International Conference | Colloque international[1]

New Ways of Analyzing Dialectal Variation |
Nouveaux regards sur la variation dialectale

Sorbonne Université (1 rue Victor Cousin, 75005 Paris)
November | Novembre 21-23, 2019

Organizing committee | Comité d'organisation

Mathieu Avanzi (STIH, Paris-Sorbonne), Alice Millour (STIH, Paris-Sorbonne)
& André Thibault (STIH, Paris-Sorbonne)

Aims of the conference

The conference "New Ways of Analyzing Dialectal Variation | Nouveaux regards sur la variation dialectale" aims to bring together dialectologists from various backgrounds in order to evaluate how new technologies can enhance our knowledge of dialectal variation in Europe. 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. Finally, it must 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.

Objectifs scientifiques du colloque

Le colloque international « New Ways of Analyzing Dialectal Variation | Nouveaux regards sur la variation dialectale » propose de réunir des spécialistes en vue de faire le point sur l'apport des nouvelles technologies à notre connaissance de la variation géographique des dialectes primaires et secondaires des langues d'Europe. Les progrès informatiques réalisés au cours des deux dernières décennies, de même que le développement du web 2.0, ont bouleversé notre façon de faire de la dialectologie. Sur le plan de la collecte des données, de nombreuses alternatives aux classiques enquêtes de terrain ont pu voir le jour : grâce à Internet et aux réseaux sociaux, il est désormais possible d'interroger les locuteurs sur leurs pratiques sans se déplacer ; les microphones de haute qualité, intégrés dans les téléphones portables, permettent même de récupérer leur voix à distance, de façon simultanée, même à des milliers de kilomètres. Sur le plan de l'analyse et de la visualisation des données, il est aujourd'hui nettement moins coûteux en temps de créer un atlas, de le mettre à la disposition de la communauté scientifique et d'en interpréter les données : les logiciels de visualisation de données et de statistiques permettent aujourd'hui des traitements qui n'auraient jamais été envisageables il y a de cela quelques décennies à peine. Enfin, il faut le souligner, ces progrès ont permis de porter un regard nouveau sur des matériaux récoltés par nos prédécesseurs, à des époques plus ou moins anciennes, et donc d'enrichir nos connaissances sur les changements linguistiques qui ont eu cours dans l'histoire des langues d'Europe.

[1] https://sites.google.com/view/nwadv2019

Program | programme (1/2)

21.11.2019 (THURSDAY | JEUDI) – ROOM | SALLE J-636	
14.00	**Yves Charles MORIN (Université de Montréal)** L'Atlas linguistique de la France informatisé (ALFi)
14.45	**Wim REMYSEN (Université de Sherbrooke)** Revisiter les données dialectologiques de la Société du parler français au Canada (1904-1906) : enjeux et perspectives
15.30	**Esther BAIWIR (Université de Lille)** Atlas linguistiques et analyse sémantique : le cas du projet APPI
16.15	*coffee break | pause-café*
16.45	**Alice MILLOUR &Karën FORT (Sorbonne Université)** Recettes de Grammaire : production participative de ressources linguistiques variées pour l'alsacien
17.30	**Fabio ARMAND (Université catholique de Lyon)** Du terrain au numérique : évolution du traitement des données de l'*Atlas linguistique et ethnographique du lyonnais*

22.11.2019 (FRIDAY MORNING | VENDREDI MATIN) – ROOM | SALLE J-636	
09.00	**Jack GRIEVE (University of Birmingham)** Word Frequencies as Dialect Features
09.45	**Dave BRITAIN (Universität Bern)** Going, going but not gone: Evidence about dialect levelling from the English Dialect App
10.30	*coffee break | pause-café*
11.00	**Mathieu AVANZI & André THIBAULT (Sorbonne Université)** Towards an automated geolocation of French speakers
11.45	**Philippe BOULA DE MAREÜIL (LIMSI-CNRS, Orsay)** Pour une cartographie des langues/dialectes d'Italie et des variétés régionales d'italien
12.30	*meal | repas (Le Cosi)*

Program | programme (2/2)

21.11.2019 (THURSDAY AFTERNOON | JEUDI APRÈS-MIDI) – ROOM | SALLE J-636	
14.00	**Stephan LÜCKE** (Ludwig-Maximilians-Universität München) VerbaAlpina. Digital Geolinguistics Dedicated to the Lexical Analysis of the Alpine Region
14.45	**Robert MÖLLER** (Université de Liège) An online atlas of colloquial German: The *Atlas zur deutschen Alltagssprache*
15.30	*pause-café / coffee break*
16.00	**Christoph PURSCHKE (Université du Luxembourg) & Dirk Hovy (Università di Milano)** Regional variation and the socio-pragmatics of online writing. A case study in the German-speaking area
16.45	**Adrian LEEMANN (Universität Bern)** Apps for mapping language variation and change in German-speaking Europe
17.30	**Yves SCHERRER (Université d'Helsinki)** Interactive dialect maps for German-speaking Switzerland and other European dialect areas

23.11.2019 (SATURDAY | SAMEDI) – ROOM | SALLE F-366	
09.00	**Delphine BERNHARD (Université de Strasbourg)** Natural Language Processing for Regional Languages of France: Lessons Learned from the RESTAURE Project
09.45	**Amélie DEPARIS (INALCO, Paris)** Étude des parlers du Croissant via la cartographie informatisée
10.30	*coffee break | pause-café*
11.15	**Xulio SOUSA (St-J. de Compostelle)** From field notebooks to the computer screen: the digital edition of the Atlas Lingüístico de la Península Ibérica
11.45	**Mónica CASTILLO LLUCH (Université de Lausanne)** Dialectos del español : une application pour l'étude de la variation morphosyntaxique dans le monde hispanophone

Abstracts | Résumés

*****

Du terrain au numérique : évolution du traitement des données de l'*Atlas linguistique et ethnographique du lyonnais*

Fabio Armand (Institut Pierre Gardette, Université Catholique de Lyon)

L'*Atlas linguistique et ethnographique du Lyonnais* (ALLy), l'un des premiers atlas linguistiques par région, a été réalisé au cours des années 1960-1976. Depuis 2015, une collaboration entre l'Institut Pierre Gardette (UCLy) et le laboratoire Dynamique du Langage (DDL – Lyon 2) a permis de mettre en œuvre un chantier de numérisation des cartes de cet atlas afin d'assurer la préservation des données linguistiques et les rendre accessibles à la communauté scientifique et au public intéressé par le patrimoine linguistique. La numérisation étant aujourd'hui terminée, nous présenterons les avancées les plus récentes du projet concernant principalement les questions de la transposition des données linguistiques – de l'alphabet phonétique Rousselot-Gilliéron, utilisé dans l'ALLy, à l'alphabet phonétique international (API-IPA) – et la mise en place d'un processus automatique de traitement, via des scripts en Python, se basant sur des techniques de reconnaissance de caractères. Des outils d'interrogation et d'exploitation des cartes traitées sont en cours de réalisation afin de pouvoir développer des analyses phonétiques et/ou lexicales des données linguistiques.

*****

Towards an automated geolocation of French speakers

Mathieu Avanzi & André Thibault (Sorbonne Université)

Until the 1970s, French dialectologists focused their attention on Gallo-Romance primary dialects, which was an urgent task due to the severe attrition that characterized their use throughout the 20th century. This might be one of the reasons why the study of regional variation in the standard language was rather neglected back then. During the last decades of the century—which witnessed the death of most 'patois'—scholars' attention finally shifted to regional French, and a number of valuable dictionaries based on philological criteria became available, with a historical and comparative perspective as well as rich textual documentation. Nevertheless, the representation of regional French diatopic differentiation in the form of maps, the way it used to be done with the traditional 'patois' in geolinguistic atlases, was still in its embryonic stage—mainly due to technical limitations.

In the past few years, the democratization of the web has enabled the successful creation and development of several platforms. First, there are websites that explicitly rely on the use of games with a specific purpose, such as ZombiLingo or Phrase Detective. There are also websites aimed at gathering dialectal material, such as Citizen Linguistics (in which the main author took part) or the VerbaAlpina project. Finally, there are websites that host online surveys whose aim is to map the area and test the vitality of a certain number of dialectal forms (words, expressions, pronunciations). Such websites exist for American English (Harvard Dialect Survey), German (Atlas zur deutschen Alltagssprache) and Italian (Atlas della Lingua Italiana QUOTidiana). As for French, we conducted more than 15 surveys that were launched on the Français de nos Régions website to gather information regarding French regional features around the world.

The material gathered through the Français de nos Régions surveys allowed us to obtain a great deal of relevant data to document the vitality and geographical extent of hundreds of linguistic features that are known to be specific to some regions of the French-speaking world. Aside from the well-known *chocolatine/pain au chocolat* debate, there are many objects (pencil, plastic bag, end of a piece of bread, pot of water at the canteen, mop, etc.) or activities (closing a door, kissing someone, intensifying a word in a sentence, choking on food, etc.) whose names vary across regions. In this talk, we will show how a selection of this material was used to train an algorithm that aims to predict French speakers' regional origins across the entire *Francophonie*.

*****

Atlas linguistiques et analyse sémantique : le cas du projet APPI

Esther Baiwir, ALITHILA (Université de Lille)

En terme d'enquêtes et d'atlas linguistiques, le domaine galloroman bénéficie sans doute de l'une des couvertures les plus denses qui soit. Depuis l'ALF, diverses enquêtes ont eu lieu, dans le cadre des atlas par région ou en-dehors. Quant à la valorisation des matériaux ainsi recueillis, elle va de la publication de volumes physiques à la numérisation des ressources, en passant par la conception de projets de seconde génération (l'on pense au projet Cartodialect ou au THESOC pour le domaine galloroman, à l'ALiR ou à VerbaAlpina dans un cadre plus large).

Dans ces travaux, la valeur sémantique des matériaux atlantographiques constitue souvent l'information la moins satisfaisante, celle-ci étant définie par un étiquetage au moyen de lexèmes de la langue-toit. Dans les marges apparaissent parfois des precisions sémantiques ou pragmatiques, mais par défaut, l'absence de marge doit être comprise comme la parfait équivalence des notions. Même lorsque celle-ci est relativement juste, un tel étiquetage ne représente cependant jamais une définition scientifique satisfaisante.

C'est ce dogme de l'équivalence que nous souhaitons dénoncer, en illustrant les dangers qu'il induit dans un projet de rassemblement des matériaux de plusieurs sources. Le projet ANR APPI (*Atlas pan-picard informatisé*) a pour ambition de rassembler les matériaux des trois atlas présentant des matériaux picards, l'ALF, l'ALPic et l'ALW. Nous montrerons pourquoi l'analyse de la valeur des matériaux constitue le préalable à l'établissement de la macrostructure de cet atlas, et pourquoi l'extension éventuelle de ce modèle aux parlers limitrophes ne pourra faire l'économie de cette étape.

*****

Natural Language Processing for Regional Languages of France: Lessons Learned from the RESTAURE Project

Delphine Bernhard (Université de Strasbourg)

The RESTAURE project[2] (2015-2018) aimed at providing computational resources and natural language processing (NLP) tools for three regional languages of France: Alsatian, Occitan and Picard. It brought together researchers from four research units located in Strasbourg (LiLPa), Toulouse (CLLE-ERSS), Amiens (Habiter le monde) and Orsay (LIMSI).

In this presentation, we will discuss the results of the RESTAURE project, focusing on the obstacles and challenges, the successful outcomes and the next steps. We will assess to what extent the project followed recent recommendations on improving digital language vitality for under-resourced and minority languages, to which Alsatian, Occitan and Picard belong [Soria *et al*., 2013, Ceberio Berger *et al*., 2018]. In particular, we will show how the cooperation between the research units involved made it possible to compensate, to some extent, for the lack of human resources and specialists for the regional languages under study. We will also explain how we re-used existing standards and proven methods in the process of resource building. Finally, this presentation will serve as an opportunity to detail the resources and tools developed during the project, which have been made available on the Zenodo platorm (https://zenodo.org/communities/restaure/) under a Creative Commons Attribution Share Alike 4.0 license.

References

Ceberio Berger, K., Gurrutxaga Heraiz, A., Baroni, P., Davyth, H., Kruse, E., Quochi, V., Russo, I., Salonen, T., Sarhimaa, A., and Soria, C. (2018). *Digital Language Survival Kit. The DLDP Recommendations to Improve Digital Vitality.* Technical report. http://www.dldp.eu/sites/default/files/documents/DLDP_Digital-Language-Survival-Kit.pdf.

Soria, C., Mariani, J., and Zoli, C. (2013). Dwarfs sitting on the giants' shoulders—how LTs for regional and minority languages can benefit from piggybacking major languages. In *Proceedings of XVII FEL Conference*, pages 73-79.

*****

[2] http://restaure.unistra.fr/, Project funded by the ANR, convention ANR-14-CE24- 0003.

Pour une cartographie des langues/dialectes d'Italie et des variétés régionales d'italien

Philippe Boula de Mareüil (LIMSI-CNRS)

Le but de cette communication est double : nous décrirons d'abord un atlas linguistique sonore qui prend la forme d'un site web présentant une carte d'Italie interactive, où l'on peut cliquer sur plus d'une centaine de points d'enquête pour écouter autant d'échantillons de parole et lire une transcription de ce qui est dit, en langues locales (ou régionales). Nous avons enregistré la fable d'Ésope « La bise et le soleil » (utilisée depuis un siècle par l'Association Phonétique Internationale pour illustrer nombre de langues du monde) dans une quinzaine de variétés italo-romanes, dans des variétés d'occitan, de francoprovençal et de catalan, en sarde, en frioulan, ladin, ainsi que dans des variétés non-romanes comme le grico, l'arbëresh et le walser. Les langues ou dialectes italoromans pour lesquels nous disposons d'au moins un enregistrement transcrit sont les suivants : piémontais, ligure, lombard, vénitien, émilien-romagnol, toscan, marchois, molisan, pouillais, salentin, calabrais, sicilien, lucanien, campanien, romanesco et gallurais. Les transcriptions orthographiques ont été fournies par les locuteurs eux-mêmes et homologuées par des linguistes.

Nous avons fait figurer, en plus des frontières de régions administratives, les limites entre domaines linguistiques, nous inspirant de la classification de Pellegrini [1], avec une signalétique particulière pour les langues non-romanes. En l'absence d'enregistrement dans des parlers centraux comme l'ombrien et le sabin, nous n'avons pas (encore) fait figurer ces étiquettes sur la carte. Nous avons pris des contacts pour combler les lacunes comme l'absence d'enregistrement en dialecte bavarois du Sud-Tyrol. Nous comptons lancer une campagne de mailing et espérons trouver le succès qu'a rencontré (avec plus d'un demi-millions de visiteurs) l'*Atlas des langues régionales de France* [2], auquel ce site est depuis peu associé.

Le second objectif du travail rapporté ici est de mettre sur pied une grande enquête de type crowdsourcing pour documenter et cartographier des variantes de prononciation en italien régional : une liste de 70 mots dont la prononciation dépend phonologiquement de la région a été établie et enregistrée auprès d'un acteur, avec à chaque fois deux possibilités (ex. *posto* avec le premier 'o' une fois ouvert, une fois fermé). Une interface en ligne a été développée, et une expérience a été lancée à travers des listes de diffusion et les réseaux sociaux. Plus de 600 personnes y ont pris part, de plus de 80 provinces italiennes (sur 109). Nous présenterons les premiers résultats, qui auront vocation à être cartographiés à l'instar du site cartopho [3] pour le français.

Références

[1] Pellegrini, G. (1977), *Carta dei dialetti d'Italia*, Pacini, Pisa.

[2] Boula de Mareüil, Ph., Vernier, F., Rilliard, A. (2017), « Enregistrements et transcriptions pour un atlas sonore des langues régionales de France », *Géolinguistique* 17, 23-48 (disponible à l'adresse https://atlas.limsi.fr).

[3] https://cartopho.limsi.fr

Going, going but not gone:
Evidence about dialect levelling from the English Dialect App

Dave Britain (Universität Bern)

Despite widespread claims of ongoing dialect levelling and supralocalisation in England, it is important to remember that:

a) These are ongoing processes, not *faits accomplis* – local dialect variation is still traceable;

b) Although levelling is driven by 'mobility' and by social and demographic change, such processes are socially diverse, engaged in more by some than others, and consequently its effects are uneven;

c) There has been relatively little empirical evidence of intra-regional levelling from multi-locality studies with spatial scope – many of the claims of levelling come from analyses of change within one or two sites, rather than comparative analyses across many sites

In this paper, I address a) and b) above, by engaging in c) – a large-scale, geographically-broad, multi-locality study of English accents and dialects. To do so, we developed the English Dialect App (EDA), a mobile application that, through a short quiz, asks people about their use of 26 different phonological, grammatical and lexical features, as well as collects spoken data from them (Leemann, Kolly & Britain 2018). Launched in January 2016 for iOS and Android, within one year over 50,000 people across England had completed the quiz, and over 4000 had completed the spoken data task. All these "subjects" submitted social metadata. It is the quiz evidence that we examine here.

On the one hand, the EDA results show that there has indeed been considerable dialect levelling with many features that were robust at the time of the Survey of English Dialects 50 years ago now in rapid decline: the use of non-prevocalic /r/ (e.g. 'arm [arm]'), the use of third person singular present tense zero (e.g. 'she feed'), or the use of lexemes other than 'autumn' to denote the season after summer. On the other hand, some features in some regions show much less retreat or indeed robust maintenance of non-standard local forms: the use of what Wells (1982) calls 'velar nasal plus' (e.g. 'tongue [tʌŋg]'), certain forms of ditransitives (e.g. 'I gave it him'), 'clear' /l/ in non-prevocalic environments, and 'spelk' as a variant of 'splinter' (a small piece of wood under the skin). The EDA also showed that local lexis was generally most vulnerable to levelling, with local phonology least susceptible.

Levelling is especially vigorous in those areas of the country that have seen dramatic and ongoing population mobility – such as those areas of the rural South which have experienced considerable levels of inward counter-urbanisation since World War 2, but it is somewhat less advanced in the urban North East and parts of the North West which have experienced a rather different demographic dynamic over the past century. Our data also show that levelling is socially differentiated – more advanced among some social groups than others. We conclude by pointing to the important ways in which surveys like the EDA can complement more locally-based ethnographic studies of dialect levelling and linguistic change in general.

References

Leemann, A., Kolly, M.-J. and Britain, D. (2018). "The English Dialects App: The creation of a crowdsourced dialect corpus", *Ampersand* 5, 1-17.

Wells, J. (1982). *Accents of English*, Cambridge, Cambridge University Press.

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

Mónica Castillo Lluch (Université de Lausanne)

Dialectos del español est une application gratuite disponible sur www.dialectosdelespanol.org et sur Google Play. Elle a été conçue par Miriam Bouzouita (Université de Gand), Mónica Castillo Lluch (Université de Lausanne) et Enrique Pato (Université de Montréal) sur le modèle des applications telles que *Voice Äpp* pour l'étude des variétés suisses-allemandes et *English Dialects App* pour l'anglais britannique. Face à ces applications dialectales développées pour d'autres langues, *Dialectos del español* vise à couvrir un public beaucoup plus large – le monde hispanophone dans son ensemble – et dépasse donc de loin la perspective régionale ou nationale précédemment utilisée, ce qui engendre des défis particuliers. L'application privilégie les questions grammaticales, généralement oubliées dans les atlas linguistiques traditionnels. Une autre innovation est l'importance accordée aux dynamiques migratoires des hispanophones, qui peuvent être révélées à l'aide de questions détaillées sur la mobilité des utilisateurs de l'application.

Le but de cette communication sera de présenter le projet d'un point de vue : 1) scientifique (le cadre dans lequel il naît, ses objectifs, la démarche adoptée pour préparer les 26 questions de l'application et le code mis en place pour que l'application tente de deviner le dialecte des utilisateurs hispanophones) ; 2) médiatique (comment depuis son lancement le 15 mai 2019 nous avons essayé de faire de la publicité sur l'application en Espagne et en Amérique pour obtenir la plus large participation) ; 3) des résultats (présentation des premiers résultats sur des phénomènes déjà étudiés dans le passé mais aussi sur des innovations linguistiques encore inexplorées).

*****

Étude des parlers du Croissant via la cartographie informatisée

Amélie Deparis (INALCO, Paris)

L'aire que l'on nomme le Croissant linguistique – terme venant de Ronjat (1913) – correspond à la frange nord du Massif Central (Brun-Trigaud, 1990). Plus précisément, d'ouest en est, cette zone traverse les départements de la Charente, la Vienne, la Haute-Vienne, l'Indre, la Creuse, le Cher, l'Allier et le Puy-de-Dôme. Les parlers de cette aire ont la particularité d'être en contact avec des variétés d'occitan (limousin et auvergnat), des langues d'oïl (poitevin, berrichon et bourbonnais), dans une moindre mesure le francoprovençal, ainsi qu'avec la langue nationale, le français. Les parlers du Croissant présentent donc simultanément des traits caractéristiques des différentes formes de galloroman mentionnées ci-dessus. Les premiers à s'intéresser à cette zone et à essayer d'en déterminer les limites sont les chercheurs C. Tourtoulon et O. Bringuier (1876), mais leur étude n'a pu être achevée. Le projet « Les parlers du Croissant : une approche multidisciplinaire du contact oc-oïl » (http://parlersducroissant.huma-num.fr/) a pour but de développer les connaissances disponibles sur les parlers du Croissant, en combinant l'apport des travaux passés à de nouvelles recherches (enquêtes de terrain, études descriptives, comparatives et multi-disciplinaires). Dans le cadre de cette communication, je présenterai dans une première partie le projet « Les parlers du Croissant » et son action. Puis, dans une seconde partie, je présenterai quelques caractéristiques marquantes de la zone du Croissant. Dans une troisième

partie, je me centrerai sur ma propre recherche doctorale qui, dans le cadre dudit projet, vise notamment à sélectionner des traits caractéristiques des parlers du Croissant présentant un intérêt au niveau (i) du lexique (ii) de la morphologie, (iii) de la sémantique, (iv) de la phonologie et de la phonétique. Je montrerai, en m'appuyant sur des exemples précis, comment ces traits peuvent donner lieu à l'élaboration de représentations cartographiques (conçues grâce à des outils informatiques contemporains) afin d'illustrer leur variation dans l'espace. Je conclurai ensuite sur l'intérêt de ces expériences cartographiques pour l'analyse de la variation dialectale, dans le Croissant et ailleurs.

Références

Brun-Trigaud, G. (1990). *Le Croissant : le concept et le mot. Contribution à l'histoire de la dialectologie française au XIX^{e} siècle*. (Thèse de doctorat). Université Jean- Moulin-Lyon-III.

Ronjat, J. (1913). *Essai de syntaxe des parlers provençaux modernes*. Mâcon, Protat Frères.

Tourtoulon, C & Bringuier, O. (1876). *Dossier sur la mission en France ayant pour but d'étudier la limite entre la langue d'oc et la langue d'oïl*. Paris, Archives Nationales.

*****

Word Frequencies as Dialect Features

Jack Grieve (University of Birmingham)

Dialectologists generally focus on mapping different linguistic forms that are used to express the same meaning. For example, well know regional patterns in English include the different words used to refer to a 'soft drink' in the US and a 'bread roll' in the UK. These types of 'alternation variables' are the standard in dialectology, because they allow us to observe variation in form, while controlling for variation in meaning. It is also easy to collect data on alternations through dialect surveys, where we can simply ask informants, for example, which word they use to refer to a given concept. These days, however, it is even easier to collect large amounts of regionalised natural language data online, especially from Twitter. This approach to data collection opens up the possibility of mapping the frequencies of individual words across a region of interest, measured relative to the total number of words at each location. In this presentation, I argue for the value of this modern approach to dialectology. Focusing on variation in the relative frequencies of common words in a multi-billion word corpus of geolocated US Twitter data, I will demonstrate that this type of analysis provides a new and important perspective on regional variation in linguistic form.

*****

Apps for mapping language variation and change in German-speaking Europe

Adrian Leemann (Universität Bern)

Methods in sociolinguistics have been evolving since the beginning of the discipline in the 1960s. Typically, these methods included a researcher conducting surveys, interviews, questionnaires, ethnographic and participant observation, and systematic note-taking and record keeping (Hernández-Campoy 2014). This was a time where notepads and microphones were the central tool of linguistic data collection. Today, the majority of citizens of westernized countries have access to a smartphone, and thereby a microphone and notepad,

facilitating the collection of linguistic data. In Germany alone in 2018, 79% of the population owned a smartphone (Newzoo 2019). Global coverage is uneven: in 2018, smartphone user penetration as percentage of total global population was 34.7%, predicted to rise to 40% by 2021 (Statista 2019).

Given the ubiquity of smartphones, we wish to discuss how these devices can be exploited for linguistic data collection and how we can use the geographically fine-grained data collected to map regional variation. In 2012, we began developing a smartphone app for Swiss German, called *Dialäkt Äpp* (Leemann 2013; Kolly and Leemann 2013). *Dialäkt Äpp* has two core functionalities: predicting where the user is from based on their choice of dialect words and recording the user's dialect. The app served as a springboard for a Swiss National Science Foundation project called *Voice Äpp* (Leemann 2015; Leemann *et al*. 2015). Aside from predicting a user's regional origin using automatic speech recognition, *Voice Äpp* features an analysis of speaking rate and fundamental frequency; it further illustrates hearing pathologies as well as speech phenomena like the Cocktail Party (Cherry 1953) and McGurk Effects (McGurk and MacDonald 1976). Finally, in 2015, we developed the web-app *Grüezi, Moin Servus* (Leemann *et al*. 2018) in collaboration with Zurich-based *Tagesanzeiger* and *Spiegel Online* which predicts users' regional origins within all of German-speaking Europe.

In all these apps, users provided current data on regional variation in German-speaking Europe. In this talk we discuss how we can use this data from nearly 1M participants to map language variation and change by showcasing selected results. In equal measure the talk focuses on problems we encountered with different mapping techniques and discusses why we opted for certain methods over others. The talk also addresses more general limitations and pitfalls of smartphone app-based data collection. Finally, we attempt a look into the future, discussing potential directions and trends in using apps for linguistic data collection in the context of studies on language variation and change.

References

Cherry, E. Colin. 1953. Some experiments on the recognition of speech, with one and with two ears. *Journal of the Acoustical Society of America* 25(5). 975–979.

Hernández-Campoy, Juan Manuel. 2014. Research methods in Sociolinguistics. *AILA Review* 27(1). 5–29.

Kolly, Marie-José & Adrian Leemann. 2013. Dialäkt Äpp: Dialektologie vermitteln – Dialekte ermitteln. *Jahresbericht des Schweizerischen Idiotikons* 2013. 21–44.

Leemann, Adrian. 2013. *Dialäkt Äpp*. https://itunes.apple.com/ch/app/dialakt-app/id606559705?mt=8.

Leemann, Adrian. 2015. *Voice Äpp*. https://itunes.apple.com/ch/app/voice-app/id950037736?mt=8.

Leemann, Adrian, Marie-José Kolly, Jean-Philippe Goldman, Volker Dellwo, Ingrid Hove, Ibrahim Almajai, Sarah Grimm, Sylvain Robert & Daniel Wanitsch. 2015. Voice Äpp: a mobile app for crowdsourcing Swiss German dialect data. *Proceedings of Interspeech 2015*. 2804–2808.

Leemann, Adrian, Stephan Elspaß, Robert Möller & Timo Grossenbacher. 2018. *Grüezi, Moin, Servus: Wie wir wo sprechen*. Hamburg: Rowohlt.

McGurk, Harry, & MacDonald, John. 1976. Hearing lips and seeing voices. *Nature* 264. 746–748.

Newzoo. 2019. *Top 50 Countries/Markets by Smartphone Users and Penetration* https://newzoo.com/insights/rankings/top-50-countries-by-smartphone-penetration-and-users/

Statista. 2019. *Smartphone user penetration as percentage of total global population from 2014 to 2021*. https://www .statista.com/statistics/203734/global-smartphone-penetration-per-capita-since-2005/

*****

VerbaAlpina
Digital Geolinguistics Dedicated to the Lexical Analysis of the Alpine Region

Stephan Lücke (Ludwig-Maximilians-Universität München)

Since 2014 the DFG-funded long term project VerbaAlpina (VA) is run at the Ludwig-Maximilians-University of Munich (LMU). VA is a cooperation of the Institute of Romance Studies and the LMU Center for Digital Humanities (DH; IT-Gruppe Geisteswissenschaften). The project focuses on lexical variation throughout the Alpine area as defined by the so-called Alpine Convention (https://www.alpconv.org/). Whereas geolinguistic research within the Alpine region is traditionally orientated towards the spread of national languages and towards political borders, VA takes the homogeneous natural environment of the mountaneous region and the resulting uniform habitat conditions and ways of living as the guiding parameters defining its area of research. VA is conceptualized as a strictly digital project that uses web technology for various purposes such as documentation, publication and visualisation. VA takes its data from traditional geolinguistic publications, mainly linguistic atlases and suitable dictionaries (i.e. dictionaries providing geographic information). The strictly digital approach is associated with several challenges starting from the difficulties regarding the transcription of the sometimes complex phonetic characters that are used especially in some of the linguistic atlases. VA has developed a series of specific reusable and freely available online tools that are used within the workflow of digitizing data from the printed sources. Another tool, the so-called Crowdsourcing tool, was built for gathering speech data from online users with the aim of filling documentation gaps that result from inconsistencies of the available printed sources. An interactive online map that is using performant up-to-date graphical technology (WebGL) offers suggestive qualitative and quantitative visualisation of geographic distribution patterns from onomasiological and/or semasiological perspectives. These can also be combined with non linguistic data such as the sites of latin inscriptions. In addition to the geolinguistic core themes of the project, VA is providing methodological reflexion on many of the issues deriving from the strictly digital orientation that should be of interest also beyond the borders of the project and even beyond the field of geolinguistics. In general, VA is looking for perspectives and solutions that allow the linkage of lexical data across so far isolated domains of geolinguistic research projects with the option of real interoperability (the “I” in the acronym FAIR). This talk will provide more detailed information on the mentioned aspects of the project VerbaAlpina.

*****

Recettes de Grammaire :
production participative de ressources linguistiques variées pour l'alsacien

Alice Millour & Karën Fort (Sorbonne Université)

La production participative pour le traitement automatique des langues (TAL) a fait ses preuves dans le cadre de projets de construction collaborative de ressources linguistiques diverses (réseaux lexicaux, corpus annotés en syntaxe de dépendances, corpus annotés en anaphore[3]), pour des langues telles que le français ou l'anglais.

Cette méthode, plaçant les locuteurs au cœur de la création de ressources linguistiques, peut servir un objectif double lorsqu'elle est appliquée à un continuum dialectal présentant des variations spatiales et scripturales. Outre la mise à contribution des locuteurs pour réaliser une tâche linguistique définie (par exemple l'annotation d'un corpus en morphosyntaxe), elle permet également de profiter du contact établi avec les participants pour recueillir auprès d'eux des informations sur les mécaniques de la variation linguistique à l'œuvre. Traditionnellement, les outils de TAL sont peu robustes à la variation, quelle qu'elle soit, et sont d'autant plus performants que les textes qu'ils traitent sont conformes à un standard orthographique défini. Or, afin qu'ils soient utilisables par les locuteurs, des outils tels que les moteurs de recherche, claviers numériques, ou dictionnaires en ligne, doivent prendre en compte ces phénomènes de variation, qu'elle soit dialectale, scripturale, ou une accumulation des deux.

Les projets Bisame, puis Recettes de Grammaire, ont été mis au point pour permettre la production participative de trois ressources : (i) des corpus bruts, sous la forme de recettes de cuisine, (ii) des corpus annotés en morphosyntaxe, selon un schéma d'annotation simplifié basé sur les catégories universelles[4], (iii) des graphies alternatives pour les mots présents sur le site. Les deux plateformes, conçues comme indépendantes de la langue d'instanciation, ont été développées pour l'alsacien dans un premier temps, puis pour les créoles guadeloupéen et mauricien.

Dans le cas de l'alsacien, 347 variantes scripturales ont été proposées pour 148 mots différents grâce à la fonctionnalité « Moi, j'aurais dit ça comme ça ! » permettant de proposer des graphies alternatives à celles saisies par les participants partageant des recettes. Par exemple, le mot original « *Erdbeere* » (« fraise ») a été associé à quatre variantes : « *Arbeere* », « *Erdbéére* », « *Arbere* », « *Ardbeera* ».

L'étude de ces entrées nous permet de retrouver des motifs connus de la variation dialectale, tels que l'alternance des voyelles finales « e » et « a » propres respectivement aux variantes bas-rhinoise et haut-rhinoise. Elles nous permettent également d'identifier plus largement d'autres motifs de variation, liés notamment aux habitudes scripturales des locuteurs. Sur la base de la ressource produite par les participants, nous avons extrait un certain nombre de règles correspondant aux motifs de substitution identifiés. Ces règles ont ensuite été utilisées pour générer automatiquement des couples de variantes scripturales potentielles sur la base des mots présents dans les textes disponibles en ligne et saisis sur la plateforme. Ont ainsi été

[3] Voir les projets JeuxDeMots (www.jeuxdemots.org), ZombiLingo (https://zombilingo.org/), et Phrase Detectives (https://anawiki.essex.ac.uk/phrasedetectives/), par exemple.
[4] Voir https://universaldependencies.org/u/pos/all.html.

générées 876 paires de variantes additionnelles qui ont été évaluées par un professeur d'alsacien. Cette ressource produite automatiquement est encore imparfaite, mais nous espérons pouvoir affiner notre algorithme d'application des règles en intégrant notamment à notre algorithme des informations sur la provenance des locuteurs lorsqu'ils la renseignent dans leur profil.

La description de ce travail sera accompagnée d'une présentation des résultats de l'enquête interrogeant les pratiques des locuteurs de l'alsacien en ligne intitulée « l'alsacien, Internet, et moi », réalisée en ligne en janvier et février 2019 et ayant recueilli 1 224 réponses.

*****

An online atlas of colloquial German: The *Atlas zur deutschen Alltagssprache*

Robert Möller (Université de Liège)

The language used in everyday speech in the German-speaking countries is still largely marked by geographical variation, and what makes the situation more complex is the fact that there are also large regional differences in the degree of use of dialects, regional standard or intermediate varieties. In order to document the variation in colloquial contemporary German, the *Atlas zur deutschen Alltagssprache* ('Atlas of colloquial German') follows the pragmatic approach of Eichhoff (1977 ff.), asking the informants to indicate which words and grammatical constructions "one would normally hear" in informal situations in their place, "be it more dialect or more standard German".

Unlike Eichhoff's survey, the *Atlas zur deutschen Alltagssprache* can use the Internet for data collection, which makes it possible to reach about 10.000 informants and to submit new questionnaires to them regularly, while presenting the maps resulting from the previous survey. As the results reveal, the high number of participants compensates for the uncertainty arising from the uncontrolled recruitment of informants. The maps give a detailed picture of variation phenomena in colloquial German (mainly as to the lexicon, but also concerning grammatical constructions). Moreover, being exactly comparable with Eichhoff's data, they provide insights into the changes over the last 30-40 years.

References

Elspaß, Stephan/Robert Möller (2003 ff.): *Atlas zur deutschen Alltagssprache (AdA)*. URL: <www.atlas-alltagssprache.de>

Eichhoff, Jürgen (1977 ff.): *Wortatlas der deutschen Umgangssprachen.* Bd. I/II: Bern [1977/78]; Bd. III: München et al. [1993]; Bd. IV: Bern/München [2000]: K.G. Saur.

Möller, Robert / Stephan Elspaß (2015): Atlas zur deutschen Alltagssprache. In: Kehrein, Roland / Alfred Lameli / Stefan Rabanus (eds.): *Regionale Variation des Deutschen – Projekte und Perspektiven*. Berlin, Boston: de Gruyter, 519-540.

Möller, Robert / Stephan Elspaß (2019): Die rezente Dynamik im arealsprachlichen Lexikon. In: Joachim Herrgen / Jürgen Erich Schmidt (eds.): *Language and Space. An International Handbook of Linguistic Variation. Vol. 4: German.* Berlin, Boston: de Gruyter (Handbücher zur Sprach- und Kommunikationswissenschaft, 30.4). 756-781.

L'Atlas linguistique de la France informatisé (ALFi)

Yves Charles Morin (Université de Montréal)

Le projet d'informatisation de l'ALF est un travail en cours qui s'est développé par la consolidation des saisies faites dans différents projets de recherches depuis 1990. Dans cette communication, je présenterai ses objectifs principaux et le bilan de l'entreprise.

L'objectif primitif était de concevoir un outil permettant de produire rapidement des cartes dialectologiques permettant d'évaluer des hypothèses d'évolution phonétique des dialectes galloroman d'oïl spécifiques à chacun de ces projet. Il s'est développé progressivement pour constituer un outil de référence de toutes les formes de l'ALF (mais non de son supplément) qui ont fini par être intégrées dans une base de données relationnelle interrogeable.

Il a été procédé à une édition quasi philologique des formes colligées dans l'ALF en n'apportant de corrections (dûment signalées) que de façon exceptionnelle et en s'imposant de conserver intégralement l'information présente dans l'ALF. Le respect des sources porte sur la transcription phonétique originale et la segmentation en unités morphosyntaxiques. Des algorithmes permettent la conversion vers d'autres transcriptions phonétiques (p. ex. API), tandis que la base contient des champs supplémentaire pour y porter des découpages morphosyntaxiques originaux.

Les cartes de l'ALF contiennent deux types d'objets : (1) un mot-forme isolé (*fauteuil*) ou une expression simple (*un bel homme*) recueillis hors contexte en réponse aux questions de l'enquêteur – parfois accompagné des formes fléchies (complètes ou abrégées) du féminin ou du pluriel (*bossu, bossue*) des noms ou adjectifs –, (2) des extraits d'une expression plus large, souvent une phrase ou une série de phrases (*il partit au bout d'une semaine, nous ne le revîmes plus*), portés séparément dans plusieurs cartes (c. 976 : *il partit* + c. 162 *au bout* + c. 1214 *d'une semaine,* + c. 1154 *nous ne le revîmes plus*). Une même carte peut rassembler des objets obtenus en réponse à des questions différentes ; p. ex., la c. 1214 collige les deux objets suivants : 1. la forme obtenue (ou les formes obtenues, dans les cas de réponses multiples) pour l'équivalent dialectal de *La semaine*, et 2. la partie correspondant à *d'une semaine* pour celui de la série de phrases précédentes (*il partit … revîmes plus*). Un système de coordonnées uniques pour chacune des formes permet de reconstruire pour chacun des points l'expression complète d'où ont été extraits les objets (le cas échéant) ; il permet également de sélectionner l'ensemble des réponses, soit à l'ensemble des questions comme ils apparaissent sur la carte imprimée primitive, soit à chacune des questions.

La base comporte une double typisation des formes : une typisation primaire étymologique, et une typisation secondaire empruntant le classement du FEW. La typisation primaire exige l'élaboration d'un type commun à l'ensemble des dialectes galloromans pour chacune des formes *fléchies* enregistrées dans l'ALF. Les types primaires répondent le plus souvent à des besoins mnémoniques et sont repris dans un lexique dédié, où l'on en trouve la définition : étymons des formes non restructurées (mots héréditaires ou emprunts) et sinon identification des affixes ou opérations morphologiques qui le définissent. Un type temporaire peut (trop fréquemment, malheureusement) être assigné aux formes qui résistent totalement ou partiellement à l'analyse. Les types primaires peuvent être modifiés à tout moment. Les types secondaires sont ceux du FEW, p.ex. FEW 15.1196b (II.2.a), dont l'utilisation n'implique pas

nécessairement leur adoption inconditionnelle. Le FEW a complètement dépouillé l'ALF et constitue un outil indispensable à son analyse. Pour compléter l'arrimage avec le FEW, l'ALFi enregistre systématiquement toutes les formes de l'ALF qui ont été répertoriées dans le FEW (opération limitée en ce moment à celles dont le point a été identifié par son numéro).

Les moteurs de recherches disponibles dans la base permettent de sélectionner les formes en fonction du type primaire, du type secondaire, des unités phonétiques dans les contextes phonétiques définis par des expressions régulières, des informations morphologiques, des marques d'usage et de toute combinaison booléenne de ces propriétés .

*****

Regional variation and the socio-pragmatics of online writing.
A case study in the German-speaking area

Christoph Purschke (Université du Luxembourg) & Dirk Hovy (Università di Milano)

In this talk, we discuss regional variation in online writing as a linguistic resource and pragmatic strategy. Social media writing differs considerably from other domains of language use. In online social spaces like Twitter, Facebook, and WhatsApp, people employ a variety of linguistic resources and develop complex and often hybrid writing styles to communicate effectively and playfully. At the same time, online writing preserves all the socio-pragmatic functions of language in practice, i.e., identity building, social positioning, the negotiation of relationships, and discourse organization. For German, online writing has been analyzed mostly with respect to the use of linguistic resources typical of digital culture (like emojis, abbreviations, or non-standard forms). Regional variation, on the other hand, has so far been considered only to a limited extent, e.g., for Switzerland or local IRC groups. However, due to the hybrid nature of online writing, social media is likely to offer a range of novel insights for the study of regional variation.

Starting from this premise, we use an integrated approach that combines computational methodology with in-depth sociolinguistic analysis. Our study analyzes a corpus of more than 3 million anonymous discussions collected from the social media platform "Jodel" in the entire German speaking area (Hovy/Purschke 2018, Purschke/Hovy forthcoming). Using computational linguistics methods like neural networks and representation learning, we analyze the data without assuming a specific linguistic structure for regional variation or the pragmatic organization of conversations. Nonetheless, our analysis reveals clear-cut regional clusters of language use that can be interpreted against the backdrop of linguistic and socio-cultural spatial structures (dialect division, socioeconomic mobility, sociocultural orientation, attitudes).

Our analysis of the revealed spatial structures shows that (anonymous) social media communication of young adults in the German speaking area can indeed be characterized by "digital regiolects" which are a) regionally distinct, b) structured by the use of specific linguistic resources, and c) closely linked to the overall structure of the German regional languages (as well as other socioeconomic and socio-cultural factors). The typical language use of different user groups also mirrors region-specific topic profiles. These profiles provide information about the socio-pragmatic organization of practice in the respective regional communities. Taken together, region-specific writing styles and topic profiles shed light on

different aspects of social dynamics in digital communication, e.g., regarding the spread and establishment of new regional or group-specific variants in the “Jodel” community.

References

Hovy, Dirk / Purschke, Christoph (2018). Capturing Regional Variation with Distributed Place Representations and Geographic Retrofitting. Proceedings of EMNLP 2018. Brussels.

Purschke, Christoph / Hovy, Dirk (forthcoming). Lörres, Möppes, and the Swiss. (Re)Discovering Regional Patterns in Anonymous Social Media Data. In: Journal of Linguistic Geography, 7(2).

*****

Revisiter les données dialectologiques de la Société du parler français au Canada (1904-1906) : enjeux et perspectives

Wim Remysen (Université de Sherbrooke)

La Société du parler français au Canada est surtout connue pour son *Glossaire* publié en 1930, véritable pièce maîtresse de la lexicographie québécoise (Mercier 2002). Trois enquêtes ont mené à la parution de cet ouvrage, parmi lesquelles une consultation dialectologique menée de 1904 à 1906 auprès d’environ 175 collaborateurs à travers la province. Réalisée une soixantaine d’années avant la mise sur pied des grands chantiers dialectologiques qui ont marqué la linguistique au Québec dans les années 1970 (v. notamment l’ALEC), cette enquête est restée méconnue, mis à part quelques études (Lavoie 1979, Laflamme 2004, Mercier 1999, 2002). Les données qu’elle a permis de recueillir au sujet de près de 4 000 emplois, essentiellement des particularismes lexicaux et sémantiques propres à la langue populaire des Québécois, demeurent ainsi largement sous-exploitées, malgré leur intérêt évident pour l’histoire du français québécois.

L’exploitation des relevés dialectologiques de la Société permet notamment de revisiter la division et l’évolution du territoire québécois en deux grandes aires linguistiques (Remysen 2016). En outre, contrairement à ce qu’ont fait les dialectologues de l’ALEC, la Société n’a pas écarté les milieux urbains dont on sait le rôle important du point de vue des changements linguistiques : l’enquête menée par la Société fournit ainsi des indices qui permettent de s’interroger sur l’influence que l’histoire de l’occupation de l’espace québécois (développement de réseaux, migrations à l’intérieur de la province, urbanisation de la population, etc.) a pu avoir sur la variation régionale de l’époque. Cela nous paraît d’autant plus pertinent que, dans l’histoire de l’occupation du territoire québécois, la période couvrant la deuxième moitié du 19^{e} siècle et le début du 20^{e} constitue une période charnière : elle correspond à l’élargissement progressif du territoire habité de la vallée laurentienne (Courville 1996), phénomène qui a donné naissance à de nouveaux espaces régionaux dont certains ont formé des expressions identitaires assez distinctes, ce qui n’a pas été sans effet sur le plan des pratiques linguistiques.

Compte tenu de leur richesse documentaire et de leur valeur patrimoniale, nous avons entrepris en 2016 un projet de numérisation en vue d’exploiter et de cartographier les données dialectologiques recueillies par la Société. Dans le cadre de cette communication, nous nous proposons de réfléchir à quelques problèmes d’ordre méthodologique que soulève le traitement des données et leur interprétation, en plus de montrer quelques études de cas qui

illustrent tout leur potentiel. Nous aborderons plus particulièrement les quatre aspects suivants : 1º réalisation de l'enquête et aperçu des données, 2º traitement numérique des données, 3º analyse de quelques sous-échantillons, 4º perspectives de développement. Il s'agit en somme de montrer comment la dialectologie historique permet de revisiter des données du passé pour répondre à de nouvelles préoccupations de recherche (v. à ce sujet Alcorn *et al.* 2019).

Références

Alcorn, R., J. Kopaczyk, B. Los et B. Molineaux (dir.), *Historical dialectology in the digital age*, Edinburgh, Edinburgh University Press.

ALEC : Dulong, G. et G. Bergeron (1980), *Le parler populaire du Québec et de ses régions voisines : atlas linguistique de l'Est du Canada*, Québec, Ministère des communications.

Courville, S. (dir.) (1996), *Population et territoire*, Sainte-Foy, Presses de l'Université Laval.

Laflamme, C. (2004), « Distribution de quelques variantes géolinguistiques dans les parlers populaires de l'Est du Canada : essai de comparaison diachronique », dans L. Mercier (dir.), *Français du Canada – français de France VI*, Tübingen, Max Niemeyer, p. 123-149.

Lavoie, Th. (1979), « Le projet d'atlas dialectologique de la Société du parler français au Canada », *Protée*, vol. 7, nº 2, p. 11-45.

Mercier, L. (1999), « Informatisation et édition des relevés de l'enquête géolinguistique de la Société du parler français au Canada (1904-1907) », *Dialangue*, vol. 10, p. 9-15.

Mercier, L. (2002), *La Société du parler français au Canada et la mise en valeur du patrimoine linguistique québécois (1902-1962) : histoire de son enquête et genèse de son glossaire*, Québec, Presses de l'Université Laval.

Remysen, Wim, « La valorisation et l'exploitation de la documentation linguistique produite par la Société du parler français au Canada : l'exemple de ses relevés géolinguistiques », dans Wim Remysen et Nadine Vincent (dir.), *La langue française au Québec et ailleurs : patrimoine linguistique, socioculture et modèles de référence*, Frankfurt am Main, Peter Lang, p. 41-69.

*****

Interactive dialect maps for German-speaking Switzerland and other European dialect areas

Yves Scherrer (Université d'Helsinki)

The web site dialektkarten.ch has been online since 2014. It presents various interactive visualisations of the Swiss German dialect atlas SDS (*Sprachatlas der deutschen Schweiz*). Besides a set of digitized feature maps, it also provides results of dialectometric analyses as well as prototypes of machine translation and dialect identification systems for the entire German-speaking area of Switzerland. I will present recent developments of this web site, including technical updates as well as the addition of further digitized SDS maps.

The same interface has been applied to new sister projects which present interactive visualisations of British English, Gallo-Romance, Italian and Romansh dialect data. These

sister projects have been realized in collaboration with the Salzburg dialectometry group under Prof. Hans Goebl and are based on the SED (*Survey of English Dialects*), ALF (*Atlas linguistique de la France*) and AIS (Atlante linguistico ed etnografico d'Italia e della Svizzera meridionale), respectively.

Finally, I will give some examples of recent corpus-based research that shows how automatically trained dialect normalization tools can provide useful insights about variation patterns.

*****

From field notebooks to the computer screen:
the digital edition of the *Atlas Lingüístico de la Península Ibérica*

Xulio Sousa (Instituto da Lingua Galega, Universidade de Santiago de Compostela)

The *Linguistic Atlas of the Iberian Peninsula* (ALPI) was a linguistic geography project begun at the beginning of the 20th century in the Centre for Historic Studies by Ramón Menéndez Pidal and Tomás Navarro Tomás. This linguistic atlas was conceived within the framework of traditional dialectology and following the model of the pioneering works of linguistic geography published in Europe in the early 1900s. Its main objective was to describe the rural Romance varieties spoken in the peninsula in the first half of the last century. The field work of the project was interrupted by the Spanish Civil War and the first and only volume published on paper did not see light until 1962. Since then, 90% of the collected documentation had remained hidden and unavailable to the scientific community.

The presentation will show the general characteristics of the current ALPI digital edition project, launched at the Superior Centre for Scientific Research under the coordination of Pilar García Mouton and with the collaboration of researchers from different universities. The digital edition project of the edited and unpublished materials of this work is based on the methodological principles of digital humanities and open science. The aim is to offer users access to different information formats (images of notebooks, transcripts, cartographic displays, etc.) that enable their analysis and treatment. The exposition focuses on the tasks related to the digitisation and computerisation of the linguistic information of the ALPI.