

La banque de données langue corse

Marie José DALBERA-STEFANAGGI
Université de Corse Pasquale Paoli - UMR 6240 (L.I.S.A.)

Résultat de plus de vingt-cinq années d'enquêtes de terrain et d'un partenariat entre l'université de Corse, le CNRS et la collectivité territoriale de Corse, la banque de données langue Corse (BDLC) regroupe une série d'informations d'ordre linguistique et ethnolinguistique présentées sous forme de dictionnaires, d'atlas, de textes sonorisés, de tableaux morphologiques, d'analyses étymologiques, de recherches lexicales... L'accès à ces informations est facilité par la numérisation et la mise en ligne d'une partie de la base. La démarche de la BDLC consiste explicitement à recueillir, stocker, analyser, restituer la langue (et la culture qu'elle véhicule) de manière à faire apparaître les relations de l'homme à son milieu linguistique et culturel.

Depuis les années 80 en effet, la BDLC recueille, à travers des enquêtes de terrain, les témoignages linguistiques de l'oralité. Elle analyse, à partir de ces données, la langue corse à travers ses variations diatopique, diachronique, diastratique. Le paramètre spatial est déterminant dans l'analyse linguistique : on a montré, dès le XIX^e siècle, que la diffusion linguistique obéit à une pertinence géographique, que l'espace « fait sens », que les phénomènes linguistiques, loin d'y être disséminés de manière aléatoire, obéissent à des modes de propagation dont l'analyse s'efforce de démontrer les mécanismes. Les cartes des atlas linguistiques ont abondamment illustré la chose et cela ressort encore plus clairement des cartes modernes dites « interprétatives ». C'est à travers l'analyse spatiale que l'on a accès, bien souvent, à une vision diachronique des faits linguistiques, tant il est vrai — et cela est particulièrement précieux dans le cas de langues sans tradition écrite — que, un peu comme dans une carte géologique, en matière de langue, l'espace reflète le temps. Ruptures et continuités, articulées sur des réalités historiques et/ou géographiques, révèlent ainsi de manière plus ou moins nette la structuration d'une aire donnée.

L'analyse diachronique s'appuie également, dans l'enquête dialectale, sur le paramètre diastratique : à chaque génération, à chaque communauté ses caractéristiques linguistiques, qui permettent d'aborder la diachronie en temps réel. Les variations linguistiques en synchronie renvoient bien souvent en effet à des changements passés ou à venir. Autant dire que, plus que toute autre approche, la dialectologie a le souci d'insérer fortement l'homme dans ses repères spatio-socio-temporels. Mais surtout, le dialectologue privilégie, dans l'enquête linguistique, le registre « familier » : c'est en effet dans le vernaculaire que l'on saisit le plus sûrement l'adéquation de l'homme à son milieu naturel et culturel. Sans vouloir faire de hiérarchie, il faut reconnaître — et le dialectologue en fait l'expérience quotidienne — que c'est le recueil de savoirs populaires, transmis à travers l'enquête orale, effectuée en situation, qui permet d'atteindre le plus sûrement et le plus profondément la langue. D'où la priorité donnée, dans l'enquête, aux thèmes relatifs à la vie quotidienne, aux réalités domestiques, à l'environnement villageois, aux pratiques traditionnelles, etc. La toponymie en fait partie ; elle constitue d'ailleurs un module autonome de la BDLC. Sa fonction est de révéler, à travers les dénominations de l'espace et leur variation dans le temps, comment l'homme s'y projette et s'y inscrit culturellement.

Dans son état actuel, la BDLC se présente donc comme un trésor informatisé qui met en connexion des données (brutes et analysées) issues de recherches pluridimensionnelles sur le terrain corse. Sur des données linguistiques totalement originales (elles représentent la mémoire de la dernière moitié du XX^e siècle) sont développées un certain nombre d'analyses et de mises en forme linguistiques : dictionnaire corse-français (interactif) et français-corse, cartographie phonétique et lexicale, analyse morphologique, lexique étymologique, lexiques spécialisés ; un module spécifique est consacré à la morphologie verbale, un

autre à la microtoponymie, recueillie également sur la base d'enquêtes orales. Ces données linguistiques et ethnolinguistiques sont enrichies et illustrées de ressources sonores (dictionnaire, banque de textes et atlas « parlants »), photographiques et vidéo, également originales, qui font de la BDLC un outil multimédia.

Le travail a pris corps, à l'origine, dans le cadre d'un programme CNRS (GDR) de réalisation des atlas linguistiques de la France par régions. Poursuivi, dans les années 90, dans le cadre d'un programme régional (collectivité territoriale de Corse) de constitution d'une base de données linguistiques multimédia informatisée, il est dès lors mené conjointement avec l'UMR 6039 de Nice-Sophia-Antipolis. Simultanément, sur la base de ce programme, nous obtenons à l'université de Corse une jeune équipe puis une équipe d'accueil. Depuis 2008, la BDLC est intégrée à l'UMR Lisa. Elle a connu ces derniers mois une avancée significative avec l'arrivée d'un ingénieur d'études analyse et traitements de données.

Le travail est aujourd'hui sur la Toile (bdlc.univ-corse.fr), avec des enrichissements réguliers. On compte environ 100 000 fiches lexicales et, prochainement, sera mis en ligne un corpus de 1 500 textes, représentant environ 250 000 mots. Il s'agit de la transcription de productions orales (sur le mode « le corse ordinaire », complètement nouveau sur le domaine corse). Ce corpus devrait pouvoir faire l'objet de diverses analyses, en particulier syntaxiques ou pragmatiques. Deux collections de publications sont issues de cette base de données : atlas linguistiques (3 parus à ce jour) et volumes de type matériaux et analyses (2 volumes à ce jour) ●