Les atlas linguistiques sont-ils des corpus ?

Esther BAIWIR
Pascale RENDERS
FNRS / Université de Liège

Les atlas linguistiques, qui recueillent une part plus ou moins grande des lexèmes dialectaux d'une région donnée, peuvent-ils être considérés comme des corpus et exploités comme tels ? Cette interrogation survient inévitablement dès qu'il est question d'informatiser ces atlas afin de les rendre accessibles à la communauté scientifique, car les fonctionnalités offertes – et, donc, le traitement informatique de départ – doivent être pensées en fonction des objectifs de l'informatisation et de ce que l'on souhaite proposer à l'utilisateur. Un tel projet est à l'étude pour l'*Atlas linguistique de la Wallonie* (ALW) avec, dans un premier temps, l'intégration d'échantillons des publications dans des bases de données, exploitables *via* une interface de consultation en ligne. La question d'une exploitation comme corpus se révèle particulièrement délicate dans le cas de cet atlas, qui a la particularité de proposer une analyse des matériaux qu'il rassemble en les structurant au sein de notices. Pour rendre ces matériaux atteignables par des moteurs de recherche, il est nécessaire de les transformer en séries de listes sur lesquelles peuvent porter des requêtes, traitement qui a pour conséquence directe la perte des informations apportées par l'analyse. La question se pose, dès lors, de savoir si la réduction ainsi opérée est pertinente et si certaines de ces informations ne doivent pas être conservées. Par ailleurs, l'analyse a pu modifier le corpus initial d'une façon qui ne permet plus l'accès aux données brutes de départ. Après quelques considérations d'ordre général, nous donnerons deux exemples de cas qui posent question dans une optique de création d'un corpus atlantographique informatisé.

Corpus n°12 « Dialectologie : corpus, atlas, analyses » (2013), 27-37

1. L'atlas comme corpus

Il semble légitime de considérer qu'un atlas linguistique, qui recueille une part plus ou moins grande des lexèmes dialectaux d'une région donnée, puisse servir de corpus pour des études portant sur le dialecte considéré. Pourtant, catégoriser ce type de corpus atlantographique n'est pas évident. Il s'agit, certes, d'un corpus d'unités de langue et non d'un corpus de textes, mais ces unités sont atteintes via des réalisations concrètes qui sont de l'ordre du discours. Ces matériaux relèvent de l'oralité, mais ne peuvent être édités qu'à travers leur transcription. Le corpus n'est jamais exhaustif, mais il tend parfois à l'exhaustivité, selon l'ambition de l'enquête de départ. Il n'est pas nécessairement construit pour être représentatif, mais il l'est en pratique. Les données sont de caractère brut au départ, mais peuvent être présentées sous une forme traitée, selon le type d'atlas. La question de savoir si le corpus contitué par l'atlas est de l'ordre du donné ou du construit n'est pas simple : elle dépend également du type d'atlas.

La cause de ces problèmes de catégorisation provient en partie du fait que l'atlas lui-même est le résultat de l'exploitation et de l'analyse d'un corpus. La plupart des atlas linguistiques, à l'exception des atlas se basant sur des sources secondaires (tels que l'ALiR), sont établis à partir d'une enquête dialectale sur le terrain, destinée à recueillir un ensemble le plus complet possible de faits linguistiques propres à la région considérée. Cette enquête préalable est généralement constituée d'unités plus ou moins complexes, soit à traduire, soit à susciter chez le témoin. Quelle que soit la méthode d'enquête, l'ensemble des matériaux recueillis constitue un corpus proprement dit, qui peut être considéré comme représentatif en ce qu'il devrait permettre de se faire une idée assez complète et précise des particularités linguistiques propres aux parlers de la région. Evidemment, nul n'ambitionne de recueillir l'ensemble d'un parler et représentativité ne signifie pas prédictibilité ; les unités du lexique qui n'auront pas été intégrées à l'enquête ne pourront que rarement être déduites des autres unités appartenant au même champ sémantique ou à la même famille lexicale, eu égard à l'inscription de notre objet dans l'humain et dans l'histoire. Notre postulat

est pourtant que le corpus constitué par l'enquête offre bien un échantillon représentatif de faits linguistiques. Il s'agit d'un corpus brut, non encore remodelé par l'analyse (v. Dalbera 2002 : 2) ; il s'agit d'un corpus relevant de l'oralité, mais néanmoins soumis à une transcription écrite, ce qui peut poser, en pratique, certains problèmes sur lesquels nous reviendrons.

Les atlas, qui éditent et publient les matériaux récoltés par l'enquête, sont-il également susceptibles de servir de corpus ? La réponse est différente selon le type d'atlas pris en compte. Une distinction est généralement opérée entre les atlas proposant des matériaux d'enquête bruts, non interprétés, et ceux dits interprétatifs, dont l'ambition est de guider le lecteur en lui fournissant les clés de la matière. Les premiers proposent une reproduction exacte – sous une forme principalement cartographique, même si des franges de matériaux annexes sont souvent relégués en marge des cartes – des matériaux de l'enquête : ils peuvent de ce fait (et moyennant un encodage des formes) être considérés comme un corpus au même titre que cette dernière. Quand il y a eu soit une transcription, soit une homogénéisation des pratiques des enquêteurs, soit encore une sélection des matériaux (par exemple l'extraction d'un mot du contexte phrastique de l'enquête ou l'exclusion d'un individu peu sûr), on peut déjà considérer ce corpus comme légèrement différent du corpus initial, un filtrage s'étant opéré au sein de la masse des matériaux, mais le corpus est toujours de l'ordre du donné.

Le second type d'atlas constitue un cas plus complexe. Les matériaux publiés dans un atlas interprétatif tel que l'ALW ont en effet été enrichis par l'apport de diverses informations, de type phonétique, étymologique, historique, voire ethnographique (v. Boutier 2008 : 309-310). Les notices qui en résultent constituent des monographies onomasiologiques, rédigées, proposées à la lecture linéaire, qui ne constituent pas *stricto sensu* un corpus, mais une analyse du corpus constitué par les enquêtes initiales. Un atlas interprétatif se situe donc à un second niveau par rapport au corpus. Cependant, au-delà de l'aspect monographique, l'atlas est aussi perçu comme un *thesaurus*, rassemblant tous les lexèmes d'un ou de plusieurs dialectes. Cette perception comme *thesaurus* explique qu'il puisse être

considéré comme un corpus au même titre que l'enquête et ce malgré sa présentation monographique, peu apte à permettre cette fonction.

Dans le cadre d'un projet d'informatisation qui vise à mettre les données linguistiques wallonnes à la disposition des utilisateurs sous une forme plus accessible et, éventuellement, comme un corpus utilisable en tant que tel, on peut se demander si ce n'est pas la liste des mots recueillis par l'enquête qui devrait être informatisée, plutôt que les notices de l'ALW. Considérant (1) que l'atlas a pour objectif initial et premier d'éditer l'enquête et de la présenter au lecteur de façon accessible ; (2) qu'il n'y a aucune différence quantitative entre l'enquête et l'atlas, la totalité des unités lexicales de l'enquête étant reprises dans ce dernier ; (3) que le traitement effectué dans l'atlas ajoute à ce contenu lexical des informations sans en retrancher aucune, la réponse nous semble évidente. La publication des matériaux récoltés par l'enquête fait effectivement de l'ALW, malgré sa forme monographique, un véritable *thesaurus* des dialectes belgo-romans. Le chercheur qui refuserait à l'ALW le statut de corpus se priverait du seul matériau assuré pour ce champ d'étude.

Il apparaît donc pertinent de permettre à l'ALW d'être utilisé comme un corpus d'unités de langue, en gardant à l'esprit le fait que l'édition des matériaux bruts fournis par l'enquête y a été accompagnée d'un enrichissement et que les deux niveaux, celui des matériaux et celui de l'analyse, sont intimement mêlés dans la structure même des notices de l'atlas. Cette particularité rend nécessaire l'informatisation de l'atlas, cette informatisation étant le seul moyen efficace de donner accès à ce corpus en l'extrayant de la couche d'analyse.

2. L'accès au corpus atlantographique

Donner accès au corpus revient, en pratique, à extraire de l'atlas la liste des mots de l'enquête accompagnés de leur localisation (l'enquête associant à chaque forme le point d'enquête où elle a été recueillie). Cette extraction n'est pas triviale, la présentation monographique de l'atlas ayant pour conséquence de rendre implicite cette relation.

[1] Ainsi, à la notice BAVARDER (ALW 17, not. 113), après l'introduction et les divers renvois bibliographiques, l'un des paragraphes de matériaux se présente comme suit :
F. 1. **+calôder**...:[33] *kalődé* Mo 1, 42, 79; S 19, 31, 37; Ch 16, 26, 63, '64, 72 (arch.); [...].

L'opération consistant à extraire de l'ALW le corpus de matériaux revient à effectuer une réduction par rapport à l'enrichissement opéré par l'analyse. En particulier, l'introduction explicative située en tête de notice a été perdue, ainsi que la note 33 ou la classification (*F.*1.) intégrant cette forme dans une vision globale des verbes édités dans la notice. En outre, l'automate devra restituer la totalité de la localisation géographique, pour partie perdue dans la version papier ; si le premier point est bien Mo 1, le deuxième est Mo 42 (et non 42), puis Mo 79, etc.

Dans ce processus de réduction informatisée, il nous semble toutefois qu'une part de l'analyse gagnerait à être conservée. Nous en donnons ci-dessous deux exemples, à propos de la graphie des formes et à propos de leur étymologisation.

2.1 Transcription phonétique des formes

Les listes de formes dialectales sont, dans les atlas de faits, simplement transcrites dans des systèmes phonétiques. La transformation de ces atlas en corpus informatiques semble à la portée de tout informaticien, et pourtant. De telles entreprises, en France par exemple, ont rapidement été confrontées à diverses difficultés, parmi lesquelles la saisie des formes, eu égard à la multiplicité des signes diacritiques :

> [L]a disparité des notations phonétiques a amené à la création pour chacun des atlas de son propre alphabet phonétique avec codage particulier. Une comparaison attentive de plusieurs atlas a en effet révélé que chaque auteur a apporté pour son propre usage des modifications au système Rousselot, non seulement en fonction des particularités de son terrain d'enquête, mais aussi à partir de ses propres conceptions de la notation et donc, en dernière analyse, de sa personnalité. (Le Dû 1992 : 304)

Il n'en va guère autrement des enquêtes pour l'ALW. Celles-ci furent menées par de multiples enquêteurs, sur une longue période (entre 1924 et 1959). Chacun, comme au sein des atlas de France par régions, a forgé son petit système personnel. Il suffira pour s'en convaincre de comparer les documents ci-dessous, représentant la même réalisation phonétique :

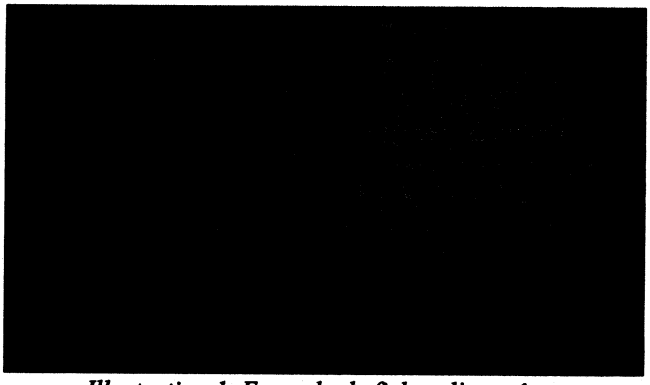

Illustration 1. Exemple de fiches d'enquête : question 1820 « pile ou face »

L'expression est traduite, dans les quatre cas, par une locution composée de type <pile ou ...>. Le premier élément se prononce de la même façon dans les quatre cas. Deux enquêtes (To 37 et To 99) conservent un *e*, uniquement graphique, issu de la connaissance que les enquêteurs ont du français ; ces enquêteurs utilisent un système graphique qui vogue entre précision phonétique et représentation plus intuitive par rapport à la graphie de la langue source qui, en l'occurrence, est aussi une langue sœur. En revanche, la partition est différente pour l'indication de brièveté du *-i-* : ce sont les enquêteurs de To 37 et de B 22 qui sont plus précis. Ce signe de brièveté n'est pourtant pas nécessaire ; par défaut, les *-i-* de l'enquête sont toujours brefs. Si aucune des deux divergences n'est problématique pour un lecteur rompu à l'exercice, il est intéressant de constater quatre graphies distinctes pour une seule réalisation phonétique. Au passage se voit encore, pour la traduction de la conjonction, l'opposition entre notation phonétique (un *u* spécifiquement en usage parmi les dialectologues wallons) et analogie avec le français (*ou*). L'accolade ajoutée à l'enquête de To 99, où l'en-

quêteur mentionne les deux versions, montre bien l'instabilité du système.

Comment, dès lors, envisager des requêtes informatiques transversales sur des listes aussi disparates ? N'est-ce pas conduire l'utilisateur dans un mur ? Diverses options se présentent pour l'informatisation. La première, qui est exploitée dans la rétroconversion du FEW (v. Renders 2011 : 137), est celle de longues listes de pseudo-équivalences permettant la neutralisation d'oppositions graphiques. La machine peut par exemple ignorer les diacritiques ou considérer « β » et « b » comme interchangeables. Cette solution permet une certaine souplesse d'utilisation ; même l'utilisateur maîtrisant mal les conventions phonétiques de tel ou tel enquêteur peut espérer une réponse à sa requête.

Dans un atlas interprétatif comme l'ALW, ces divergences ont pour particularité d'être neutralisées par les rédacteurs eux-mêmes. Les données de l'enquête sont en effet transcrites dans le système orthographique d'usage pour les dialectes belgo-romans : l'orthographe dite *Feller*, du nom de son inventeur (v. Feller 1905 ; Pierret 1992 ; il s'agit de la forme en caractère gras de l'exemple 1). Ce système graphique est basé sur un double principe : il s'appuie sur l'orthographe française, mais tente de noter précisément la prononciation. Les variantes phonétiques minimes sont neutralisées – par exemple, la différence de timbre entre API a, *ä* et *a*, de même que des différences graphiques représentant le même son, comme *é*, *éÚ* ou *ÈÚ*. Simplification, réduction d'un côté, mais enrichissement de l'autre : le système Feller intègre, comme l'orthographe du français, des signes muets donnant des informations morphologiques ou permettant de distinguer des homophones.

[2] Ainsi, *Nos-èstans firs d'èsse Lîdjwès* correspond à la notation phonétique suivante : *noz èstã fi:r d ɛs li:dzwɛ*. Le *-s* de *nos* permet l'identification du pronom personnel, mais également la liaison avec le verbe qui le suit. Le *-s* de celui-ci identifie le morphème *-ans* comme un morphème verbal de première personne du pluriel (comp. fr. *-ons*) ; le *-s* de *firs* est la marque du pluriel. Quant au *-s* de *Lîdjwès*, il permet de subodorer la forme de l'adjectif féminin (*Lîdjwèse*).

On le voit, le système Feller est en lui-même une interprétation. Il convient d'analyser l'énoncé avec justesse pour pouvoir le graphier. Ce traitement modifie le corpus de départ, mais il améliore la lisibilité des unités, tant au niveau visuel que morphologique. Prendre pour départ d'une informatisation une version en orthographe Feller des matériaux belgo-romans écarte le problème des habitudes individuelles des enquêteurs, mais aussi la question des signes diacritiques complexes des transcriptions phonétiques (en italique dans l'ALW ; v. exemple 1), inexistants en Feller. Remarquons que l'existence de cette orthographe normalisée n'interdit pas de mettre en place un système informatique de neutralisation des oppositions graphiques. Dans le cas d'une variété de tradition orale, le public n'est guère rompu à l'exercice d'écriture et il n'est pas raisonnable de demander à l'utilisateur d'une interface de consultation informatique d'introduire correctement les accents sur les voyelles, par exemple.

2.2 Etymologisation des formes

Au point de vue technique et dans nos disciplines, un corpus informatisé prend généralement la forme de documents textuels électroniques qui sont interrogeables par ordinateur. En fonction des interrogations prévues, les divers types d'information contenus dans ces documents textuels peuvent être balisés, c'est à dire annotés, de façon à permettre à un automate de les identifier et de les retrouver. Ce balisage permet également d'élaborer des index, qui optimisent la recherche d'informations dans le corpus en fonction des besoins de l'étude.

Une recherche de matériaux dans une base de données textuelles (le terme « base de données » étant ici à comprendre au sens large) nécessite donc que soient définis les types d'information pertinents pour l'analyse. Dans le cas de l'ALW, il s'agit par exemple de la liste des unités lexicales éditées dans l'atlas, ainsi que les points d'enquête (localisation géographique) où ont été relevées ces formes. Ces deux types d'information correspondent aux matériaux de l'enquête initiale ; leur annotation (ou balisage) permet d'accéder au thesaurus lexical qu'est l'atlas. Que penser, en revanche, des types d'information ajoutés par l'analyse aux matériaux de l'enquête, tels que les étymons, qui relèvent de l'analyse et du mode monographique de

l'atlas ? Ils ne font pas partie du corpus proprement dit, mais ils constituent une voie d'entrée permettant d'accéder aux matériaux selon un angle d'étude particulier ; à ce titre, ils sont évidemment d'un grand intérêt pour l'utilisateur. En outre, ils sont facilement balisables et constituent une facette précieuse des matériaux.

Le travail d'étymologisation des formes effectué par les rédacteurs de l'ALW peut donc être intégré dans l'ensemble des informations offertes par le corpus. L'apport de cette intégration est important, car il donne une visibilité plus grande au travail d'étymologisation. En effet, en l'état actuel, cette information est donnée au fil du texte et se découvre, pour le lecteur, uniquement dans une consultation de type monographique de l'ouvrage. L'outil que constituent les index étymologiques, présents dans les volumes les plus récents, permet de dépasser cette lecture, mais pas complètement ; il reste en effet cantonné aux seuls domaines sémantiques traités dans le volume consulté. Une informatisation intégrant cette analyse permet de porter un regard plus large sur la matière.

[3] Examinons par exemple l'étymon latin GAMBA, traité FEW 2, 111 sous CAMBA. Si l'on peut assez facilement soupçonner qu'une requête informatique pointera vers le sens 'jambe' (dont toutes les variantes phonétiques sont présentées ALW 1, not. et c. 52 JAMBE), on sera plus surpris de trouver un sens 'lobe ou cuisse de noix' (v. ALW 6, not. 80 AMANDE DU NOYAU, n. 3 et not. 123 BROU DE LA NOIX, β et n. 57), ayant parfois évolué en 'amande du noyau (d'un fruit)' (v. ALW 6, not. 80, *H*). Un automate permettra même au non initié de faire le lien avec le type <jambot>, signifiant parfois 'bébé' (ALW 17, not. 26 BEBE, *J* et n. 12), plus souvent 'enfant' (ALW 17, not. 34 GARÇON, *E*, not. 38 gamin, *D*, etc.) ou avec le collectif <jamboterie> 'marmaille' (ALW 17, not. 33, *D* et n. 7). Un autre type suffixé en <-illon> apparaît ALW 15, sous la forme *gambiyons* 'mal aux cuisses, d'avoir marché' (ALW 15, not. 53 «ENTREFESSON», n. 8). L'étymon est encore illustré par des formations verbales, telles que <jamb(i)er>, apparaissant sous plusieurs réalisations phoniques (ALW 17, not. 99 je les ai fait DÉGUERPIR, n. 65 ou not. 105 SANS

HÉSITER, SANS TORTILLER, SANS DÉTOUR, *R'* 12) ou *s'ingâmbier* 'trébucher' (ALW 15, not. 29, *L* et n. 16).

Nul doute que les utilisateurs d'une base de données produisant en temps réel une telle liste de formes issues du même étymon découvriront rapidement l'intérêt de l'outil.

Conclusion

Parmi les sources géolinguistiques, les corpus exploitables informatiquement sont à l'image des atlas : divers, plus ou moins construits, bruts ou enrichis. On l'aura compris, la richesse des corpus informatisés et consultables émanant des divers atlas linguistiques dépendra donc de la conception même de ces atlas ; il revient à chaque équipe la responsabilité de s'interroger sur les données utiles, voire nécessaires à une « corpus-isation » de leur atlas.

Les deux types d'enrichissements présentés ci-dessus, l'étymologisation des formes et la transcription en orthographe Feller, ne sont que des exemples d'informations intégrables dans le cadre de l'informatisation de l'*Atlas linguistique de la Wallonie*, qui rendent le corpus plus riche que le corpus constitué par les enquêtes initiales. D'autres apports des rédacteurs mériteraient certainement d'être analysés en vue d'une intégration dans le projet d'informatisation.

Il reste à souhaiter que ces nouvelles voies d'accès à l'ALW en tant que *thesaurus* incitent les utilisateurs à entrer dans la dimension monographique de l'œuvre.

Références bibliographiques

ALiR = Contini M. *et al.* (éd.) (1996-). *Atlas Linguistique Roman*. Roma : Istituto Poligrafico e Zecca dello Stato.

ALW = Remacle L., Legros E. *et al.* (1953-.). *Atlas linguistique de la Wallonie* (10 volumes). Liège : Université de Liège.

Boutier M.-G. (2008). « Cinq relations de base pour traiter la matière géolinguistique : réflexions à partir de l'expérience de l'Atlas linguistique de la Wallonie », *Estudis Romànics* 30 : 301-310.

Dalbera J.-P. (2002). « Le corpus entre données, analyse et théorie », *Corpus* [En ligne], 2|2002, mis en ligne le 15 décembre 2003, consulté le 1er mars 2013. URL : http://corpus.revues.org/10.

Feller J. (1905). *Règles d'orthographe wallonne.* Liège : Vaillant-Carmanne (BSW 41/2 : 45-96).

Le Dû J. (1992). « L'informatisation des atlas linguistiques de la France », *in Actes du Congrès International de Dialectologie*, IKER 7. Bilbao : Académie de la langue basque, 299-317.

Pierret J.-M. (1992). « La notation courante des langues romanes : "l'orthographe Feller" », *in* W. Bal (coord.) *Līmēs I. Les langues régionales romanes en Wallonie.* Bruxelles : Traditions et Parlers populaires Wallonie-Bruxelles, 25-33.

Pop S. (1950). *La dialectologie. Aperçu historique et méthodes d'enquêtes linguistiques.* Louvain : chez l'auteur, 1950.

Renders P. (2011). Modélisation d'un discours étymologique. Prolégomènes à l'informatisation du Französisches Etymologisches Wörterbuch. Liège : Université de Liège (thèse de doctorat).

Bases, Corpus, Langage (UMR 7320)

CORPUS 12

2013

Dialectologie

Corpus, atlas, analyses

Rita Caprini

E. Baiwir, P. Renders

S. Canobbio

H. Goebl

D. Le Bris

P. Sauzet, G. Brun-Trigaud

A. Malfatto

J.-Ph. Dalbera

J. Bucci

R. Faure, M. Oliviéri

A. Ledgeway

Sommaire

Prix : 20 €

ISSN 1 638-9808
Edizioni dell'Orso