Travaux de Linguistique et de Philologie

Nouveaux regards sur la variation dialectale

New Ways of Analyzing Dialectal Variation

ELiPhi

TraLiPhi – Travaux de Linguistique et de Philologie

André Thibault, Mathieu Avanzi,
Nicholas Lo Vecchio, Alice Millour (éds.)

Nouveaux regards sur la variation dialectale

New Ways of Analyzing Dialectal Variation

ELiPHI
EDITIONS DE LINGUISTIQUE ET DE PHILOLOGIE

Ouvrage publié avec le soutien financier de l'Équipe d'Accueil 4509 (Sens, Texte, Informatique et Histoire), de l'École Doctorale V (Concepts et langages – ED 0433) et du Conseil Académique de Sorbonne Université, ainsi que du projet de recherche CAMARADERIE (porteur : Mathieu Avanzi) subventionné dans le cadre du programme de soutien à la recherche 'Émergence' de Sorbonne Université.

ISBN 978-2-37276-056-0

EAN 9782372760560

Table des matières / Table of contents

A.2. Combiner numérisation et gains empiriques / Combining digitization with empirical gains

B. Nouveaux projets numériques / New digital projects

B.1. Cartographie basée sur le crowdsourcing / Mapmaking based on crowdsourcing

B.2. Études spécifiques basées sur les nouvelles technologies web / Specific studies based on new web technologies

Nouveaux regards sur la variation dialectale : Présentation

Ce volume vise à rassembler des articles de romanistes et germanistes de divers horizons afin d'évaluer comment les nouvelles technologies peuvent améliorer notre connaissance de la variation dialectale en Europe et au-delà. Au cours des dernières années, des progrès technologiques rapides et la démocratisation du Web 2.0 ont conduit à la création et au développement de plusieurs initiatives dialectologiques. En ce qui concerne le recueil de données, plusieurs alternatives au travail de terrain traditionnel sont apparues : grâce à Internet et aux réseaux sociaux, il est désormais possible de réaliser des enquêtes sans avoir à se déplacer ; les téléphones portables ont des microphones intégrés de haute qualité qui permettent d'enregistrer des voix à distance partout dans le monde. En termes d'analyse et de visualisation des données, il est désormais beaucoup moins coûteux de créer des atlas, de les mettre à la disposition de la communauté universitaire et d'en interpréter le contenu. De nos jours, les logiciels de statistiques et de visualisation des données permettent un traitement qui n'aurait jamais été possible il y a seulement quelques décennies. Il faut également souligner que ces avancées nous ont permis de porter un nouveau regard sur les matériaux recueillis par les prédécesseurs, à la fois récents et anciens, et ainsi d'enrichir notre connaissance des changements linguistiques intervenus dans l'histoire des langues européennes.

Dans ce contexte, il n'est pas surprenant que de nombreuses équipes d'universitaires aient pu concevoir des projets innovants ces dernières années, dont beaucoup sont présentés dans cette publication. Les articles de la première moitié du volume (A) portent sur le développement de cartes géolinguistiques à l'aide de nouveaux outils numériques. De nombreux projets sont basés sur des travaux existants (A.1) mais certains d'entre eux ajoutent des données nouvellement recueillies aux ressources déjà existantes (A.2). Les articles de la seconde moitié du volume (B) traitent de nouveaux projets numériques entièrement basés sur des données originales récemment collectées (B.1), dont beaucoup visent des objectifs spécifiques (B.2) tels que l'étude de la variation syntaxique, du nivellement dialectal ou du traitement automatique du langage naturel.

A. Cartographier à l'aide de nouveaux outils numériques

De nos jours, les dialectologues ont souvent une énorme quantité de données à leur disposition, recueillies par d'éminents prédécesseurs et généralement présentées sous forme de cartes, compilées dans des atlas. Mais de tels atlas sont rarement accompagnés de commentaires analytiques, et même lorsqu'ils le sont, il est extrêmement long de réutiliser ces sources à d'autres fins : les anciennes transcriptions phonétiques doivent être modernisées, le matériel doit être numérisé et saisi dans des tableurs afin d'automatiser leur traitement. Bon nombre des atlas les plus célèbres ont fait l'objet d'entreprises de numérisation : pour n'en nommer que quelques-uns, l'ALF (*Atlas Linguistique de la France*), l'AIS (*Atlante italo-svizzero*), l'ALPI (*Atlas Lingüístico de la Península Ibérica*) et le SDS (*Sprachatlas der deutschen Schweiz*). Ces projets tentent souvent d'ajouter du nouveau contenu aux données existantes ou de remettre en question certains choix méthodologiques de leurs prédécesseurs.

1. Numérisation des œuvres traditionnelles

Dans la première contribution (« A digital, retro-standardized edition of the *Tableaux Phonétiques des Patois Suisses Romands* (TPPSR) »), Hans Geisler, Robert Forkel et Johann-Mattis List présentent une édition rétro-standardisée des *Tableaux Phonétiques des Patois Suisses Romands*, un recueil de matériaux dialectaux élaboré par Louis Gauchat, Jules Jeanjaquet et Ernest Tappolet au début du 20e siècle et publié en 1925. Bien que Gauchat et ses collègues n'aient jamais pu mener à bien leur ambition de présenter ces matériaux sous forme d'atlas, faute de financement, les auteurs montrent comment de solides techniques de numérisation, adossées à des approches de rétro-standardisation assurant une transparence dans le traitement des données, permettent de transformer les matériaux des TPPSR en un atlas dialectal moderne et interactif.

L'ALF, publié au début du siècle dernier, a été suivi un demi-siècle plus tard par une série d'atlas régionaux français au réseau de localités plus dense. Dans « Du terrain au numérique : évolution du traitement des données dans l'*Atlas linguistique et ethnographique du Lyonnais* », Fabio Armand explique comment un alphabet phonétique traditionnel a été transposé en API à l'aide d'un processus d'océrisation automatisée. De nouveaux outils analytiques sont en cours d'élaboration afin de tirer le meilleur parti des données numérisées.

Esther Baiwir, dans « Quel *sens* ont les unités lexicales des atlas linguistiques ? Une exploration sémantique dans le domaine (gallo)roman », adopte une perspective critique sur la gestion de la sémantique lors de l'assemblage de matériaux issus de travaux géolinguistiques antérieurs. Dans le projet APPI (*Atlas pan-picard informatisé*), qui vise à agréger les matériaux picards de trois atlas (ALF, *Atlas linguistique et ethnographique picard* et *Atlas linguistique de la Wallonie*), l'analyse de la sémantique des formes est fondamentale pour établir l'architecture du projet.

Dans certains cas, les enquêtes dialectologiques n'ont pas abouti à la publication d'un atlas, mais plutôt à celle d'un glossaire. Néanmoins, il est toujours possible de faire revivre les données et de les transformer en cartes. Telle est l'ambition de Wim Remysen dans « Revisiter les données dialectologiques de la Société du parler français au Canada (1904-1906) : enjeux et perspectives ». L'un des avantages de cet ensemble de données est qu'il comprenait déjà des zones urbaines au début du XX^{e} siècle. Cela a permis à l'auteur d'identifier des processus de nivellement du français en milieu urbain (en particulier à Montréal), contrairement à ce que l'on peut observer en milieu rural.

L'atlas linguistique de la péninsule ibérique a été l'une des nombreuses victimes de la guerre civile espagnole : contrairement à l'ALF, il n'a jamais été entièrement publié, avec un seul volume paru sous forme imprimée (1962). Dans son article (« Odres nuevos para vino añejo: la edición digital del *Atlas Lingüístico de la Península Ibérica* »), Xulio Sousa présente le projet d'édition numérique de l'ALPI, qui permet la génération automatique de cartes et l'exportation de données sous d'autres formats.

2. Combiner numérisation et gains empiriques

L'équivalent italo-roman de l'ALF, l'AIS (*Atlante italo-svizzero*), a été numérisé à l'Université de Zurich par Michele Loporcaro *et al.* ; dans « AIS, *reloaded*: A digital dialect atlas of Italy and southern Switzerland », l'objectif est double : d'une part, développer une version numérique consultable de l'atlas ; de l'autre, incorporer de nouvelles données d'enquêtes menées dans les régions italophones de la Suisse, dans les mêmes localités et avec le même questionnaire, mais pratiquement cent ans plus tard.

En Suisse alémanique, le *Sprachatlas der deutschen Schweiz* a été numérisé dans une version qui permet la visualisation interactive de nombreuses caractéristiques dialectales, mais dans « Les cartes dialectométriques interactives de dialektkarten.ch », Yves Scherrer présente également d'autres méthodes de visualisation reposant sur l'analyse dialectométrique, les prototypes de traduction automatique et l'identification des dialectes ; en outre, l'interface développée pour le SDS est actuellement appliquée à d'autres conglomérats dialectaux.

En fait, toute la région alpine est bien dotée en matière de ressources dialectales. Dans « VerbaAlpina: Digital geolinguistics dedicated to the lexical analysis of the Alpine dialects », Stephan Lücke présente le projet VerbaAlpina, une interface électronique puissante et polyvalente qui utilise la technologie Web pour stocker les informations de diverses publications géolinguistiques traditionnelles (atlas, et glossaires), les rendre disponibles en ligne et les visualiser sur des cartes. De plus, il enrichit les données traditionnelles avec des techniques de *crowdsourcing* en ligne.

Une autre région riche en dialectes est le “Croissant”, à la limite nord du Massif Central en France, où les trois groupes de dialectes fondamentaux de la Galloromania – Oïl, Francoprovençal et Occitan – convergent. Dans « Étude des parlers du Croissant à travers la cartographie informatisée », Amélie Deparis montre comment une équipe de chercheurs travaille actuellement dans une perspective multidisciplinaire sur les langues de la région, combinant la documentation existante avec de nouvelles données empiriques.

B. Nouveaux projets numériques

1. Cartographie basée sur le crowdsourcing

Alors que les projets décrits ci-dessus visent principalement la numérisation et l'enrichissement d'œuvres déjà existantes, d'autres visent à produire des œuvres entièrement nouvelles à partir de données récentes recueillies en ligne par le biais d'enquêtes électroniques – l'approche dite du « crowdsourcing ».

Robert Möller (avec Stephan Elspaß) est en charge de « An online atlas of colloquial German: The *Atlas zur deutschen Alltagssprache* ». On doit cet atlas en ligne à la participation régulière d'environ 10 000 informateurs qui ont participé à des questionnaires en ligne successifs portant principalement sur le lexique, mais aussi sur la grammaire. Comme les questions sont les mêmes que dans le *Wortatlas der deutschen Umgangssprachen* de J. Eichhoff, qui a commencé en 1977, les résultats permettent des comparaisons diachroniques.

Mathieu Avanzi et André Thibault (« Cartographier l'amuïssement et la restitution des consonnes finales en français grâce à la production participative ») ont répliqué la procédure décrite dans l'article précédent et l'ont appliquée aux régions de la Francophonie d'Europe et du Canada. À ce jour, plus de 15 enquêtes ont été menées auprès de plusieurs dizaines de milliers de locuteurs. Les premiers résultats sont disponibles depuis 2016 sous la forme de cartes dans des billets de blog (« Français de nos régions ») ainsi que dans plusieurs ouvrages grand public. Plus récemment, ils ont commencé à étendre leurs enquêtes à des territoires insulaires d'outre-mer, et l'objectif final est de couvrir l'ensemble du monde francophone. La présente contribution montre ce que leurs données permettent d'apporter à la question de l'aréologie et de l'évolution de la prononciation des consonnes finales dans la francophonie d'Europe, en croisant les axes diatopiques et diagénérationnels.

Philippe Boula de Mareüil, en collaboration avec une équipe binationale, supervise un double projet : « For a mapping of the languages/dialects of Italy and regional varieties of Italian ». La première phase traite de la documentation des variétés dialectales et des langues régionales de l'Italie par l'élicitation d'échantillons de discours qui peuvent être entendus et téléchargés en ligne, ainsi que des transcriptions. La deuxième phase se concentre spécifiquement sur l'étude de l'italien régional

à travers le crowdsourcing, en se basant sur des oppositions phonologiques qui peuvent être visualisées à l'aide d'un programme dédié (Cartopho).

2. Études spécifiques basées sur les nouvelles technologies web

Tous les projets ne visent pas à cartographier de longues listes de caractéristiques linguistiques diversifiées. Les nouvelles technologies peuvent également être utiles dans l'étude de problèmes linguistiques plus ciblés.

Miriam Bouzouita, Mónica Castillo Lluch et Enrique Pato, par exemple, présentent un nouveau projet (« *Dialectos del español* : une application pour l'étude de la variation linguistique dans le monde hispanophone ») visant spécifiquement à étudier la variation syntaxique de l'espagnol. Bien que cela puisse sembler très ciblé comme objectif, la portée géographique en est, en revanche, très large : l'enquête est menée dans l'ensemble du monde hispanophone. Techniquement, elle est gérée par une application téléphonique qui s'est déjà avérée efficace avec les dialectes du suisse alémanique et de l'anglais britannique.

Cela nous amène à la contribution de David Britain, Tamsin Blaxter et Adrian Leemann, « Dialect levelling in England: Evidence from the *English Dialects App* ». Les auteurs montrent, grâce aux données recueillies auprès de plus de 50 000 participants dans toute l'Angleterre, que certaines caractéristiques (lexicales pour la plupart) ont subi un nivellement considérable au cours des cinquante dernières années, tandis que d'autres (principalement phonologiques) résistent à cette tendance, en particulier dans les zones où la mobilité est plus faible.

Enfin, dans le dernier article, Delphine Bernhard (« Corpus annoté en parties du discours pour les dialectes alsaciens : comparaison avec l'allemand standard et le français ») montre comment une équipe de spécialistes travaillant sur l'alsacien – mais aussi l'occitan et le picard – réussit à aider à maintenir ces langues minoritaires en vie sur le Net, en développant l'automatisation de programmes servant à annoter syntaxiquement les morphèmes des dialectes alsaciens.

Sorbonne Université

André THIBAULT
Mathieu AVANZI
Nicholas Lo VECCHIO
Alice MILLOUR

New Ways of Analyzing Dialectal Variation: Presentation

This volume aims to bring together articles from Romance and Germanic dialectologists with diverse backgrounds in order to evaluate how new technologies can enhance our knowledge of dialectal variation in Europe and beyond. In the past few years, rapid technological progress and the democratization of the Web 2.0 have led to the successful creation and development of several dialectological initiatives. In terms of data collection, several alternatives to traditional fieldwork have emerged: thanks to the Internet and social networks, it is now possible to conduct dialectological surveys without having to travel; mobile phones have high-quality built-in microphones that make it possible to record voices remotely anywhere in the world. In terms of data analysis and visualization, it is now much less expensive to create atlases, make them available to the academic community, and interpret the data. Nowadays, data visualization and statistical software allow processing that would never have been possible just a few decades ago. It must also be emphasized that these advances have allowed us to take a new look at materials collected by predecessors, both recent and historical, and thus to enrich our knowledge of the linguistic changes that have taken place in the history of European languages.

In this context, it is not surprising that many teams of scholars have been able to devise innovative projects in recent years, many of which are presented in this publication. The articles in the first half of the volume (A) focus on the development of geolinguistic maps using new digital tools. Many projects are based on existing works (A.1), but some of them add newly collected data to the already existing resources (A.2). The articles in the second half of the volume (B) deal with new digital projects entirely based on recently collected original data (B.1), many of which are aimed at specific goals (B.2), such as the study of syntactic variation, dialect levelling, or natural language processing.

A. Mapmaking using new digital tools

Dialectologists nowadays often have an enormous amount of data at their disposal, collected by prominent predecessors and usually presented in the form of geolinguistic maps, compiled in atlases. But such atlases rarely come with analytical commentary, and even when they do, it is extremely time-consuming to reuse the data for other purposes: old phonetic transcriptions must be modernized, the material has to be digitized and entered into spreadsheets, so that searches can be conducted and new maps automatically generated. Some of the most renowned atlases have

undergone digitization processes: to name a few, the *ALF* (*Atlas Linguistique de la France*), the *AIS* (*Atlante italo-svizzero*), the *ALPI* (*Atlas Lingüístico de la Península Ibérica*), and the *SDS* (*Sprachatlas der deutschen Schweiz*). These projects often try to add new content to the existing data, or to question some of the methodological choices of their predecessors.

1. Digitizing traditional works

In the first study, Hans Geisler, Robert Forkel and Johann-Mattis List present «A digital, retro-standardized edition of the *Tableaux Phonétiques des Patois Suisses Romands* (TPPSR)», an early collection of dialect data of the 'Suisse romande', which was compiled by Louis Gauchat, Jules Jeanjaquet and Ernest Tappolet at the beginning of the 20th century and later published in 1925. While the plan of Gauchat and his collaborators to turn their data into a dialect atlas could never be realized due to lack of funding, the authors show how consistent techniques for digitization, accompanied by transparent approaches to retro-standardization, can be used to turn the original data of the TPPSR into a modern interactive dialect atlas.

The *ALF*, published at the beginning of the last century, was followed half a century later by a series of French regional atlases with a denser network of localities. In «Du terrain au numérique : évolution du traitement des données de l'*Atlas linguistique et ethnographique du Lyonnais*» Fabio Armand explains how a traditional phonetic alphabet was transposed into IPA using automated OCR. New analytical tools are on the way in order to make the most out of the digitized data.

Esther Baiwir, in «Quel *sens* ont les unités lexicales des atlas linguistiques ? Une exploration sémantique dans le domaine (gallo)roman», adopts a critical perspective on the handling of semantics when assembling materials from previous geolinguistic works. In the APPI project (*Atlas pan-picard informatisé*), which aims to aggregate Picard materials from three atlases (*ALF*, *Atlas linguistique et ethnographique picard*, and *Atlas linguistique de la Wallonie*), the analysis of the forms' semantics is fundamental for establishing the project's architecture.

In some cases, dialectological surveys didn't lead to the publication of an atlas, but rather to that of a glossary. Nevertheless, it is still possible to revive the data and turn it into maps. This is Wim Remysen's ambition in «Revisiter les données dialectologiques de la Société du parler français au Canada (1904-1906) : enjeux et perspectives». One of the advantages of this data set is that it already included urban areas at the beginning of the twentieth century. This has allowed the author to identify levelling processes in urban settings (particularly in Montreal), in contrast with the situation in rural areas.

The linguistic atlas of the Iberian Peninsula was one of the many casualties of the Spanish Civil War: unlike the *ALF*, it was never fully published, with only one volume appearing in print (1962). In his paper («Odres nuevos para vino añejo: la edición digital del *Atlas Lingüístico de la Península Ibérica*»), Xulio Sousa introduces the digital

edition project of the *ALPI*, which already enables automatic mapmaking and will, at last, make the collected documentation available to all.

2. Combining digitization with empirical gains

The Italo-Romance equivalent of the *ALF*, the *AIS* (*Atlante italo-svizzero*), has been digitized at Zurich University by Michele Loporcaro *et al.*; in «AIS, *reloaded*: A digital dialect atlas of Italy and southern Switzerland», the objective is twofold: on the one hand, developing a searchable digital version of the atlas; on the other, incorporating fresh data from surveys conducted in southern Switzerland, in the same localities and with the same questionnaire—but practically a hundred years later.

In German-speaking Switzerland, the *Sprachatlas der deutschen Schweiz* has been digitized in a version that allows interactive visualization of many dialectal features, but in «Les cartes dialectométriques interactives de dialektkarten.ch», Yves Scherrer also reveals new initiatives that have been undertaken, such as dialectometric analyses, prototypes of machine translation, and dialect identification; furthermore, the interface developed for the *SDS* is currently being applied to other dialectal conglomerates.

In fact, the whole Alpine region is well endowed when it comes to dialect resources. In «VerbaAlpina: Digital geolinguistics dedicated to the lexical analysis of the Alpine dialects» Stephan Lücke introduces the VerbaAlpina project, a powerful and versatile electronic interface that uses web technology to store information from various traditional geolinguistic publications (atlases and glossaries), make it available online, and visualize it. Moreover, it enriches the traditional data with online crowdsourcing techniques.

Another region rich in dialects is the so-called "Croissant," on the northern edge of the Massif Central in France, where the three fundamental dialect groups of Gallo-Romania—Oïl, Francoprovençal, and Occitan—converge. In «Étude des parlers du Croissant à travers la cartographie informatisée» Amélie Deparis shows how a team of scholars are currently working in a multidisciplinary perspective on the languages of the region, combining existing documentation with new empirical input.

B. New digital projects

1. Mapmaking based on crowdsourcing

While the projects depicted above mainly pursue the digitization and enrichment of already existing works, others aim to produce entirely new works based on recent data collected online through electronic surveys—the so-called crowdsourcing approach.

Robert Möller (together with Stephan Elspaß) is in charge of «An online atlas of colloquial German: The *Atlas zur deutschen Alltagssprache*». This has been driven by the regular participation of about 10,000 informants who have taken part in successive online questionnaires mainly dealing with the lexicon, but also with grammar. Since the questions are the same as in J. Eichhoff's *Wortatlas der deutschen Umgangssprachen*, which started in 1977, the results enable a comparison over time.

Mathieu Avanzi and André Thibault («Cartographier l'amuïssement et la restitution des consonnes finales en français grâce à la production participative») are doing the same thing, but for the French-speaking European countries, as well as Canada. More than 15 surveys have been conducted so far, reaching several tens of thousands of speakers. Initial results have been available since 2016 in the form of a blog (*Français de nos régions*) as well as in three published volumes. More recently, they've started extending their surveys to overseas French-speaking islands, and the final goal is to cover the whole French-speaking world. In this contribution, they show what their data can bring to the question of the areal distribution and evolution of final consonant pronunciation in European French-speaking countries, by combining regional and generational figures.

Philippe Boula de Mareüil, together with a binational team, oversees a twofold project in «For a mapping of the languages/dialects of Italy and regional varieties of Italian». The first phase deals with documenting Italy's dialectal varieties and regional languages through the elicitation of speech samples that can be heard and downloaded online, along with transcripts. The second phase focuses specifically on the study of regional Italian through crowdsourcing, based on phonological oppositions, which can be visualized using a dedicated program (Cartopho).

2. Specific studies based on new web technologies

Not all projects are aimed at mapping long lists of diversified linguistic features. New technologies can also be useful in the study of more targeted language issues.

Miriam Bouzouita, Mónica Castillo Lluch and Enrique Pato, for instance, introduce a new project («*Dialectos del español* : une application pour l'étude de la variation linguistique dans le monde hispanophone») specifically aimed at studying Spanish syntactic variation. While this might seem a narrow focus, the geographic span is very broad: the survey is being conducted across the entire Spanish-speaking world. Technically, it is managed by a phone app that has already proved to be quite effective with Swiss German and British English dialects.

Which brings us to David Britain, Tamsin Blaxter and Adrian Leemann's contribution, «Dialect levelling in England: Evidence from the *English Dialects App*». These authors show, thanks to data collected from more than 50,000 participants all over England, that certain features (mostly lexical) have undergone considerable levelling in the last fifty years, while others (mainly phonological) resist this tendency, especially in areas where mobility is weaker.

The last article deals with natural language processing for regional languages in France. Delphine Bernhard, in «Corpus annoté en parties du discours pour les dialectes alsaciens : comparaison avec l'allemand standard et le français», shows how a team of specialists working on Alsatian – as well as Occitan and Picard – have managed to help keep these minority languages alive on the Net, by developing programs that allow the grammatical tagging of their constituents.

Sorbonne Université

André THIBAULT
Mathieu AVANZI
Nicholas Lo VECCHIO
Alice MILLOUR

Dialectos del español : une application pour l'étude de la variation linguistique dans le monde hispanophone

1. Introduction

La méthode traditionnelle de collecte de données en dialectologie, basée sur des enquêtes *in situ* auprès de locuteurs de différentes enclaves de la zone explorée, malgré le fait qu'elle implique des coûts élevés en temps, en ressources humaines et financières, a été privilégiée jusqu'à présent car on a considéré que les enquêtes par correspondance pratiquées à un stade préscientifique de cette discipline (entre la fin du XVIII[e] et la fin du XIX[e] siècle) ne fournissaient pas de données d'une fiabilité comparable (cf. Águila Escobar 2012 : 121 et Bergounioux 1992 : 8-10). Cela dit, les recherches modernes obtenues à partir d'enquêtes à distance ne manquent pas lorsqu'il s'agit de projets embrassant de très vastes territoires, comme c'est le cas, par exemple, de *El léxico del español de América*, publié par Manuel Alvar en 1966, ou encore de *The Atlas of North American English*, élaboré par Labov / Ash / Boberg (2006) avec des données recueillies par téléphone dans les années 1990. Au cours de ce nouveau siècle ont d'ailleurs surgi de nombreuses études variationnistes avec un soutien empirique dans les réseaux sociaux (Twitter et Facebook surtout), les blogs, les forums et autres sources écrites et orales disponibles sur Internet. Dans ce cadre des humanités numériques, de nouveaux outils d'étude dialectologique ont vu le jour, tels que ceux présentés dans ce volume, y compris des applications spécialement conçues pour être utilisées sur des smartphones (cf. les pionnières *Dialäkt Äpp* et *Voice Äpp*, Kolly / Leemann 2015 et Leemann 2021, ainsi que *English Dialects App*, Leemann / Kolly / Britain 2018 ; voir le numéro spécial de *Linguistics Vanguard* édité par Hilton / Leemann 2021 et leur introduction pour en découvrir d'autres).

Cet article présente une nouvelle application pour smartphone destinée à l'étude dialectologique de la langue espagnole : *Dialectos del español*, mise en service en mai 2019 et disponible sur www.dialectosdelespanol.org et Google Play. Sont exposés en premier lieu (§ 1) les aspects scientifiques et techniques du projet : comment il a vu le jour, ses objectifs, les 26 questions de l'application, l'encodage des réponses et le fonctionnement de la prédiction ; deuxièmement (§ 2) on fournit des informations à propos de la diffusion médiatique et de la participation du public ; quelques exemples des données obtenues, qui permettent d'entrevoir le potentiel de l'outil, sont enfin présentés ainsi que les défis actuels du projet (§ 3). Les conclusions résument les contenus de ces pages.

2. Origine du projet, objectifs et méthodologie

Lors du 21e Symposium de Sociolinguistique de la *Linguistic Society of America*, qui s'est tenu à l'Université de Murcia en 2016, nous avons eu l'occasion de découvrir deux applications : *Dialäkt Äpp*, développée pour l'étude des variétés suisses alémaniques, et *English Dialects App*, pour l'étude des variétés anglaises britanniques, présentées respectivement par Adrian Leemann et David Britain, alors en poste aux universités de Zurich et de Berne (cf. Leemann *et al.* 2016 et Britain 2016). Ces applications étaient basées sur une série de questions phonétiques et lexicales, les composantes de la langue les plus traditionnellement étudiées, et leur principal objectif était de recueillir du *big data* pour faciliter la connaissance de l'évolution de certaines variables de ces langues étudiées auparavant par d'autres projets de dialectologie.

Les auteurs de ces lignes ont immédiatement partagé l'intérêt de développer une application similaire pour étudier la variation grammaticale de l'espagnol, l'aspect le moins exploré par la dialectologie hispanique, aspirant, par ailleurs, à toucher à l'ensemble du monde hispanophone. Pour commencer, nous avons mis en place une étude pilote qui nous a permis de tester un ensemble de questions au moyen d'un formulaire web (Google Forms) distribué à environ 500 participants dans plusieurs pays hispanophones. Dans Bouzouita / Castillo / Pato (2018), nous avons exposé les objectifs et la méthodologie de l'application qui était au stade de projet à ce moment-là et nous avons présenté les résultats obtenus dans le cadre de cette étude préliminaire.

Moyennant cette première expérience et l'intervention technique de Daniel Wanitsch – le même informaticien qui avait conçu les applications de l'équipe d'Adrian Leemann et David Britain[1], nous avons développé l'application *Dialectos del español* disponible sur www.dialectosdelespanol.org et sur Google Play[2].

Notre objectif principal est d'obtenir des données qui permettent d'étudier des phénomènes de variation morphosyntaxique de l'espagnol dans l'ensemble du monde hispanophone[3]. Certains de ces phénomènes ont déjà fait l'objet d'une description auparavant, et dans ce cas il s'agira de vérifier si cet état de la question correspond bien aux usages actuels. D'autres phénomènes n'ont été décrits que partiellement, et dans ce cas-là l'enjeu consistera à compléter la description et l'analyse. Enfin, quelques-unes des variables grammaticales de notre liste n'ont jamais fait l'objet d'une étude et nous souhaitons les explorer pour la première fois.

---

[1] Nous tenons à remercier Adrian Leemann pour sa disponibilité et son attitude tout à fait ouverte pour partager avec nous des informations précieuses pour développer notre projet.

[2] Les données recueillies sont stockées sur un serveur de l'Université de Lausanne, ce qui garantit la sécurité et la pérennité de leur conservation.

[3] Les questions posées aux participants et les réponses prévues pour la prédiction de leur géolocalisation tiennent compte des variantes constatées en Europe et en Amérique, mais nous souhaiterions aussi compter sur une participation d'hispanophones d'Afrique et d'Asie.

Voici ces trois catégories de questions avec les variables objet de notre étude :

(1) Questions classiques (déjà abordées par la dialectologie traditionnelle)
- Diminutifs
- *La / el sartén*
- *A Pedro le / lo vi*
- *A los jugadores les / los ves / Ø ves*
- *El libro lo / le tengo*
- *Había(n) muchos estudiantes*
- *Habíamos / estábamos / éramos*
- *¿Qué dices / qué tú dices / qué decís?*
- *A ti / vos te gusta el invierno*
- *Bailas / Bailás / Bailái(s)*
- *Fue en Bogotá donde / que se conocieron…*

(2) Questions moins classiques (on dispose de certaines informations, mais pas de résultats pour l'ensemble du monde hispanophone)
- *Comprásemos / compráramos*
- *Tuviese / tuviera / tendría*
- *Manuel está enfrente de mí / mío / mía*
- *No sé si vendrá / venga*
- *Isa me dijo ayer que viniéramos / vengamos / de venir hoy*
- *Nada más / más nada*
- *Yo ya / Ya yo había salido*

(3) Questions inédites
- *Esta / este agua*
- *Van a venir / vendrán hoy*
- *No la vamos a reconocer / no vamos a reconocerla*
- *Cuanto / contra / mientras más…*
- *Habla mal de mí / mío / mía*

Le principe de ce type d'application consiste à solliciter la collaboration du public en échange d'une récompense ludique : d'après les réponses, l'application tente de deviner l'origine dialectale des participants. De ce fait, un deuxième objectif de notre projet consiste à identifier correctement les dialectes des participants à partir de leurs réponses. Le défi n'est pas banal dans la mesure où bon nombre des variables de notre étude ne permettent pas du tout de géolocaliser les locuteurs (la plupart des

questions « inédites » et « moins classiques ») et quand elles permettent la géolocalisation, celle-ci n'est que trop vague (les diminutifs, *sartén* masculin, le *leísmo*, *haber* accordé avec un objet pluriel, le *voseo*, *más nada*, *ya yo*…). Les réponses grammaticales qui pointent vers un pays américain ou une aire de la péninsule sont finalement très exceptionnelles : *(vos) bailái(s) conmigo* (Chili, Vénézuela), *Se conocieron en Bogotá fue* (République Dominicaine), *enfrente mía* (Galice, Andalousie, Madrid).

Ce contrat passé avec le public, affiché sur la page d'accueil de l'app (« Dime cómo hablas y te diré de dónde eres » : « Dis-moi comment tu parles et je te dirai d'où tu viens »), nous a contraints à inclure quelques questions de lexique qui orientent de façon bien plus fine vers un pays et même parfois vers une région précise. Ces questions lexicales sont au nombre de trois : la formule utilisée pour répondre au téléphone et les mots équivalents à *petit pois* et *lézard* (à partir d'une image du légume et de l'animal). Par ailleurs, les diverses formes du diminutif proposées avec une image d'un petit chien peuvent aussi être considérées comme relevant du lexique dialectal.

Derrière la couche visible des 26 questions, l'app contient un encodage plus ou moins précis des réponses, servant à la prédiction du dialecte du participant. Voici trois exemples concrets de cet encodage :

(1) … agua es potable.
   (a) Esta…
   (b) Este…
   (c) Esta / Este…

(2) ¿La sartén o el sartén?
   (a) La sartén [ESPAÑA]
   (b) El sartén [AMÉRICA]
   (c) La / el sartén [AMÉRICA]

(3) Manuel está enfrente…
   (a) de mí. [AMÉRICA: México; AMÉRICA: Guatemala; AMÉRICA: Honduras; AMÉRICA: Nicaragua; AMÉRICA: El Salvador; AMÉRICA: Costa Rica; AMÉRICA: Panamá; AMÉRICA: República Dominicana; AMÉRICA: Puerto Rico; AMÉRICA: Venezuela; AMÉRICA: Colombia; AMÉRICA: Estados Unidos; ESPAÑA: Aragón; ESPAÑA: Islas Baleares; ESPAÑA: Asturias; ESPAÑA: Cantabria; ESPAÑA: Castilla y León; ESPAÑA: Extremadura; ESPAÑA: La Rioja; ESPAÑA: Murcia; ESPAÑA: Navarra; ESPAÑA: País Vasco; ESPAÑA: Castilla La Mancha]

(b) mío. [ESPAÑA: Galicia; ESPAÑA: Andalucía; ESPAÑA: Madrid; AMÉRICA: Argentina; AMÉRICA: Uruguay; AMÉRICA: Chile; AMÉRICA: Ecuador; AMÉRICA: Perú; AMÉRICA: Bolivia; AMÉRICA: Paraguay; ESPAÑA: Cataluña; ESPAÑA: Comunidad Valenciana; AMÉRICA: Panamá]

(c) mía. [ESPAÑA: Galicia; ESPAÑA: Andalucía; ESPAÑA: Madrid]

(d) de mí / mío. [ESPAÑA: Galicia; ESPAÑA: Andalucía; ESPAÑA: Madrid; AMÉRICA: Argentina; AMÉRICA: Uruguay; AMÉRICA: Chile; AMÉRICA: Ecuador; AMÉRICA: Perú; AMÉRICA: Bolivia; AMÉRICA: Paraguay; ESPAÑA: Cataluña; ESPAÑA: Comunidad Valenciana; AMÉRICA: Panamá; AMÉRICA: Costa Rica]

(e) de mí / mía. [ESPAÑA: Galicia; ESPAÑA: Andalucía; ESPAÑA: Madrid]

(f) mío / mía. [ESPAÑA: Galicia; ESPAÑA: Andalucía; ESPAÑA: Madrid]

(g) (de mí / mío / mía. [ESPAÑA: Galicia; ESPAÑA: Andalucía; ESPAÑA: Madrid]

Comme on peut le voir, l'encodage est parfois impossible (exemple 1), car aucune étude préalable ne nous renseigne sur la diffusion des formes féminine ou masculine du démonstratif devant les noms commençant par /á/ ; dans d'autres cas, il est très vague (exemple 2)[4], et ce n'est que lorsque l'on dispose d'informations très précises sur un phénomène (exemple 3, cf. Salgado / Bouzouita 2017 et Marttinen Larsson / Bouzouita 2018) qu'il peut vraiment contribuer à la prédiction du dialecte des participants.

Le principe de fonctionnement interne de l'app pour cette prédiction est simple : l'addition de points par pays américain ou par province d'Espagne, de telle sorte que, par exemple, si un participant répond avec les réponses encodées « [AMÉRICA] » et celles qui suivent, il sera identifié comme Colombien :

- *Enfrente de mí*
- *Juan habla mal de mí*
- *Sí, lo vi ayer*
- *Los veo [a los jugadores]*
- *El libro lo tengo en casa*
- *Arvejas*
- *A ver*
- *No, se conocieron fue en Bogotá*
- *Yo ya / ya yo había salido*
- *A ti / vos te gusta el invierno*
- *Qué dices / Qué decí(s) (tú / vos)*

[4] Nous avons délibérément simplifié dans ce cas et omis de considérer que dans certains dialectes méridionaux européens (andalou, canarien) *sartén* peut aussi avoir le genre masculin, car cette forme n'est généralement pas utilisée par les usagers majoritaires de l'application, de niveau universitaire (cf. *infra*). Cependant, la combinaison de ce trait avec d'autres dans le questionnaire peut prédire que l'utilisateur de l'app est bel et bien Andalou ou Canarien.

Nous avons constaté que le langage additif de l'app était parfois insuffisant et pouvait donner lieu à de grossières erreurs de géolocalisation, ce qui nous a menés à demander à notre informaticien d'ajouter lors du calcul des points quelques conditions éliminatoires. Ces conditions, implémentées pour la première fois dans ce type d'application, sont au nombre de 3 : 1) si le *voseo* pronominal ou verbal est employé, l'Espagne est exclue de la prédiction ; 2) la non inversion du pronom sujet dans les interrogatives (*¿Qué tú dices?*) exclut l'Espagne ; 3) *ser* focalisateur exclut l'Espagne.

À la suite du questionnaire et de notre prédiction de l'origine dialectale du participant (3 géolocalisations offertes par ordre de vraisemblance), nous sollicitons de sa part quelques données extralinguistiques : son genre (homme / femme / autre), son âge (-18 / 18-35 / 36-55 / +55), son niveau d'études (primaires, secondaires, universitaires), s'il est locuteur natif ou non (et dans ce cas, quelle est sa langue maternelle) et s'il est locuteur natif d'une autre langue en plus de l'espagnol. Enfin, nous requérons l'aide du participant pour améliorer l'application en lui demandant de situer sur une carte son lieu de naissance, celui où il a grandi et celui où il habite actuellement. Nous lui demandons également combien de fois il a changé de ville (jamais / 1 fois / 2-3 fois / plus de 3 fois) et de pays (*idem*) tout au long de sa vie. Il peut aussi évaluer au moyen de 5 étoiles (en 10 tranches) la précision de la prédiction de l'application, peut laisser des commentaires et partager l'app sur Facebook et Twitter. Pour nos deux objectifs (obtention de données dialectales et amélioration de la prédiction de l'application), logiquement, seuls les questionnaires accompagnés des données cartographiques et sociolinguistiques des participants sont utiles. Ces données nous permettent de lier des réponses à des localisations et de reproduire ainsi le fonctionnement des questionnaires de la dialectologie classique.

3. Diffusion médiatique et participation du public

Le succès de tout projet de *crowdsourcing* (production participative) dépend étroitement de la publicité qu'on peut lui donner. *Dialectos del español* a eu la chance de bénéficier d'une diffusion médiatique exceptionnelle dès son lancement en mai 2019, qui lui a valu un record de participation en très peu de temps. En effet, dès l'annonce faite par Lola Pons (2019) dans un article (« Todos hablamos dialecto y no una lengua ») paru dans *Verne – El País* (28.5.2019), les informations dans la presse et la radio espagnoles se sont multipliées et la participation du public a été déclenchée sans commune mesure par rapport à celle des autres applications similaires existantes. En un mois, *Dialectos del español* avait obtenu plus de 300.000 questionnaires remplis, alors que *Dialäkt Äpp* et *English Dialects App* avaient recueilli respectivement 98.000 et 100.000 questionnaires la première année[5]. C'était pour nous sans doute une excellente nouvelle, car le but d'une app comme la nôtre est de recueillir un maximum de *big data*. Il faut toutefois signaler que sur les 302.497 questionnaires du premier mois, 51,3 % (155.191) étaient inutilisables pour nos recherches, notamment en

[5] Cette information nous a été transmise personnellement par D. Wanitsch en juillet 2019.

raison de l'absence de l'information finale fournie par le participant à propos de son lieu de naissance, de vie pendant son enfance-adolescence et de résidence actuelle[6]. Ces questionnaires non géolocalisés ou de toute évidence géolocalisés de façon erronée (points en pleine mer, par exemple) ont dû être éliminés, en partie par nos soins, ce qui a représenté un nombre important d'heures de travail.

Par ailleurs, si les médias espagnols ont très vite amplifié l'annonce parue dans *El País* (entre autres *Heraldo de Aragón* le 29.5.2019, *El Confidencial* le 30.5.2019 et *ABC* le 31.5.2019 en format papier et numérique ; *Rioja 2* le 30.5.2019, *Aragón Radio* le 1.6.2019, *Cadena Ser* le 3.6.2019 et *La Cope* le 20.6.2019 sur les ondes)[7], nous n'avons pour l'instant pas eu le même succès auprès des médias américains (sauf *Radio Canada International* le 19.07.2019). Malgré les contacts entrepris avec des collègues et des médias outre-Atlantique, les réactions se font encore attendre et ce n'est que plus récemment (novembre 2020) que nous avons obtenu une notice sur BBC Mundo (Hernández Velasco 2020), dont le public cible est majoritairement hispano-américain. De ce fait, le nombre de participants américains représente une part encore limitée des données valides recueillies à ce jour[8]. Cela nous mène à réfléchir à des stratégies de communication spécifiques pour toucher plus de participants américains lors d'une prochaine étape de notre projet.

Concernant le genre des participants, on peut reporter à partir des 125.715 questionnaires valides du premier mois en Espagne que la proportion de participants masculins est très légèrement supérieure à celle des femmes (respectivement 67.153 – 53,4 % et 58.103 – 46,2 %) et que l'option « autre » apparaît également (459 – 0,4 %). L'âge le plus représenté dans tous les pays est 19-35, suivi de 36-55 ; par exemple, en Espagne, 77.941 (62 %) ont entre 19 et 35 et 34.545 (27,5 %) entre 36 et 55. Sans surprise, pour ce qui est du niveau d'études, on retrouve les mêmes tendances observées dans des projets similaires (cf. Leemann / Kolly / Britain 2018) : ce sont majoritairement des universitaires qui répondent au questionnaire (79,3 % des questionnaires valides recueillis le premier mois pour tous les pays – 113.842 / 143.563) et les participants du niveau primaire ne sont que très exceptionnels (1,3 % – 1.908 / 143.563).

Pour ce qui est de la distribution géographique des participants, en Espagne elle est généralisée et dans toutes les provinces l'échantillon est quantitativement repré-

---

6 Selon D. Wanitsch (communication personnelle en juillet 2019), le taux de 48,7 % (147.306 / 302.497) de questionnaires valides est tout de même élevé par rapport à d'autres applications qu'il a développées.

7 En plus, nous avons créé un compte Facebook et un autre Twitter pour interagir avec le public : d'une part, nous y publions la revue de presse de l'app et d'autres informations concernant les études dialectologiques et, d'autre part, nous répondons aux questions qui nous sont posées et exprimons notre gratitude envers les participants qui partagent leur résultat de l'app par le moyen conventionnel de ce réseau, en leur envoyant un petit cœur.

8 Ils étaient moins de 15 % du total des 302.497 questionnaires du premier mois, mais l'app est à nouveau devenue virale en novembre 2020 (après la parution de la notice de BBC Mundo) et il est très probable que la participation américaine y ait contribué (nous disposerons de ces données très prochainement).

sentatif par rapport à l'ensemble de la population, car il dépasse de loin le chiffre de 0,025 % – c'est-à-dire 25 locuteurs pour cent mille habitants – établi par Labov (1966, 170-171). En Amérique, le nombre de participants n'atteint pas encore ce seuil de représentativité sauf pour le Costa Rica, ce qui nous met face au défi de mener à l'avenir une campagne plus ciblée dans ce continent afin d'obtenir un volume suffisant de données permettant une description et une analyse sur des bases quantitatives convenables.

Dialectos del español compte à ce jour (février 2021) plus de 615.000 participations, ce qui est une preuve du grand intérêt que la société démontre envers la variation géographique de la langue espagnole et de l'excellent accueil qu'elle réserve à des outils de vulgarisation des connaissances sur le sujet.

4. Résultats préliminaires et défis du projet

Nous entamons actuellement la phase du traitement et analyse des données recueillies avec l'application *Dialectos del español* mais nous disposons déjà de quelques aperçus des résultats dont nous exposerons ici trois exemples (à partir des données extraites des premiers 143.563 questionnaires valides)[9].

Une question comme celle de « comment répondez-vous au téléphone (lorsque vous ne savez pas qui est au bout du fil) » permet d'entrevoir la variation interne aux différents pays, invisible, par exemple, dans le DVD *Las voces del español: tiempo y espacio* (RAE / ASALE 2011) qui inclut un enregistrement audio avec une seule formule par pays hispanophone. La carte suivante montre comment parmi les réponses *¿Qué hay?*, *¿Sí?*, *A ver*, *Aló*, *Bueno*, *Diga / dígame*, *Hola*, plusieurs sont documentées et alternent dans le même pays. Cette variation a été décrite dans la *Nueva gramática de la lengua española* (RAE / ASALE 2009, 2508) mais sans indication des fréquences et avec certaines informations désuètes (par exemple, pour l'Espagne *díga / dígame* est décrite comme l'expression par défaut, alors que *sí*, comme nous le voyons, est de nos jours majoritaire ; *a ver* est indiquée comme formule propre à la Colombie, or elle n'apparait pas pour l'instant dans les données de l'app relatives à ce pays).

[9] Le travail d'extraction des données a pu se faire grâce au soutien financier des Universités de Gand et de Montréal.

Figure 1 : réponse au téléphone

Une étude menée par Salgado / Bouzouita (2017) qui abordait l'usage des possessifs avec des adverbes dans des locutions locatives en espagnol péninsulaire constatait l'insuffisance de données dans les corpus de référence pour pratiquement tout le nord de l'Espagne. Grâce à la question « Manuel está *enfrente… de mí / mío / mía…* » de l'application, nous disposons maintenant d'un grand volume de données sur la construction avec cet adverbe qui nous font découvrir, premièrement, que la variante standard « Manuel está enfrente de mí » est minoritaire dans la pratique en Espagne (44 % des cas en moyenne). Un deuxième constat est que, toujours en moyenne, le complément possessif est plus souvent féminin (*enfrente mía*, 30,4 %) que masculin (*enfrente mío*, 25,6 %). Les chiffres les plus élevés de *enfrente mía* se trouvent d'ailleurs dans le nord-ouest de l'Espagne, en Galice (35 %), alors que *enfrente mío* atteint en Catalogne presque 50 % des réponses. Le morphème {-a} semble donc être un trait occidental propre au nord de l'Espagne – en plus de l'Andalousie.

En parallèle de ces structures adverbiales possessives, on a récemment aussi documenté l'utilisation d'un complément possessif dans le syntagme verbal : *él habla suyo*, *ella depende mío*, *tú gustas mío* (Casanova 2020, Bertolotti 2017). Ce phénomène récent et propre à la langue parlée spontanée, encore une fois, n'apparaît pas dans les corpus linguistiques conventionnels et exige de la part des chercheurs le recours à des sources alternatives d'information qui contiennent des usages non normatifs (blogs, forums, réseaux sociaux comme Twitter). Dans ce cas-là, les données fournies par *Dialectos del español* révèlent que le complément possessif au lieu du régime prépositionnel habituel du verbe *hablar* est extrêmement minoritaire en Espagne (seulement 2,8 % des cas) face à la variante standard (« Juan habla mal *de mí* » 97,2 %) et que c'est la variante avec possessif féminin (« Juan habla mal *mía* ») qui domine légèrement sur

celle avec possessif masculin (1,6 % contre 1,2 %). C'est une fois de plus en Galice que l'on retrouve le pourcentage le plus élevé de ces structures verbales avec complément possessif – 4,5 % en moyenne – qui atteint presque 7 % dans certaines provinces (La Coruña) et toujours avec une préférence pour le possessif féminin.

Ces derniers exemples montrent comment une application comme *Dialectos del español* peut être un outil d'une grande puissance pour sonder de manière ciblée des variantes grammaticales non-standards généralement peu ou pas du tout documentées dans les corpus de référence pour la langue espagnole. La profondeur de ce sondage est sans commune mesure par rapport à celle que nous permettent les méthodes traditionnelles de recueil de données en dialectologie et peut avoir une portée plus large que celle de la structure dont il s'agit : certes avec la question *enfrente… de mí / mío / mía* nous obtenons des données avec une granularité très fine pour l'utilisation du possessif dans cette locution avec cet adverbe précis et non pas avec d'autres (*detrás*, *delante*, *encima*…), mais cela nous renseigne déjà de façon très détaillée sur un cas de figure du paradigme qui peut servir à éclairer des aspects intéressants à explorer par la suite dans l'ensemble de celui-ci avec une recherche plus vaste.

L'avantage essentiel de l'outil est sans doute sa puissance pour procurer en peu de temps des données massives dont on ne disposait pas auparavant pour une forme ou une structure linguistique précise, mais une autre de ses qualités réside dans le fait de pouvoir l'adapter à des recherches sur de possibles corrélations de phénomènes. Un cas précis dans le cadre de notre application est celui des compléments possessifs dans le domaine adverbial et verbal que nous venons d'exposer et que nous traiterons de manière conjointe lors de notre analyse. Le croisement des variables externes (outre les coordonnés spatiales, l'âge, le genre, la mobilité au long de la vie et le niveau d'études) avec les diverses réponses aux variables linguistiques pourra, par ailleurs, fournir des informations essentielles pour analyser les usages et les changements linguistiques en cours.

Idéalement, si le temps et les ressources le permettent, une application comme celle-ci nous semble être une méthode de recueil de données qu'il convient d'utiliser de façon complémentaire à d'autres plus classiques d'observation directe des usages spontanés, car il ne faut pas perdre de vue que les données recueillies avec *Dialectos del español* ou toute autre application de ce type correspondent à des usages déclarés et non pas réels. Cependant, s'il n'est pas exclu que les participants ne répondent pas à certaines questions en reflétant plutôt leur connaissance des usages normatifs que leur façon réelle de parler, le contexte de production des réponses anonyme et stimulé par la curiosité de vérifier dans quelle mesure ils peuvent obtenir de la web-app une prédiction juste de leur dialecte, les prédisposerait à répondre de façon plutôt réaliste, en accord avec la consigne initiale « te pedimos que reflejes tu modo real de hablar ». En revanche, le fait que l'espace réservé aux réponses non prévues soit très restreint peut parfois constituer une vraie limite de l'application et avoir une incidence sur la qualité des données (le locuteur qui ne trouve pas sa forme dans la liste des réponses répondra de la façon la plus proche, mais l'information ne correspondra

pas à un usage réel). Nous avons néanmoins pris cette décision de ne pas permettre les réponses ouvertes, car cela exige un traitement des données plus sophistiqué et plus lourd que nous souhaitions éviter lors de cette première version de l'application.

Plusieurs défis se présentent à notre équipe à l'heure actuelle. Comme nous l'avons déjà indiqué plus haut, nous espérons encore pouvoir obtenir avec cette version de l'application suffisamment de données en Amérique (un échantillon s'élevant au moins à 0,025 % de la population de chaque pays) permettant la comparaison avec les données espagnoles. Cela exigera de redoubler d'efforts pour diffuser une information plus efficace sur l'application outre-Atlantique. Par la même occasion, nous envisageons d'ajuster quelques détails, concernant notamment l'encodage des réponses, afin d'améliorer la performance de la prédiction. Ces améliorations concerneront le contenu – grâce au feedback (commentaires explicites) obtenu de la part de certains utilisateurs et aux réponses déjà enregistrées dans notre base de données nous pourrons mieux fixer la géolocalisation de certaines réponses –, mais aussi la forme de notre code – en ajoutant certaines conditions éliminatoires nous éviterons des prédictions erronées.

Par ailleurs, dans un délai assez court (2022) nous projetons de publier une monographie sur la variation grammaticale en espagnol européen à partir des données recueillies en Espagne (Bouzouita / Castillo / Pato en cours), suivie à plus longue échéance par une publication qui fasse le point sur les résultats en Amérique. Mais pour ce continent nous envisageons, à l'avenir, de développer une version spécifique qui nous permette d'explorer la variation grammaticale qui lui est propre de manière plus fine et qui soit plus performante en ce qui concerne la prédiction (actuellement nous arrivons à une prédiction régionale pour l'Espagne et uniquement nationale pour l'Amérique). À ce moment-là, il conviendra de concevoir un langage informatique plus sophistiqué que celui de *Dialectos del español*, qui est pour l'heure fondamentalement plat. Il s'agira, dans l'idéal, de produire une application avec un design arborescent, permettant de géolocaliser progressivement les participants en cours d'usage de l'application et de continuer avec des questions à chaque pas plus ciblées concernant la variété détectée.

5. Conclusions

Les pages qui précèdent présentent la première application pour smartphone développée afin d'étudier la variation dialectale de la langue espagnole. *Dialectos del español*, conçue sur le modèle des applications *Dialäkt Äpp* et *English Dialects App* (servant respectivement à l'étude des variétés suisses alémaniques et de l'anglais britannique), vise la variation grammaticale de l'espagnol européen et américain. Par le biais de 26 questions, cette application recueille des données dialectales sur des variables grammaticales à propos desquelles nous disposons de plus ou moins de connaissances à ce jour et offre en échange aux participants une prédiction de leur géolocalisation linguistique.

Cet article expose l'origine, les objectifs et les contenus de l'application, son fonctionnement technique, la façon dont elle a été diffusée, l'accueil que le public lui a réservé et quelques exemples des données obtenues. Sont également évoquées les limites de l'outil et les modifications envisagées pour améliorer cette app ainsi que pour en produire une future version plus ciblée pour l'Amérique. Enfin, nous nous référons aux projets de publications prévues avec les données massives obtenues avec *Dialectos del español* grâce à une très large participation du public, laquelle prouve l'intérêt et la curiosité dont les hispanophones témoignent envers la variation dialectale de leur langue.

Humboldt-Universität zu Berlin — Miriam BOUZOUITA
Université de Lausanne — Mónica CASTILLO LLUCH
Université de Montréal — Enrique PATO

Références bibliographiques

Águila Escobar, Gonzalo, 2012. « La encuesta dialectal como narración y el modo de preguntar en el ALEA », *Letral* 8, 118-137.

Alvar, Manuel, 1966. *El léxico del español de América*, Granada, CSIC.

Bergounioux, Gabriel, 1992. « Les enquêtes de terrain en France », *Langue française* 93 (*Enquête, corpus et témoin*), 3-22.

Bertolotti, Virginia, 2017. « Pronombres posesivos y cambios gramaticales en español. Análisis en la variedad rioplatense », in : Company Company, Concepción / Huerta Flores, Norohella (ed.), *La posesión en la lengua española*, Madrid, Consejo Superior de Investigaciones Científicas, 325-349.

Bouzouita, Miriam / Castillo Lluch, Mónica / Pato, Enrique, 2018. « *Dialectos del español :* una nueva aplicación para conocer la variación actual y el cambio en las variedades del español », *Dialectologia* 20, 63-85.

Bouzouita, Miriam / Castillo Lluch, Mónica / Pato, Enrique, 2019. Web-app *Dialectos del español* www.dialectosdelespanol.org.

Bouzouita, Miriam / Castillo Lluch, Mónica / Pato, Enrique, en cours. *La variación gramatical en el español europeo, hoy. Datos de* Dialectos del español.

Britain, David, 2016. « *Up, app and away? :* Social dialectology and the use of smartphone technology as a data collection strategy », conférence prononcée lors du 21e Sociolinguistics Symposium (Universidad de Murcia, 16.07.2016).

Casanova, Vanessa, 2020. « El uso del complemento posesivo verbal por el complemento de régimen preposicional en español actual », in : Bouzouita, Miriam / Marttinen Larsson, Matti (ed.), *Possessive Constructions in Romance*, *Moderna språk*, 115.

Hilton, Nanna Haug / Leemann, Adrian, 2021. « Editorial : using smartphones to collect linguistic data », *Linguistics Vanguard* 7.s1 ‹https://doi.org/10.1515/lingvan-2020-0132›.

Hilton, Nanna Haug / Leemann, Adrian (ed.), 2021. *Using Smartphones to Collect Data for Linguistic Research*, *Linguistics Vanguard* 7.s1 ‹https://www.degruyter.com/journal/key/LINGVAN/7/s1/html›.

Hernández Velasco, Irene, 2020. « Lengua, dialecto, geolecto y sociodialecto: ¿hay alguien que hable realmente español? », *BBC News / Mundo* 2.11.2020 ‹https://www.bbc.com/mundo/noticias-53864492›.

Kolly, Marie-José / Leemann, Adrian, 2015. « *Dialäkt Äpp :* Communicating dialectology to the public – crowdsourcing dialects from the public », in : Leemann, Adrian *et al.* (ed.), *Trends in Phonetics and Phonology. Studies from German-speaking Europe*, Bern, Peter Lang, 271-285.

Labov, William, 1966. *The Social Stratification of English in New York City*, Washington D.C., Centre for Applied Linguistics.

Labov, William / Ash, Sharon / Boberg, Charles, 2006. *The Atlas of North American English: Phonetics, Phonology and Sound Change*, Berlin / New York, Mouton de Gruyter.

Leemann, Adrian, 2021. « Apps for capturing language variation and change in German-speaking Europe: Opportunities, challenges, findings, and future directions », *Linguistics Vanguard* 7.s1 ‹https://doi.org/10.1515/lingvan-2019-0022›.

Leemann, Adrian *et al.*, 2016. « Crowdsourcing Big Data in dialectology – the case of Swiss German », conférence prononcée lors du 21e *Sociolinguistics Symposium* (Universidad de Murcia, 17.07.2016).

Leemann, Adrian / Kolly, Marie-José / Britain, David, 2018. « The English Dialects App: The creation of a crowdsourced dialect corpus », *Ampersand* 5, 1-17.

Marttinen Larsson, Matti / Bouzouita, Miriam, 2018. « *Encima de mí* vs. *encima mío*: un análisis variacionista de las construcciones adverbiales locativas con complementos preposicionales y posesivos en Twitter », *Moderna språk* 112.1, 1-39.

Pons, Lola, 2019. « Todos hablamos dialecto y no una lengua », *Verne – El País* 28.5.2019 ‹https://verne.elpais.com/verne/2019/05/21/articulo/1558424530_527443.html›.

RAE / ASALE, 2009. *Nueva gramática de la lengua española: Morfología y Sintaxis*, Madrid, Espasa.

RAE / ASALE, 2011. *Nueva gramática de la lengua española: Fonética y fonología. DVD Las voces del español: tiempo y espacio*, Barcelona, Espasa.

Salgado, Hugo / Bouzouita, Miriam, 2017. « El uso de las construcciones de adverbio locativo con pronombre posesivo en el español peninsular: un primer acercamiento diatópico », *Zeitschrift für romanische Philologie*, 133.3, 766-794.