurs des textes contenus dans notre banque de données : la *Passion* (XV^{e} siècle ?), le *Mirouer de la Mort*, la *Vie de sainte Barbe* (XVI^{e} siècle), *l'Eloge funèbre de Michel Morin* (XVIII^{e} siècle) et le *Livr el labourer* (XIX^{e} siècle). Il permet de tirer quelques premiers enseignements positifs, bien qu'encore très généraux, sur la mécanique des mots du breton.

Quant aux statistiques, nous en arrivons à la conclusion suivante :

A. La moitié environ des mots d'un texte ne figure qu'une seule fois dans ce texte ; nous les appelons des *hapax* (du grec *hapax legomenon* signifiant 'Chose dite une fois'). L'autre moitié des entrées du dictionnaire est formée d'unités récurrentes, présentes de deux à plusieurs centaines de fois dans le texte ; nous les appelons simplement, faute de mieux, des *items*.

B. Quelques dizaines de mots reviennent si souvent qu'à eux seuls ils constituent la moitié de la matière du texte (c'est le groupe 1, celui des mots

très récurrents). L'autre moitié du texte est formée d'une part de mots moins récurrents (c'est le groupe 2), d'autre part de hapax (c'est le groupe 3).

C. Les unités du groupe 1 sont massivement des mots-outils, monosyllabiques, d'origine celtique. Les unités du groupe 3 sont majoritairement des mots référencés, polysyllabiques, d'origine non celtique (latin, roman, germanique...). Les unités du groupe 2 mêlent les caractères du groupe 1 et ceux du groupe 3.

D. La plupart des unités du groupe 1 se retrouvent d'un texte à l'autre ; nous les déclarons *athématiques* parce que leur présence fréquente n'est pas liée au thème traité dans tel ou tel texte, mais au fonctionnement général de la langue. Inversement, peu de hapax se retrouvent d'un texte à l'autre parce qu'ils sont appelés par le thème propre à un texte; ils sont donc réputés *thématiques*.

E. Un petit groupe de mots athématiques fortement récurrents n'apparaît que dans les textes antérieurs au XVII^{e} siècle : ce sont des mots-chevilles liés aux habitudes rhétoriques et aux nécessités de la versification en usage à l'époque du moyen-breton (*glan*, *dien*, *certen* etc.).

F. Les unités outils, brèves, simples, celtiques et fortement récurrentes des groupes 1 et 2 suggèrent l'existence d'un ensemble structuré, très ancien et très stable qui pourrait métaphoriquement figurer le moteur de la langue. Les unités référencées, plus longues, souvent affixées, bien moins souvent celtiques et faiblement récurrentes des groupes 2 et 3, apparaissent en contraste comme un courant produit par l'actualité extralinguistique et translinguistique, fluide instable et volatil que "consommerait" ledit moteur.

Ces premières conclusions encore toutes naïves permettront d'imaginer aisément, nous l'espérons, l'intérêt qu'un tel traitement exhaustif, partie automatique, partie manuel, de la matière lexicale de textes très divers pourra présenter dans un avenir proche. D'un côté pour l'établissement d'une grammaire de la langue qui ne soit élaborée ni à partir de modèles étrangers mal adaptés au breton ni à partir de formules théoriques souvent peu convaincantes dans la pratique, mais sur la réalité concrète de discours incontestables. D'un autre côté pour l'histoire du breton écrit dont le sous-développement chronique prolonge fâcheusement l'usage d'outils obsolètes et l'existence d'idées reçues du dix-neuvième siècle qui devraient être refondus (pour les premiers), abandonnées ou reformulées (pour les secondes).

***

Presque tous les textes moyen-bretons sont saisis, ainsi qu'un nombre non négligeable de textes modernes. Pour conclure, voici en quelques lignes une description des principaux travaux en cours, classés par ordre chronologique.

Breton du seizième siècle

Yves Le Berre termine le traitement des *Trois poèmes en moyen-breton* de divers points de vue : lexical, sémantique, grammatical. A partir de ce travail, il a établi une nouvelle traduction de ces poèmes.

Le *Mirouer de la Mort* : Yves Le Berre et J. Le Dû travaillent à l'établissement du dictionnaire complet de ce poème de 3 602 vers de 8, 10 et 12 pieds. Le texte compte 27 089 mots, répartis en 4 264 entrées. Les mutations ne sont en général pas notées. Les entrées et les exemples sont rassemblés et les traductions proposées pour les lettres de A à F.

Breton du bas-Léon du dix-huitième siècle

Eloge funèbre de Michel Morin par Claude-Marie Le Lae.

La recherche se fait au sein d'un groupe organisé dans le cadre du C2 de breton et comprenant des étudiants actuels et anciens de maîtrise et les enseignants concernés (Yves Le Berre et J. Le Dû). Il nous a semblé, en effet, que le travail en commun était formateur, à la condition d'être varié. Le groupe se réunit depuis maintenant six ans à raison d'une séance de 2 heures hebdomadaires (tous les vendredis après-midi).

Le texte est traduit en français (la traduction existante de Gaston Esnault, fourmillant de termes rares et dialectaux, n'était pas satisfaisante).

Son dictionnaire est maintenant établi. Il faudra cependant attendre l'aboutissement de la réflexion grammaticale pour préciser les définitions des entrées. La recherche porte actuellement sur le système verbal du breton du texte, chaque participant ayant en même temps la charge d'un point de grammaire particulier (genre et nombre, prépositions fléchies, etc.) qu'il présente au groupe pour être discuté en commun.

Vannetais du dix-neuvième siècle

Livr el Labourer, de Joachim Guillôme.

Le texte et sa traduction ainsi qu'une étude en ont été publiés par Y. Le Berre dans les *Cahiers du CRBC*. Le dictionnaire complet par J. Le Dû est pratiquement terminé.

Yves Le Berre, Jean Le Dû
Université de Bretagne Occidentale - Brest