Présentation du projet ANR ECLATS

Présentation du projet et des équipes

Le projet ANR ECLATS ou **E**xtraction de **C**ontenus géo**L**inguistiques d'**AT**las et analyse **S**patiale a démarré en novembre 2015 et s'achèvera en novembre 2021. Initialement prévu sur quatre ans, il a bénéficié d'une prolongation d'une année supplémentaire[1].

Le projet réunit quatre équipes :

1 : l'équipe STeamer du Laboratoire d'Informatique de Grenoble (UMR 5217) est spécialisée en informatique et en géomatique pour la conception de systèmes d'information spatio-temporelle et la géovisualisation de données multidimensionnelles et hétérogènes. Paule-Annick Davoine porte le projet et dirige la thèse financée de Clément Chagnaud, tandis que Maeva Seffar a été recrutée comme ingénieure en informatique dans le cadre du projet ;

2 : l'équipe Voix, Systèmes Linguistiques et Dialectologie du Département Parole et Cognition du GIPSA-lab de Grenoble (UMR 5216), avec les dialectologues Elisabetta Carpitelli, Carole Chauvin, Michel Contini auxquels sont associés des membres des laboratoires de Nice, BCL, avec Michèle Olivieri et Guylaine Brun-Trigaud et de Toulouse, CLE-ERSS, avec Patrick Sauzet qui avait porté le projet SYMILA sur l'ALF. E. Carpitelli co-dirige la thèse de C. Chagnaud ;

1 — http://eclats.imag.fr

3 : le Laboratoire Informatique, Image et Interaction, L3I, axe Analyse et Gestion de Contenus, de l'université de La Rochelle, qui s'intéresse à la reconnaissance automatique des contenus dans les documents. Il a également recruté un doctorant sur la partie extraction des données ;

4 : l'équipe IMAGINE du Laboratoire d'InfoRmatique en Image et Systèmes d'information, LIRIS, de Lyon effectue des recherches dans la mise en place de modèles de représentation des mots à partir des données images et dans l'élaboration de mécanismes de reconnaissance via des apprentissages spécifiques.

L'objectif du projet est d'apporter un outillage logiciel et méthodologique facilitant l'extraction, l'analyse, la visualisation et la diffusion des données contenues dans les atlas linguistiques, principalement à partir des données de l'*Atlas Linguistique de la France* (ALF), afin de permettre des recherches novatrices en dialectologie.

Le projet s'attache à :

1. développer des méthodes d'extraction de contenus par vectorisation et annotation de contenus ;
2. proposer un processus de stockage des cartes numérisées afin de faciliter leur exploitation et leur diffusion ;
3. définir des modèles de représentation des données géolinguistiques en vue de leur intégration dans un SIG ;
4. développer des méthodes d'analyse spatiale et de géovisualisation facilitant la production de cartes interprétatives et l'extraction de connaissances géolinguistiques ;
5. promouvoir une démarche collaborative afin de faciliter la mutualisation et la diffusion des données géolinguistiques.

Ces cinq objectifs ont donné lieu, pour le moment (avril 2020), à la mise en place de cinq modules et/ou applications :

1. Extraction
2. CartoDialect
3. ShinyDialect
4. ShinyClass
5. DialectoLOD

Ces outils ont été pensés pour constituer, à terme, une suite cohérente.

1. Extraction

Ce travail fait actuellement l'objet d'une thèse par Jordan Drapeau, sous la direction de Jean-Christophe Burie à La Rochelle.

Un système d'extraction du contenu de cartes de l'ALF[2] a été créé en utilisant des arbres de composants connexes. Le système prend en entrée une image (scan de la carte), et en sortie délivre les informations de la carte regroupées sous forme de couches.

Chaque couche correspond à un type d'information spécifique et à ses positions. L'approche proposée utilise un arbre de composantes connexes basé sur le niveau de gris de l'image d'entrée. Une composante connexe est un ensemble de pixels homogènes. Un objet est dit connexe s'il est fait d'un seul « morceau », ici représenté par des lettres (a). Les deux arbres morphologiques (b) et (c) de la même image (a) et, ainsi que les valeurs de gris clair (resp. foncé) représentent des valeurs entières élevées (resp. faibles). (Cf. Figure 1).

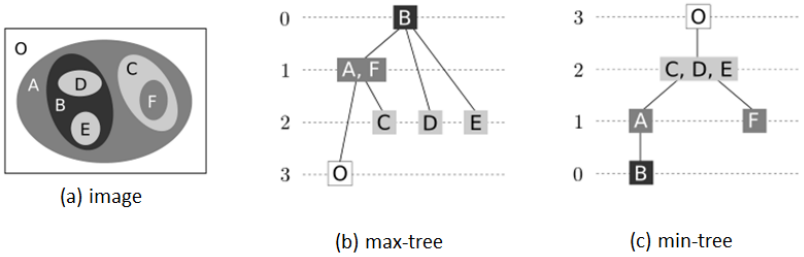

Figure 1 : Exemple de composante connexe

Un filtrage adapté de cet arbre permet d'extraire les composantes souhaitées en utilisant leurs propriétés intrinsèques (Figure 2).

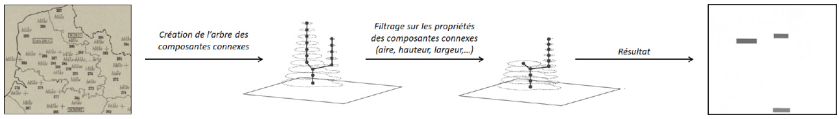

Figure 2 : Exemple de filtrage d'un arbre de composantes connexes sur une portion de carte de l'ALF

Ainsi, la méthode permet de localiser et d'extraire les composantes présentes sur l'image de carte. L'évaluation de cette méthode donne d'assez bons résultats dans l'ensemble (plus de 80 % de précision), notamment pour les noms de département qui sont détectés sans fautes.

Les principaux défauts surviennent lorsque deux types d'informations se touchent physiquement, par exemple lorsqu'un numéro

2 — Cf. J. Drapeau « Extraction », p. 3-4.

de point d'enquête est collé à une frontière. La méthode ne permet pas de séparer ces deux types d'informations.

Cette méthode comporte deux autres inconvénients. Tout d'abord, il est nécessaire de connaître la carte sur laquelle on l'applique pour fixer les seuils de façon optimale, afin de pouvoir la filtrer. Ensuite, le fait que les extractions sont successives répercute les défauts des premières étapes d'extraction sur les suivantes (Figure 3).

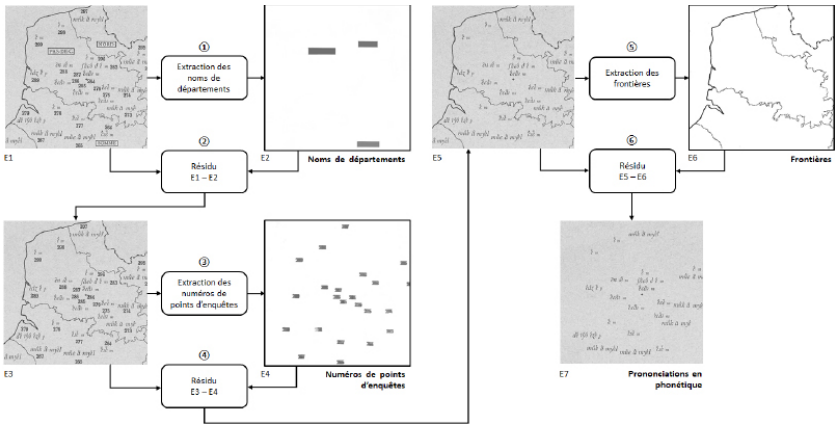

Figure 3 : Le processus complet du système de filtrage pour une carte. Seule une partie de la carte est affichée, mais la carte entière a été donnée en entrée et en sortie.

Afin d'améliorer ces résultats, des travaux basés sur le clustering[3] de composantes ont été entrepris.

Cette nouvelle méthode va extraire différentes informations sur les composantes de l'arbre (hauteur, largeur, périmètre, densité, circularité, excentricité, épaisseur de trait). Ensuite, elle va chercher à regrouper les composantes entre elles de façon à obtenir un groupe pour un type d'information. Les résultats que donne cette nouvelle méthode n'ont pas été chiffrés, mais sont relativement bons visuellement parlant.

Le principal avantage de cette nouvelle méthode est le suivant : il n'est pas nécessaire de connaître la carte sur laquelle on l'applique, car elle regroupe les composantes en fonction de leurs caractéristiques intrinsèques et non à partir de seuils déterminés manuellement comme la méthode précédente.

Cependant quelques défauts persistent, directement liés à notre travail au niveau des composantes connexes. Concrètement,

3 — Cf. D. Prastiwi, « Modification of MSDR ».

lorsqu'une composante qui doit être rattachée à la couche d'information A touche physiquement sur la carte une composante qui appartient à la couche d'information B, il reste actuellement impossible de les séparer lors de l'extraction.

2. CartoDialect

Le produit le plus rapidement mis en place a été la mise en ligne des cartes de l'*Atlas Linguistique de la France* : http://lig-tdcge.imag.fr/cartodialect5/#/ (Figure 4).

Il faut dire que le travail de numérisation avait été déjà amorcé indépendamment par les projets exploratoires *CartoDialect* et *Géodialect* financés par la Mission Interdisciplinarité du CNRS et le Labex Persyval-Lab, ainsi que par le CIRDOC à Béziers à la demande de l'équipe CLE-ERSS de Toulouse.

Cependant, bien que ces travaux aient été effectués sans concertation entre les équipes, ces deux jeux de numérisation se complètent parfaitement. En effet, les deux équipes sont reparties de versions anciennes de l'ALF en grand format, dont certaines cartes étaient endommagées : la réunion des deux jeux a permis d'en obtenir une très bonne copie, diffusable et téléchargeable en ligne.

Figure 4 : CartoDialect. Page d'accueil.

CartoDialect, dans sa version publique, propose différents modes de recherches, en dehors de la simple consultation par le numéro de carte ou par le titre (ou partie du titre). Ainsi, il est possible d'effectuer des sélections de cartes par catégories grammaticales (nom, adjectif, pronom, verbe, etc.) en les combinant avec d'autres critères (genre/nombre pour les noms ou les adjectifs, temps/personne pour les verbes). Cette sélection peut être exportée sous forme de liste.

Dans la nouvelle version mise en ligne en avril 2020, il est également possible de faire des sélections par thème lexical (oiseaux, outils, etc.) et de consulter directement les carnets d'enquêtes conservés par la BNF et mis en ligne dans le site *Gallica*.

On peut également consulter les cartes à partir d'un point de focalisation : le zoom choisi reste actif de carte en carte.

Dans l'optique d'une collaboration avec d'autres équipes de dialectologues, il a été prévu un volet "privé", accessible en se connectant avec un mot de passe. Ce mode 'interprétation' activé permet de rendre les cartes interactives et de cliquer sur les points pour annoter ou consulter les interprétations déjà établies le cas échéant (transcriptions API, lemmatisation, commentaires, etc.), avec une propagation automatique vers les réponses identiques (Figure 5).

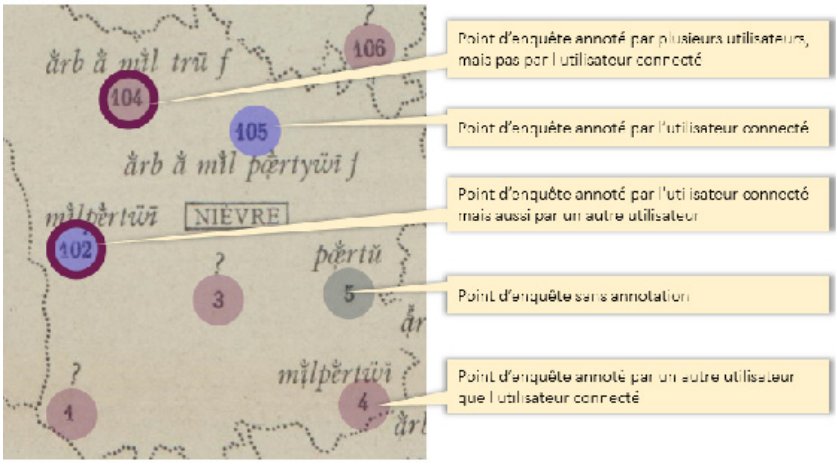

Figure 5 : CartoDialect. Volet interprétation.

Cette partie sera alimentée automatiquement au fur et à mesure que l'extraction du contenu des cartes sera disponible.

À la fin du projet, les cartes seront déposées dans Huma-Num.

3. ShinyDialect

La deuxième application utilisable est ShinyDialect, qui doit son nom au package du langage de programmation R qui a été utilisé pour l'élaborer (Figure 6).

Il permet d'importer et de sélectionner les données relatives à un lemme, et de construire par interpolation spatiale les zones linguistiques homogènes[4].

Son but avoué est de faciliter l'accès à la cartographie pour les linguistes (et les autres) qui n'ont pas d'appétence particulière pour les logiciels (type QGIS ou Cartography sous R) et autres applications en ligne (type MAGRIT[5]) qui en général demandent un apprentissage ou une maîtrise des outils informatiques et de la programmation ou qui ne sont pas adaptés à nos besoins…

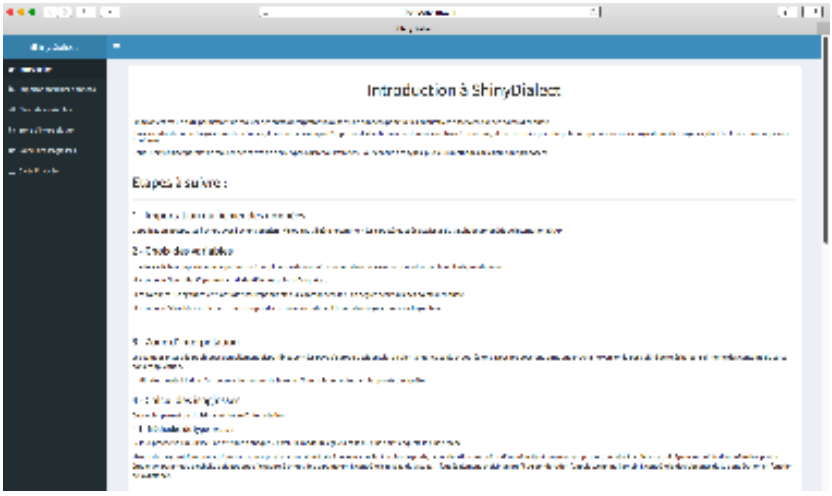

Figure 6 : ShinyDialect. Page d'accueil.

ShinyDialect permet de générer une carte en 5 étapes :

1) l'importation des données s'effectue à partir d'un fichier Excel ou csv, comportant au moins quatre colonnes dont une avec un identifiant pour les points d'enquêtes (nom ou numéro), une avec leur latitude, une avec leur longitude (en WGS84) et une ou plusieurs colonnes avec les données à cartographier (lemmes ou autre) ;

4 — C. Chagnaud, « ShinyDialect », p. 23.
5 — http://magrit.cnrs.fr

2) le choix des variables : le fichier pouvant comporter plusieurs colonnes de variables à visualiser, cette étape permet de choisir la variable à cartographier ;

3) le choix de la zone d'interpolation : pour le moment, l'application donne la possibilité de cartographier des données sur la France, l'Angleterre, les Pays-Bas, la Belgique, l'Allemagne, la Suisse, l'Italie, l'Espagne et le Portugal ;

4) le calcul des isoglosses : cet onglet permet de choisir la méthode d'interpolation qui génèrera les isoglosses à l'aide de différentes méthodes :

- soit celles de type *raster* qui à l'aide de calculs statistiques («inverse de la distance», «plus proche voisin», «gauss», etc.) permet de faire de la généralisation pour obtenir des cartes dans lesquelles les zones dialectales sont plus ou moins schématisées (Figure 7).

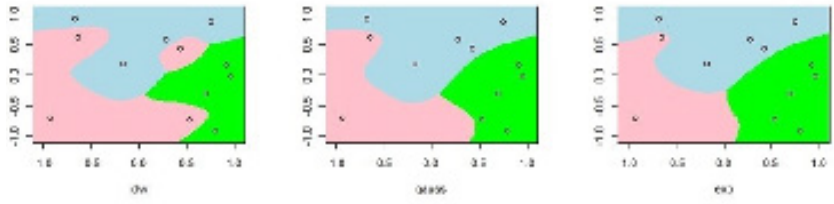

Figure 7 : ShinyDialect. Schémas des méthodes *raster*.

- soit celles de type *vectorielles* qui agrègent, en les lissant ou non, les cellules qui ont les mêmes modalités (Figure 8).

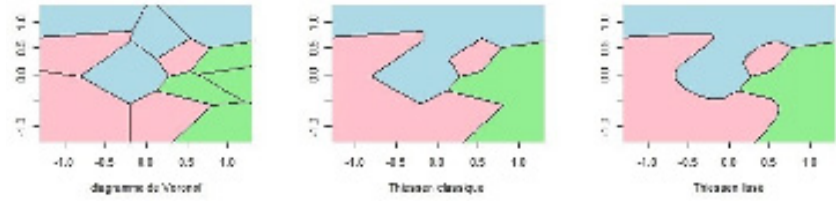

Figure 8 : ShinyDialect. Schémas des méthodes *vectorielles*.

Puis lorsque la carte a été créée, il est possible de modifier les couleurs puisque l'on dispose de 6 groupes de couleurs avec 7 nuances chacun ;

5) enfin la carte finalisée permet de choisir l'habillage de la carte avec le figuré des identifiants (points ou noms) et la légende.

La carte est exportée au format PDF et peut être utilisée sous cette forme ou retravaillée, le cas échéant (Figure 9).

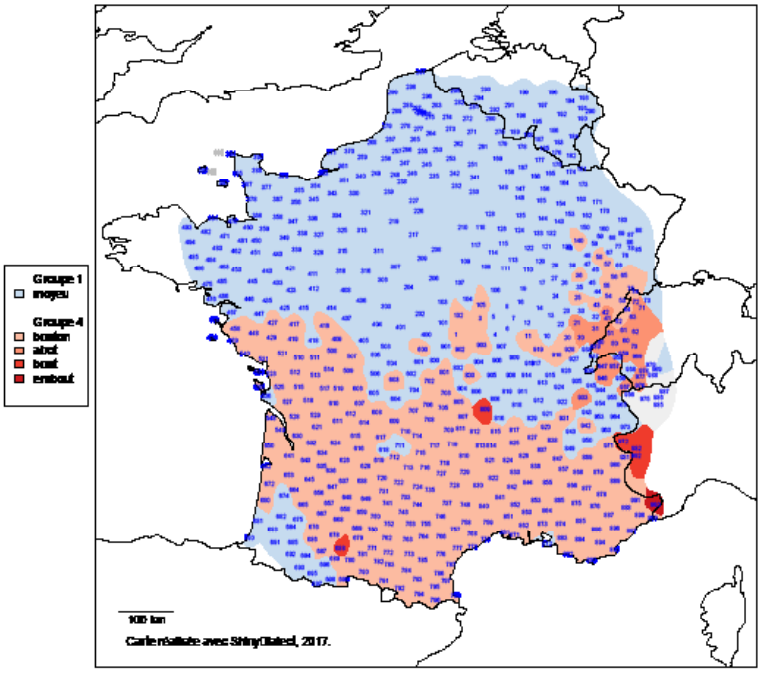

Figure 9 : ShinyDialect. Carte Moyeu d'après ALF 887.

4. ShinyClass

Le but de cet outil est de faire de l'analyse statistique couplant des méthodes de projection et de classification pour identifier des ensembles cohérents au sein d'un corpus d'entités géographiques surfaciques que l'on appelle aires de dispersion[6]. L'idée est de voir si les aires linguistiques ont des contingences, des ressemblances avec des aires géographiques (relief, bassins versants, etc.), historiques, ethnographiques, etc.

Comme pour le moment, nous ne pouvons pas encore bénéficier des données issues de l'extraction de l'ALF, nous avons utilisé un autre jeu de données qui provient d'un travail que nous avons réalisé à partir des données issues du *Thesaurus occitan* pour faire de la dialectométrie sur l'ensemble de l'occitan[7]. 250 cartes lexicales avaient été dépouillées et lemmatisées à cette occasion.

6 — C. Chagnaud, « Classification d'aires de dispersion ».
7 — G. Brun-Trigaud, « Essai de typologie ».

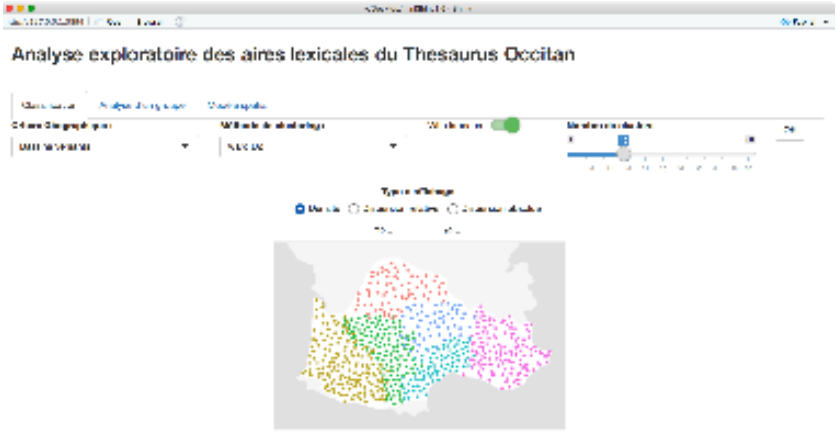

Figure 10 : ShinyClass. Accueil

Chacune des cartes lemmatisées a été traitée avec ShinyDialect et a été transférée dans QGIS pour récupérer les aires sous formes de polygones.

Puis à partir d'un fond de carte, par exemple les sous-bassins versants, un canevas a été créé avec des hexagones.

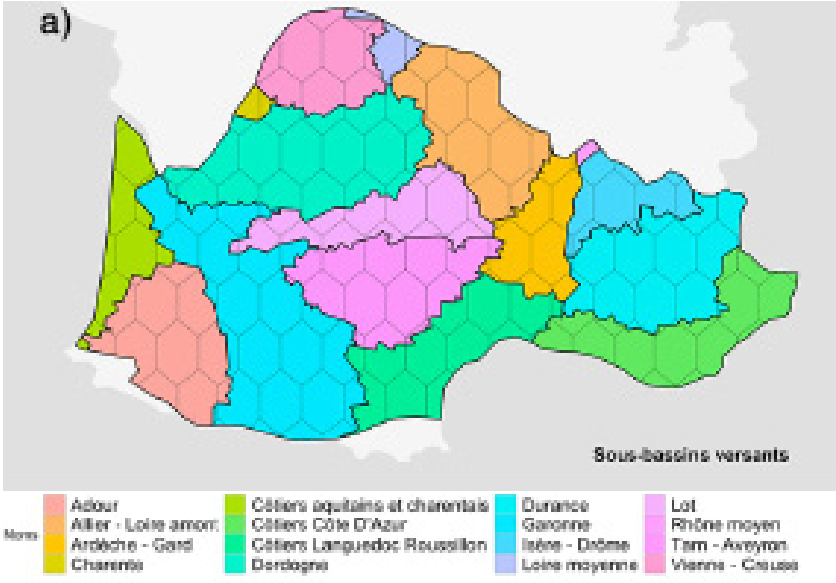

Figure 11 : ShinyClass. Canevas sur les sous-bassins versants

Chaque aire lexicale est analysée avec ces hexagones, ce qui permet par la suite de chercher les corrélations qui pourraient exister entre les aires lexicales et celles des sous-bassins versants.

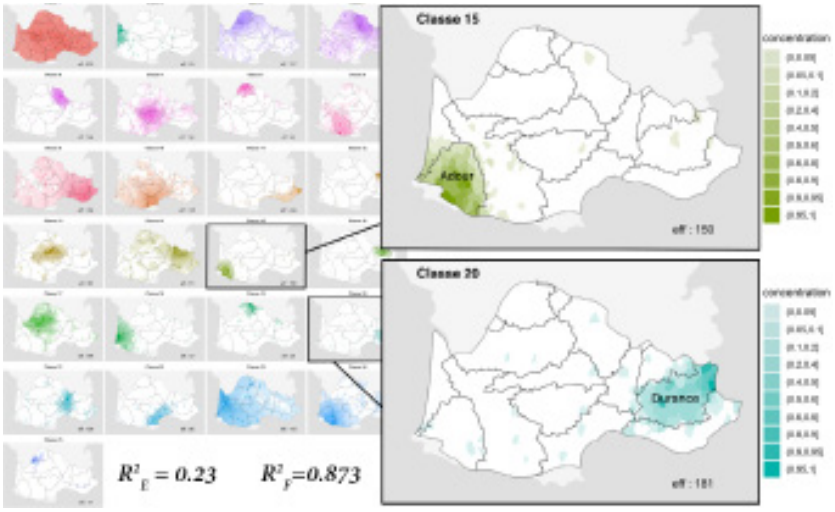

Figure 12 : ShinyClass. Classification en 25 classes à l'aide du facteur géographique des sous-bassins versants

Dans le cas présenté ici, on voit assez rapidement qu'il existe 150 aires lexicales qui s'intègrent bien dans le bassin de l'Adour ou 181 aires pour le bassin de la Durance.

Ensuite on peut entrer dans les détails et voir quels sont les pans du lexique qui sont le plus impliqués.

C'est un outil exploratoire qui devrait permettre aux dialectologues d'appréhender sur de grands corpus la validité de certains concepts concernant la répartition des aires dialectales[8] : sont-elles corrélées entre elles ou non ? Si oui, peut-on déterminer quels sont les facteurs externes qui ont pu jouer un rôle ? Existe-t-il une typologie ? Autant de questions auxquelles il était difficile de répondre jusqu'ici.

5. DialectoLOD

L'application DialectoLOD est un outil de cartographie qui repose sur la manipulation et la superposition de couches géographiques. Elle permet de faciliter la création de cartes géolinguistiques thématiques et assure les fonctionnalités suivantes :

8 — G. Brun-Trigaud, « Identification of lexical areas ».

- la manipulation des entités des cartes avec isoglosses produites sur ShinyDialect (isoler un ou plusieurs lemmes, modifier l'apparence…) afin de mettre en évidence certaines zones de la carte. Ces cartes peuvent être importées localement ou depuis une source publiée sur Geoserver[9] ;

- la superposition de différentes couches raster ou vecteur issues de données géospatiales partagées (cartes anciennes, fonds de cartes, cartes thématiques…) pour enrichir la carte d'informations thématiques ;

- l'exportation de la carte en format PNG afin de faciliter son utilisation (Cf. la figure 11).

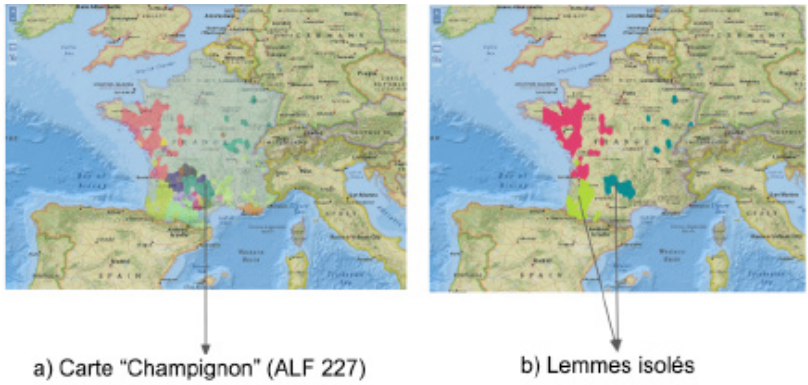

Figure 13 : a) carte des isoglosses correspondant à la répartition des lemmes de la carte ALF n° 227 'Champignon' ; b) carte avec lemmes isolés

9 — http://geoserver.org/

Conclusions

Cette suite d'outils a été conçue pour répondre aux différentes problématiques géolinguistiques, de l'extraction des données des cartes existantes aux nouvelles explorations et projections sur des supports exploitables dans l'univers actuel du numérique. Leur but est à la fois de rendre accessibles les ressources "anciennes" en proposant une interface de visualisation, mais aussi d'enrichir ces données en les reliant à d'autres données du même type ou extra-linguistiques (géographiques, historiques, ethnographiques…).

Guylaine BRUN-TRIGAUD
Université Côte d'Azur, CNRS, BCL
Clément CHAGNAUD
Université Grenoble Alpes, CNRS,
Grenoble INP, LIG
Maeva SEFFAR
Université Grenoble Alpes, CNRS,
Grenoble INP, LIG, STEAMER
Jordan DRAPEAU
Université de La Rochelle, CNRS, L3i

Bibliographie

ALF = J. Gilliéron, E. Edmont, *Atlas Linguistique de la France*, Paris, Champion, 1902-1910.

G. Brun-Trigaud, C. Chagnaud, P. Garat, « Identification of lexical areas templates throughout the Occitan domain », dans *10th International Conference on Language Variation in Europe (ICLaVE10)*, Jun 2019, Leeuwarden, Netherlands. À paraître.

G. Brun-Trigaud, A. Malfatto et M. Sauzet, « Essai de typologie des aires lexicales occitanes : regards dialectométriques » In *Fidélités et dissidences. 12e Congrès de l'Association Internationale d'Études Occitanes (Albi, 10-15 juil. 2017)*. À paraître.

C. Chagnaud, P. Garat, P-A. Davoine, E. Carpitelli, A. Vincent, « ShinyDialect: a cartographic tool for spatial interpolation of geolinguistic data » in *the 1st ACM SIGSPATIAL Workshop on Geospatial Humanities (GeoHumanities'17)*, Nov 2017, Redondo Beach, United States. pp. 23-30.

C. Chagnaud, P. Garat, P-A. Davoine, G. Brun-Trigaud, « Classification d'aires de dispersion à l'aide d'un facteur géographique : application à la dialectologie », dans *Spatial Analysis and GEOmatics - SAGEO*, Nov 2019, Clermont-Ferrand, France. À paraître.

J. Drapeau, T. Géraud, M. Coustaty, J. Chazalon, J. Burie, V. Eglin, et S. Bres, « Extraction of ancient map contents using trees of connected components », in *12th International Workshop on Graphics Recognitio, 14th IAPR International Conference on Document Analysis and Recognition*, *GREC@ICDAR 2017, Kyoto, Japan, November 9-15, 2017*. p. 3-4.

D. Prastiwi, K. Ariyanti, and T. Siswantining, « Modification of MSDR algorithm and ITS implementation on graph clustering », in eds K. A. Sugeng, D. Triyono and T. Mart, *International Symposium on Current Progress in Mathematics and Sciences 2016, ISCPMS 2016: Proceedings of the 2nd International Symposium on Current Progress in Mathematics and Sciences 2016* [030147] (AIP Conference Proceedings; Vol. 1862). American Institute of Physics Inc.