Numérisation des données dialectales d'oïl : le projet APPI comme laboratoire

Esther Baiwir et Pascale Renders
Université de Lille[1]

0. Introduction

La révolution numérique a provoqué depuis quelques années un mouvement international d'informatisation des ressources lexicographiques, visible à travers les publications des colloques scientifiques Euralex et eLex (v. http://euralex.org et https://elex.link) ainsi que sur les portails dédiés aux grandes langues européennes (voir par exemple http://ivdnt.org/onderzoek-a-onderwijs/projecten/gigant, http://woerterbuchnetz.de/ ou http://www.cnrtl.fr/). Les objectifs sont non seulement d'améliorer l'accessibilité des données, mais aussi d'ouvrir la voie à de nouveaux modes de consultation et d'exploitation, notamment via la mise en réseau de ces ressources. L'Europe encourage depuis 2014 la numérisation, la mise en ligne et la mise en réseau des dictionnaires des grandes langues européennes via l'action COST « European Network of e-Lexicography » (www.elexicography.eu).

La linguistique historique française et romane n'échappe pas à ce mouvement. Le format numérique est privilégié, que ce soit dès la création d'une ressource (par exemple pour le *Dictionnaire du Moyen Français* – DMF ou le *Dictionnaire étymologique roman* – DÉRom) ou dans le cadre de la rétroconversion d'un ouvrage au départ imprimé (par exemple pour le *Trésor de la Langue Française* – TLFi ou le *Französisches etymologisches Wörterbuch* – FEW). De nombreuses ressources ne sont toutefois pas encore disponibles sous une forme numérique, que ce soit pour des motifs humains (manque de moyens ou de porteurs de projets) ou techniques. Parmi ces ressources, les atlas linguistiques, ouvrages de référence pour la dialectologie et ressources essentielles pour la linguistique galloromane, présentent des matériaux non linéaires, contrairement aux dictionnaires. Leur numérisation ne peut donc s'envisager sans une transformation radicale de leur microstructure. Bien que le plus ancien, l'*Atlas linguistique de la France*, ait fait l'objet d'un projet de mise en ligne sous la forme d'images (voir http://cartodialect.imag.fr/cartoDialect), les atlas linguistiques par régions sont pour la plupart cantonnés à des éditions physiques, voire encore à l'état de matériaux inédits. L'intégration de ces atlas dans le réseau lexicographique en construction est pourtant nécessaire : outre la sauvegarde d'un patrimoine linguistique en voie de disparition (et des matériaux d'enquête eux-mêmes, manuscrits rarement numérisés), la mise en ligne des données dialectales permet d'accéder enfin à des données qui éclairent et corrigent les ressources lexicographiques de la Galloromania et de mettre à jour ces dernières (cf. Baiwir/Renders 2013). La mise en réseau des atlas remédierait également au cloisonnement qu'ils opérent entre les différents domaines géolinguistiques, ouvrant ainsi la voie à une meilleure compréhension des marges entre les aires linguistiques traditionnelles (v. Deparis 1973), voire à une réflexion sur les liens entre dialectes et français régionaux (v. Avanzi 2017). La question de la faisabilité d'une intégration de ressources atlantographiques dans un réseau essentiellement lexicographique se pose toutefois, ainsi que la pertinence des diverses modalités d'intégration possibles, d'un point de vue linguistique et d'un point de vue informatique.

[1] Univ. Lille, EA 1061 - ALITHILA - Analyses Littéraires et Histoire de la Langue, F-59000 Lille, France.

Édition numérique — *Atlas pan-picard informatisé* (APPI)
Projet financé par l'Agence nationale de la Recherche
Première mise en ligne le 4 avril 2019
https://appi.univ-lille.fr

C'est pour répondre à ces questions qu'a été initié le projet APPI, acronyme de *Atlas pan-picard informatisé*, qui vient de recevoir un financement de l'ANR pour les années 2018, 2019 et 2020 (appel générique 2017). Au-delà de sa focalisation sur un domaine dialectal, certes particulièrement intéressant, mais géographiquement réduit, le projet a pour objectif plus général de préparer la future intégration des ressources atlantographiques dans le réseau de lexicographie numérique en analysant tous les aspects du problème, qu'il s'agisse des modalités de mise en relation avec les ressources lexicographiques concernées ou des modalités de numérisation des données elles-mêmes, qui se présentent sous des formes très variées selon les atlas.

Après un rappel de l'état du numérique en dialectologie et en lexicographie galloromane (1), cette communication présente le projet APPI en détaillant les trois phases du projet : la constitution d'un corpus rassemblant des données hétérogènes (2.1), la numérisation proprement dite, sous un nouveau format permettant le dialogue entre l'atlantographie et la lexicographie (2.2), et l'intégration dans le réseau lexicographique (2.3). Sont finalement décrits les résultats attendus du projet (3), le plus attendu étant sans aucun doute la valorisation du patrimoine linguistique dialectal par les nouveaux modes d'exploitation qu'offre le format numérique.

1. La révolution numérique en lexicographie et dialectologie galloromanes

La lexicographie galloromane est dominée par le FEW, qui en constitue le « dizionario-tetto » (Buchi/Renders 2013). Ce véritable thesaurus rassemble l'ensemble des lexèmes du domaine galloroman (français et dialectes d'oïl, gascon, occitan, franco-provençal) des origines à nos jours. L'informatisation des 25 volumes, entamée en 2012 à l'ATILF (http://www.atilf.fr/few), a ouvert la voie à la mise en réseau des ressources galloromanes, amenées à préparer leur connexion numérique autour de cet ouvrage central (cf. Renders/Baiwir/Dethier 2015). Parmi les travaux concernés se trouvent d'une part des ouvrages à visée diachronique, qui complètent le FEW pour une période chronologique déterminée (par exemple le DMF et le TLF cités supra, ou encore le *Dictionnaire étymologique de l'ancien français* – DEAF), d'autre part des ouvrages à visée diatopique, qui examinent un domaine linguistique dialectal ou régional (*Dictionnaire suisse romand* – DSR, *Dictionnaire historique du français québecois* – DHFQ, *Atlas linguistique de la Wallonie* – ALW notamment). Toutes ces ressources, dont une partie sont déjà disponibles sous une forme numérique (DEAF, DMF, TLF) corrigent et complètent le FEW, qui leur apporte à son tour un cadre structurant et englobant.

En ce qui concerne les dialectes, la discipline la plus répandue est l'étude de la répartition diatopique des traits linguistiques, ou *géographie linguistique*. En France, c'est le suisse Jules Gilliéron (1854-1926) qui porte la discipline sur les fonts baptismaux en lançant la campagne d'enquêtes pour l'*Atlas linguistique de la France* (ALF). L'enquête, constituée de traductions des mots et phrases du français vers les parlers locaux, sera menée par Edmond Edmont, entre 1897 et 1901 (v., par exemple, Brun-Trigaud, Le Berre et Le Dû 2005). Quelques décennies plus tard, Albert Dauzat, épinglant les défauts d'une entreprise qui, au-delà de ses qualités extraordinaires, avait souffert de l'immaturité de la discipline, rassemble des équipes de scientifiques locaux et conçoit le projet d'une seconde campagne d'enquêtes linguistiques et ethnographiques dans toute la France (v. Séguy 1973 et Straka/Gardette 1973), mais en morcellant le territoire en zones *a priori* homogènes aux niveaux linguistique et culturel (ce découpage étant donc forcément antérieur aux conclusions auxquelles devraient amener ces atlas régionaux — v. par exemple Tuaillon 1976 : 27-28). D'une part, il s'agissait d'affiner les questionnaires par rapport aux réalités locales : l'interdépendance entre les *mots* et les *choses* est en effet au coeur de l'entreprise (« Wörter und Sachen »), ce qui n'était pas le cas pour l'ALF. D'autre part, l'idée était de resserrer le réseau des points d'enquête pour former un maillage plus apte à éclairer des micro-faits de langues.

La fragmentation du domaine et l'objectif de coller aux réalités locales eurent pour conséquence un éclatement des méthodes d'enquête ainsi que des collections de matériaux plus hétéroclites. Ceux-ci constituent néanmoins de précieux témoignages d'un passé linguistique ayant forgé l'identité des régions de France, aux niveaux toponymique ou anthroponymique, folklorique et culturel et éclairant l'histoire sociale et littéraire française. La disparition des locuteurs dialectaux rend aujourd'hui ces enquêtes d'autant plus irremplaçables et leur édition essentielle. Malheureusement, seule une partie est éditée ; quant aux les matériaux publiés, ils sont eux aussi peu accessibles, pour plusieurs raisons. L'éparpillement des informations, la disparité des différents corpus (voir Baiwir 2017 pour une revue systématique des entreprises atlantographiques du nord de la France) et l'opacité des matériaux non analysés rend difficile l'exploitation de ces données, aussi bien pour les linguistes non-dialectologues (les membres du projet DÉRom, par exemple) que pour les spécialistes de la sous-discipline. Cette difficulté d'accès a pour conséquence directe l'impossibilité de s'en servir pour mettre à jour les ouvrages lexicographiques de synthèse (v. Buchi/Renders 2013), en particulier le FEW.

La numérisation de ces atlas constituerait un pas important vers leur meilleure exploitation. Divers projets atlantographiques exploitent déjà la voie de l'informatisation et de l'exploitation des ressources informatiques, que ce soit pour acquérir une nouvelle dimension – c'est le cas des atlas audiovisuels en Suisse (*Atlas linguistique audiovisuel du Valais romand*) ou en France (*Atlas linguistique multimédia de la région Rhône-Alpes et des régions limitrophes*, *Thesaurus occitan* — THESOC) – ou pour exploiter d'autres voies théoriques – telles que la dialectométrie formalisée par Hans Goebl (1982 ou 1987, par exemple). Cet apport du numérique est prometteur, mais encore timide. Le pas de la rétroconversion en langage XML pour formaliser les liens avec les autres ressources est une voie qui n'a pas encore été défrichée.

À la limite entre géolinguistique et lexicographie existe également l'admirable travail effectué ces dernières années autour du *Glossaire des patois de la Suisse romande* (GPSR), balisé, rétroconverti et interrogeable en ligne (à l'adresse http://www.unine.ch/gpsr). Toutefois, si son contenu présente une dimention diatopique importante, sa structure est de type purement lexicographique et la rétroconversion s'apparente à celle de dictionnaires tels que le TLF.

Le projet le plus ancien de numérisation d'atlas galloromans est le THESOC, rassemblant les données d'atlas avec les ressources de nouvelles enquêtes. La modélisation informatique de ce dernier projet a fait l'objet de plusieurs révolutions depuis la création du projet en 1992, et l'actuelle présentation des données (voir Oliviéri / Casagrande / Brun-Trigaud / Georges 2017) ouvre des perspectives, même si ce projet n'a pas de vocation à devenir un réseau pan-galloroman.

Enfin, dans le panorama des ressources dialectales, l'ALW, qui propose à la fois des cartes et des notices, s'avère particulièrement intéressant à cause de son statut mixte atlanto-lexicographique. Son informatisation a fait l'objet de quelques études préliminaires, accompagnant une première mise en ligne de notices scannées (cf. http://alw.philo.ulg.ac.be/). Parmi ces études, citons les tentatives de structuration en langage RDF (Mazziotta 2008 et 2011), sur la base de la modélisation scientifique proposée par Boutier 2008. Grâce à l'ALW, la réflexion concernant la modélisation informatique des matériaux dialectaux d'oïl est donc entamée.

2. Le projet APPI

C'est à partir du domaine picard qu'est envisagée la réflexion sur la mise en réseau des atlas linguistiques avec les dictionnaires du domaine galloroman. Ce dialecte est en effet bien documenté, grâce aux matériaux recueillis dans le cadre de l'*Atlas linguistique et ethnographique picard* (ALPic). Les deux volumes publiés en 1989 et en 1997 sous l'égide du CNRS contiennent 660 cartes, éditant une partie des matériaux des 1150 questions de l'enquête effectuée à partir de 1980 par Fernand

Carton et Maurice Lebègue, à laquelle il convient d'ajouter les questions des enquêtes antérieures de Loriot, Dubois et Deparis (1960-1968). Les 660 faits de langue relevés en 127 points d'enquête et cartographiés ont été considérés par les auteurs de l'ALPic comme les plus représentatifs des enquêtes, un des critères de sélection étant leur répartition sur l'ensemble du domaine considéré. Le premier tome (cartes 1 à 317) couvre les notions appartenant au champ sémantique de la vie rurale, vocabulaire particulièrement menacé de disparition, tandis que le second tome (cartes 318-660) complète le corpus au niveau ethnographique, lexical, morphologique et phonétique. L'ensemble peut être considéré comme représentatif du dialecte des cinq départements picards de France que sont le Nord, le Pas-de-Calais, la Somme, l'Oise et l'Aisne.

Le domaine picard de Belgique a pour sa part été recueilli lors des enquêtes de Jean Haust pour l'*Atlas linguistique de la Wallonie* (ALW), effectuées entre 1924 et 1959 sur l'ensemble du territoire de la Belgique romane. Les rédacteurs de l'ALW ont actuellement édité la moitié des matériaux dialectaux recueillis, en 10 volumes sur les 20 prévus. Ces 10 volumes comptabilisent ensemble quelque 1670 notices et 812 cartes, couvrant une grande variété de champs sémantiques sur l'ensemble du territoire belge, qui comprend le picard, mais aussi le wallon et le gaumais (v. http://alw.philo.ulg.ac.be/publications/liste-des-volumes-publies/). La particularité de l'ALW est qu'il s'agit d'un atlas explicatif, dans lequel les cartes sont subordonnées au texte. Dans chaque notice, les rédacteurs s'attachent à livrer des matériaux intégrés dans un cadre historique galloroman, grâce à un dialogue constant avec le FEW et avec d'autres ressources lexicographiques. Les cartes apparaissent uniquement lorsque la répartition aréologique des matériaux est intéressante. Contrairement aux pratiques françaises, les cartes deviennent donc une illustration des matériaux et non plus le coeur de la microstructure, dans une démarche tirant l'atlantographie vers la lexicographie.

Le domaine picard a donc pour particularité de se répartir entre la France et la Belgique, qui l'ont traité très différemment de part et d'autre de la frontière. Il permet dès lors la rencontre entre deux méthodes différentes, l'une purement atlantographique (du côté français), l'autre intermédiaire entre diatopie synchronique et lexicographie historique (du côté belge). Outre les prémisses de l'informatisation de l'ALW, qui constitue un pont possible entre les données atlantographiques et les ressources lexicographiques, le projet s'appuie sur deux autres résultats préliminaires : tout d'abord l'informatisation en cours du FEW, dont l'étude préalable constitue un « intérêt paradigmatique pour d'autres projets et réalisations de numérisation en philologie » (M.-D. Gleβgen, en présentation de Renders 2015, quatrième de couverture) ; ensuite l'examen des possibilités techniques de la réunion des matériaux de l'ALPic et de l'ALW ainsi que des verrous d'une telle entreprise (Baiwir 2016).

Le projet se décompose en trois étapes. La première consiste à rassembler les données des deux atlas en un corpus homogène pan-picard. Une modélisation informatique adéquate permettra ensuite la création proprement dite de la ressource APPI, qui sera mise à la disposition de la communauté scientifique. Enfin, la dernière étape étudie la mise en relation de cette nouvelle ressource numérique avec le FEW et envisage l'extension du modèle à d'autres domaines linguistiques.

2.1 **Constitution d'un corpus atlantographique pan-picard**

L'objectif de la première étape est de permettre, par la réunion de matériaux éparpillés, une appréhension globale du fait picard qui ne soit plus cloisonnée par les frontières administrative et méthodologique entre la France et la Belgique.

La constitution du corpus se donne comme base de travail les 660 faits de langue répertoriés dans les volumes 1 et 2 de l'ALPic, considérés comme représentatifs de la langue dans tous ses aspects (phonétique/phonologique, morpho-syntaxique et lexical). Cette édition est *a priori* fiable ; néanmoins, des questions de découpage syntaxique ou de lemmatisation pourront apparaître. Les

archives de la Somme permettront de compléter, vérifier ou documenter les données rassemblées pour l'APPI par la consultation des cahiers d'enquêtes d'origine de l'ALPic. Y sont adjoints les matériaux picards de l'ALW présentant les mêmes faits de langue.

La réunion des matériaux de l'ALPic et de l'ALW en un corpus atlantographique homogène, interrogeable comme un tout, nécessite de résoudre un grand nombre d'obstacles liés à l'hétérogénéité des données atlantographiques elles-mêmes. Chaque atlas possède en effet ses méthodes et ses défauts propres : c'est vrai dans le nord de la France comme partout ailleurs (Pop 1950, Baiwir 2017). Dans le cas picard, les données divergent toutefois particulièrement selon qu'elles proviennent de France ou de Belgique. Les différences entre les deux ressources sont nombreuses, qu'elles concernent la chronologie des enquêtes, la densité du réseau des points d'enquête et le système de référencement de ceux-ci, le système de transcription phonétique ou la présentation des matériaux. Baiwir (2016) envisage les difficultés liées à l'homogénéisation d'un corpus constitué de ces deux ressources atlantographiques aux structures si différentes.

Les divergences de transcription phonétique seront par exemple résolues par une normalisation en alphabet phonétique international (API). L'homogénéisation des points d'enquête sera effectuée en ajoutant à la numérotation de l'ALPic des sigles renvoyant aux différents départements, de façon à assurer une numérotation cohérente des points d'enquête. Enfin, un géoréférencement de tous ces points d'enquête permettra au corpus d'être interrogé sous sa forme atlantographique, via des cartes en ligne englobant l'ensemble du domaine picard.

2.2 **Transfert vers une ressource numérique atlanto-lexicographique**

A partir de la collection atlantographique classique de matériaux bruts résultant de l'étape précédente, il s'agit ensuite de créer une ressource numérique qui puisse être intégrée dans un réseau d'ordre lexicographique. La méthodologie proposée accorde un rôle important au modèle de description dialectale offert par l'ALW, qui situe chaque fait de langue dans l'histoire de sa famille lexicale. Cette pratique de l'« étymologie-histoire du mot » permet d'inscrire pleinement les matériaux dialectaux dans la discipline de la lexicographie historique, en leur apportant une dimension diachronique dont ils sont dépourvus dans leur forme atlantographique. Le projet APPI fait l'hypothèse que l'ajout aux données atlantographiques de cette épaisseur historique est, dans le domaine galloroman, la solution la plus efficace et la plus pertinente pour intégrer les atlas dans un réseau de ressources lexicographiques numérisées autour du FEW.

Concrètement, cette épaisseur historique est apportée par l'étymologisation des données et leur intégration dans l'une des familles lexicales décrites dans le FEW. Situer les matériaux dans un article du FEW peut dès lors suffire à assurer cette inscription historique de façon minimale : il s'agit d'ajouter aux données dialectales une « référence FEW », dans les meilleures pratiques de la lexicographie historique romane. Cette référence est présente dans les notices de l'ALW ; en revanche, elle est absente de l'ALPic. La réunion des matériaux de l'ALPic et de l'ALW doit permettre une étymologisation rapide d'une partie du corpus, élevant en quelque sorte l'ALPic au niveau de l'analyse lexicographique de l'ALW. Environ 80% des matériaux devraient pouvoir être étymologisés de la sorte. Les données restantes seront traitées au moyen du classement onomasiologique que le FEW propose pour les matériaux d'origine inconnue. De nouveaux étymons pourront également être ajoutés à la nomenclature du FEW. Le projet APPI nourrira ainsi la réflexion en cours sur la façon dont le FEW numérique peut intégrer les mises à jour apportées par les ressources lexicales galloromanes qui le corrigent et le complètent.

D'un point de vue technique, la structuration informatique des données se fera durant cette étape au moyen du langage XML, particulièrement adapté à la formalisation de données textuelles. Le format TEI sera privilégié autant que possible. L'encodage des caractères phonétiques particuliers

aux différentes ressources suivra le standard Unicode ; en cas d'utilisation des zones privées, le codage défini pour la fonte du FEW (Renders 2015) servira de guide.

Dès cette deuxième étape du projet, cette nouvelle ressource numérique sera mise à la disposition de tous, chercheurs, étudiants et grand public. Sa diffusion sera effectuée en *open access*, via une interface d'interrogation en ligne qui contiendra également la bibliographie du projet et une documentation sur les nouvelles possibilités d'exploitation des données mises à la disposition de la communauté scientifique. Cette mise en ligne assurera en même temps la sauvegarde et la valorisation du patrimoine dialectal picard.

2.3 **Mise en réseau grâce au FEW et étude d'extension du modèle**

La numérisation du corpus et l'étymologisation des matériaux ne suffisent pas pour mettre les données picardes en réseau avec le FEW et les autres ressources linguistiques du domaine. Connecter l'APPI au FEW nécessite de définir concrètement les modalités d'implémentation des liens. Une première question concerne les unités à mettre en relation dans chacune des ressources : s'agit-il du lexème, de l'article, de la notice, de la carte ? En fonction de la granularité des unités ainsi identifiées, diverses modélisations informatiques sont possibles. Le FEW propose par exemple deux niveaux de description lexicographique : chacun de ses articles décrit une famille lexicale dans sa globalité, tandis qu'au sein de chaque famille, chaque unité lexicale est définie avec ses caractéristiques propres. La modélisation XML du FEW (Renders 2015) permet d'envisager des liens vers chacun de ces deux niveaux. En pratique, l'implémentation d'une relation vers un article du FEW pourra se contenter d'une URI identifiant cet article, tandis qu'un lien de lexème à lexème nécessitera peut-être la création d'un réseau lexical plus précis, qui se formalisera plus facilement sous la forme de triplets RDF, comme cela a été proposé pour l'ALW suivant les pratiques de l'Ontology Web Language (Mazziotta 2008).

Par ailleurs, l'étude envisagera divers cas de figure selon que les données auront pu, ou non, être étymologisées dans le module précédent. L'informatisation du FEW n'étant pas encore achevée, seules les données de l'APPI qui sont en relation avec les articles du FEW en ligne (actuellement les articles des volumes 16, 17 et 19) seront effectivement mises en réseau d'un point de vue technique. Les autres données du corpus seront pourvues d'une implémentation qui permettra leur mise en réseau dès que les volumes correspondants du FEW seront numérisés.

Au terme de cette troisième étape, le projet APPI constituera ni plus ni moins que la première tentative de mise en réseau entre un atlas et le FEW numérisé. Si cette expérience s'avère concluante, le modèle sera proposé pour les atlas des autres domaines linguistiques de France. Le second objectif de cette phase est en effet l'examen de la faisabilité d'une extension du modèle aux domaines linguistiques connexes (normand, wallon, parlers de l'Ile-de-France et orléanais), de façon à documenter la question des limites entre domaines atlantographiques. Cette question des marges constituera un premier pas vers une extension du modèle APPI à d'autres aires dialectales jusqu'à, peut-être, recouvrir l'ensemble du domaine d'oïl.

3. Conclusion

La transformation d'une ressource papier en une ressource numérique est toujours synonyme d'une meilleure accessibilité de son contenu. Les choix effectués dans la modélisation informatique jouent toutefois un rôle majeur dans les modalités des exploitations rendues ainsi possibles. Dans le cadre du projet APPI, l'accès aux matériaux dialectaux doit être amélioré non seulement par la mise en ligne de ces derniers, mais surtout par l'ajout de métadonnées qui viennent jeter un pont entre les

matériaux de départ et les données d'autres ressources linguistiques. L'homogénéisation des transcriptions phonétiques et des références géographiques, l'analyse historique des données lexicales et l'ajout de références au FEW permettront aux données dialectales picardes d'intégrer pleinement un réseau lexicographique en construction. Grâce au pivot que constitue le FEW, des liens hypertextes seront envisageables avec d'autres ressources, telles que le TLFi ou le DMF, qui intègre de multiples diatopismes ; grâce à ce dernier, on pourra situer l'histoire lexicale du picard moderne dans la diachronie. Les corrections et compléments que fournissent les données dialectales pourront, dans un mouvement dialectique, être intégrés dans les projets partenaires. La linguistique française et romane en retirera des matériaux traités, analysés et exploitables pour documenter les questions de linguistique historique du monde roman. En particulier, des projets pan-romans tels que le DÉRom bénéficieront grandement de l'accès à ces données dialectales.

En dialectologie, la réunion des deux domaines picards au-delà de la frontière franco-belge en un corpus homogénéisé ouvre la voie à des études qui permettront une vision globale de ce dialecte, difficile jusqu'à présent. La transcription en alphabet phonétique international lèvera en outre les difficultés de compréhension de ces matériaux phonétiques (actuellement retranscrits selon des standards différents) et les rendra accessibles à tous au-delà même du monde galloroman. Leur intégration dans des projets atlantographiques pan-romans tels que l'ALiR en sera grandement facilitée. Cette vision globale profitera aussi aux étudiants, aux amateurs et au grand public, qui seront plus à même de découvrir et de comprendre les faits linguistiques qu'ils rencontrent.

En ce qui concerne le grand public, il sera par ailleurs intéressant de voir en quoi l'offre d'une ressource numérique dialectale est susceptible de modifier la perception linguistique de ce dialecte, auquel les locuteurs du nord du domaine n'identifient par toujours leur parler (qu'ils appellent *ch'ti* ou *patois*). Cette absence de conscience picarde a été à plusieurs reprises signalée par Jean-Michel Eloy, qui insiste sur le fait que cette région « ignor[e] même que son parler [est] picard » (Eloy 1997, 80).

De façon plus large, le projet APPI permettra d'étudier en quoi les nouvelles approches permises par le format numérique peuvent conduire la discipline française de la dialectologie à renouveler ses concepts et ses méthodes. L'intégration de l'écrit et de l'oral, par exemple, impossible dans la version imprimée des atlas de France, ouvre la voie à de nouvelles exploitations et à de nouvelles possibilités d'étude du fait dialectal, dorénavant inscrit dans un univers multimedia et non plus cantonné à des fiches manuscrites perdues au fond d'un tiroir d'archive. L'inscription numérique des matériaux dialectaux les rendra par ailleurs accessibles pour de nouvelles études qui pourraient dépasser le cadre linguistique pour contribuer à des recherches pluridisciplinaires en ethnographie, en histoire sociale, en littérature.

En jetant des ponts par-delà deux types de frontières (d'une part la frontière entre des domaines géographiques contigus, mais traités séparément et différemment dans chaque atlas ; d'autre part la frontière entre un traitement atlantographique et un traitement lexicographique des matériaux linguistiques), nous espérons que le projet APPI permettra aux données dialectales picardes de s'inscrire durablement dans le champ des humanités numériques.

Bibliographie

ALAVAL = *Atlas linguistique audiovisuel du Valais romand*, http://www5.unine.ch/dialectologie/Atlas Present.html

ALF = Gilliéron, Jules & Edmond Edmont (1902-1910). *Atlas linguistique de la France*, 1920 cartes, Paris, Champion.

ALiR = Contini Michel (dir.), *Atlas Linguistique Roman* (1996), Vol. I, Tome 1 (Présentation), 232 pages; Tome 2 (Commentaires), 151 pages; Tome 3 (Atlas), 14 cartes, Rome, Istituto Poligrafico e Zecca dello Stato.

ALMURA = *Atlas linguistique multimédia de la région Rhône-Alpes et des régions limitrophes*, http://www.atlas-almura.net

ALPic = Carton, Fernand & Maurice Lebègue (1989-1998). *Atlas linguistique et ethnographique picard*, 2 vol. Editions du CNRS.

ALW = Remacle, Louis, Legros, Élisée, Lechanteur, Jean, Counet, Marie-Thérèse, Boutier, Marie-Guy, Baiwir, Esther (1953-). *Atlas linguistique de la Wallonie*, Liège, Université de Liège (10 vol.).

Avanzi, Mathieu (2017). « Le français régional à substrat picard : étude de géographie linguistique », in *Bien Dire et Bien Aprandre*, 32, numéro thématique intitulé *Le picard moderne : un état de la recherche*, Université de Lille 3, 133-158.

Baiwir, Esther, Renders, Pascale (2013). « Vingt minutes en Utopie : l'ALWi », in Baiwir, Esther & Renders, Pascale (éds), *Actes de la 5e Journée liégeoise de Traitement des Sources Galloromanes (Liège, 16 mai 2013)*, Liège, Université de Liège, publication électronique (http://lingwa.philo.ulg.ac.be/trasogal/baiwir_renders_2013.pdf), 9 p.

Baiwir, Esther (2016). « Un type picard par-delà les frontières : le <nom-jeté> », *Les dialectes de Wallonie*, t. 36, 5-24.

Baiwir, Esther (2017). « La géographie linguistique au nord du domaine d'oïl », *Bien dire et bien aprandre* 32, Le picard moderne : un état de la recherche, 73-100.

Boutier, Marie-Guy (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.

Brun-Trigaud, Guylaine, Yves Le Berre et Jean Le Dû (2005). *Lectures de l'*Atlas linguistique de la France *de Gilliéron et Edmont. Du temps dans l'espace*, Paris, CTHS.

Buchi, Éva et Renders, Pascale (2013). « 46. Gallo-romance I : Historical and etymological lexicography », in *Dictionaries. An International Encyclopedia of Lexicography. Supplementary volume : Recent developments with special focus on computational lexicography*, Gouws Rufus H *et al*. (éd.), De Gruyter, Berlin/New York (Handbooks of Linguistics and Communication Science (HSK) 5/4), 653-662.

DEAF, DEAFplus = K. Baldinger *et al*. (1974–), *Dictionnaire Étymologique de l'Ancien Français*, version informatisée : http://deaf-server.adw.uni-heidelberg.de

Deparis, Claude (1973). « Du picard au wallon: observations sur les parlers modernes de la Wallonie occidentale et du Hainaut français », in Straka, G. & Gardette, P. (éds). *Les dialectes romans de France*. Paris : Editions du CNRS, 462-472.

DÉRom = *Dictionnaire étymologique roman*, http://www.atilf.fr/DERom

DHFQ = Poirier, Claude (1998). *Dictionnaire historique du français québecois. Monographie lexicographique des québécismes*, Québec, Presses de l'Université Laval.

DMF = Martin, Robert et Bazin, Sylvie (dir.) (2015), *Dictionnaire du Moyen Français*, Nancy, ATILF/CNRS & Université de Lorraine, http://www.atilf.fr/dmf.

DSR = Thibault, André (1997/2004²). *Dictionnaire suisse romand*. *Particularités du français contemporain*, Éditions Zoé, Genève.

Eloy, Jean-Michel (1997). *La constitution du picard : une approche de la notion de langue*, Louvain-la-Neuve - Amiens, Bibliothèque des Cahiers de l'Institut de Linguistique de Louvain - Centre d'Etudes Picardes, 262 p.

FEW = Wartburg, Walther von *et al.* (1922–2002). *Französisches Etymologisches Wörterbuch. Eine darstellung des galloromanischen sprachschatzes*, 25 vol., Bonn/Heidelberg/Leipzig-Berlin/Bâle, Klopp/Winter/Teubner/Zbinden.

GPSR = Gauchat, Louis/Jeanjaquet, Jules/Tappolet, Ernest et al. (1924–). *Glossaire des patois de la Suisse romande*, Neuchâtel/Paris, Attinger. Glossaire en ligne : http://www.unine.ch/gpsr

Goebl, Hans (1982). *Dialektometrie : prinzipien und methoden des einsatzes der numerischen taxonomie im bereich der dialektgeographie*, Wien, Österreichischen Akademie der Wissenschaften.

— (1987). « Points chauds de l'analyse dialectométrique : pondération et visualisation », *Revue de linguistique romane* 51, 63-118.

Mazziotta, Nicolas (2008). « Exploitation de l'Ontology Web Language pour la rédaction des notices de l'*Atlas Linguistique de la Wallonie* », in Heiden, S. et Pincemin, B. 2008. *9es journées internationales d'analyses stratistiques des données textuelles*. Lyon,12-14 mars 2008, Lyon, Presses universitaires de Lyon, 823-835.

— (2011). « L'informatisation du *Französisches Etymologisches Wörterbuch*. Concepts pour une approche modélisée commune à l'*Atlas Linguistique de la Wallonie* », ZRP 127, 36-62.

Oliviéri, Michèle, Casagrande, Sylvain, Brun-Trigaud, Guylaine & Georges, Pierre-Aurélien (2017). « Le Thesaurus Occitan dans tous ses états », *Revue française de linguistique appliquée*, XXII-1, 89-102.

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

Renders, Pascale (2015). *L'informatisation du Französisches Etymologisches Wörterbuch. Modélisation d'un discours étymologique*, Strasbourg, ELiPhi-Editions de linguistique et de philologie.

Renders, Pascale, Baiwir, Esther et Dethier, Gérard (2015). « Automatically Linking Dictionaries of Gallo-Romance Languages Using Etymological Information », in Kosem, I., Jakubiček, M., Kallas, J. (Eds.) *et al.*, *Electronic0020lexicography in the 21st century: linking lexical data in the digital age. Proceedings of the eLex 2015 conference, 11-13 August 2015, Herstmonceux Castle, United Kingdom*, 452-460.

Séguy, Jean (1973). « Les Atlas linguistiques de la France par régions », *Langue française* 18 (Les parlers régionaux), 65-90.

Straka, G. et Gardette, P. (1973). *Les dialectes romans de France à la lumière des atlas régionaux.* Actes du Colloque National de Strasbourg, 24-28 mai 1971, Paris, CNRS.

TLF, TLFi = P. Imbs, B. Quemada (dir.) (1971–1994), *Trésor de la langue française. Dictionnaire de la langue du XIX^e^ et du XX^e^ siècle (1789–1960)*, version informatisée : http://atilf.atilf.fr/tlf.htm

Tuaillon, Gaston (1976). *Comportements de recherche en dialectologie française*, Paris, CNRS.

Pour citer cet article : Baiwir Esther & Renders Pascale, « Numérisation des données dialectales d'oïl : le projet APPI comme laboratoire », publication en ligne dans le cadre du projet APPI (Atlas pan-picard informatisé, sous la direction d'Esther Baiwir), avril 2019, 9 p. (https://appi.univ-lille.fr/data/medias/baiwirrenders2019).