Société de Linguistique de Paris

Journée Scientifique du 12 juin 2021

Les langues régionales de France: nouvelles approches, nouvelles méthodologies, revitalisation

https://zoom.us/j/98215747055?pwd=bDN2Z2pmQ0dOU1g5QmxxY0JFQ1FkZz09
ID: 982 1574 7055, Mot de passe: 122764
Trouvez votre numéro local: https://zoom.us/u/adtpI1EUFV
Version Zoom récente pour accéder aux groupes de discussion.

Organisateurs :
Annie Rialland (UMR 7018 LPP CNRS/ Univ. Sorbonne-Nouvelle),
Michela Russo (Univ. de Lyon 3 & UMR 7023 CNRS/ Paris 8),
Catherine Schnedecker (Univ. de Strasbourg LiLPa EA 1339)

Cette journée se propose de réunir des linguistes de diverses spécialités autour des questions suivantes : en quoi les études des langues régionales évoluent-elles ? Quels sont les progrès en termes de données ? En quoi les nouvelles technologies (bases de données, fouille de données, intelligence artificielle) font elles avancer l'étude des langues régionales que ce soit du point de vue synchronique ou diachronique? Quelles sont les nouvelles approches théoriques ? Comment se dessine l'avenir de ces langues?

Programme

9h-9h10 Introduction

9h10-9h50, **Ricardo Etxepare (CNRS, IKER, UMR 5478, Bayonne)**
La langue basque : défis sociaux, paris scientifiques

9h50-10h30, **Mélanie Jouitteau (CNRS, IKER, UMR 5478, Bayonne)**
Science citoyenne, science ouverte et ressources numériques sur le breton

Pause

10h50-11h30, **Delphine Bernhard (Strasbourg Univ., LiLPa, EA 1339)**
Traitement automatique des langues régionales de France : retour d'expérience sur les dialectes alsaciens

11h30-12h10, **Christophe Rey (Cergy-Paris Univ., EA 7518, LT2D)**
Peut-on revitaliser la langue picarde grâce aux nouvelles technologies?

12h10-12h50, **Nicolas Quint (CNRS, LLACAN, UMR 8135, Villejuif)**
Les parlers du Croissant : un aperçu des actions actuelles de documentation et de promotion d'un patrimoine linguistique menacé

Pause déjeuner

13h50-14h30, **Myriam Bras (Toulouse 2, CLLE, UMR 5263)**
Nouvelles perspectives pour la linguistique occitane à partir de la base textuelle BaTelòc

14h30-15h10, **Hervé Lieutard (Montpellier 3, ReSo)**
Nouvelles approches linguistiques et lexicographiques de l'occitan médiéval

Pause

15h30-16h10, **Patrick Sauzet (Toulouse 2, CLLE, UMR 5263)**
L'occitan « *pro-drop* or not *pro-drop* »: l'éclairage de la base de données en ligne SYMILA (http://symila.univ-tlse2.fr/)

16h10-16h50, **Jonathan Kasstan (University of Westminster, UK) & Michela Russo (Lyon 3/SFL UMR 7023 CNRS/Paris 8)**
Maintenance *in* shift: on nasalisation in transitional Francoprovençal and Occitan areas

16h50-17h30, **Philippe Boula de Mareüil (CNRS, LISN, Orsay), Marcel Courthiade (INALCO, Paris), Frédéric Vernier (CNRS, LISN, Orsay)**
De la Provence aux Balkans : discours épilinguistiques autour d'un atlas sonore des langues régionales ou minoritaires d'Europe

Résumés

9h10-9h50, **Ricardo Etxepare (CNRS, IKER, UMR 5478, Bayonne)**

La langue basque : défis sociaux, paris scientifiques

L'aire linguistique basque représente aujourd'hui un domaine privilégié pour l'étude de la variation linguistique et le contact des langues. D'une part, le basque est divisé en de nombreuses variétés locales et régionales, qui demeurent souvent peu explorées du point de vue scientifique ; d'autre part, les basques ont depuis environ trois ou quatre décennies une variété de basque standard (*euskara batua*), qui est utilisée, sous des formes flexibles, tant au sein du système scolaire que dans les médias majeurs et dans la plus grande partie de la production littéraire. Durant les dernières décennies, en raison notamment du système scolaire par immersion, un nouveau type de locuteurs est apparu correspondant aux personnes dont la langue maternelle n'est pas nécessairement le basque, mais le français, l'espagnol ou une langue d'immigration, et qui ont acquis leur compétence en basque à un âge précoce dans le cadre scolaire (un phénomène en claire augmentation). Ceci a produit une nouvelle source de tension entre le basque développé dans ce contexte et le standard basé sur des variétés plus traditionnelles. Le terrain linguistique basque est donc un terrain idoine pour les études sur le contact des langues, la variation, et le changement linguistique. Cette richesse linguistique constitue aussi un champ de recherche privilégié pour les études sur l'acquisition du langage dans des contextes plurilingues, sur l'esprit et le cerveau plurilingue, sur le développement atypique du langage, sur la didactique des langues et sur d'autres approches en relation avec l'étude du langage et de l'éducation plurilingue et interculturelle.

Pendant ma présentation, j'essayerai de rendre compte d'un certain nombre d'évolutions dans les études sur la langue basque, et sur la manière dans laquelle ces évolutions s'articulent autour des défis plus larges, concernant (i) la codification d'une langue commune ; (ii) la tension entre norme et usage ; (iii) la généralisation du locuteur plurilingue ; (iv) le développement des outils de traitement automatique et la création de *big data* ; (v) le devenir sociolinguistique de la langue et les bases idéologiques, juridiques et matérielles des politiques linguistiques, notamment en France.

Cette présentation privilégie des éléments d'information relevant de la situation de la langue basque en France, mais il est impossible à ce jour de dresser un panorama des études basques et des questions d'ordre sociologique qui se posent quant à son avenir, sans élever le regard vers les deux côtés du Pays basque. Le pays basque espagnol fournit d'une part, un contexte universitaire et de recherche plus riche en termes de ressources, et d'autre part, un espace juridique et politique (ayant longtemps servi comme référence) qui a encouragé les initiatives d'action-recherche autour du plurilinguisme. Les réseaux de recherche français et espagnols tournés vers la langue basque travaillent depuis longtemps en symbiose. Cette présentation se situe dans une perspective qui englobe l'ensemble des études basques.

9h50-10h30, **Mélanie Jouitteau (CNRS, IKER, UMR 5478, Bayonne)**

Science citoyenne, science ouverte et ressources numériques sur le breton

Je dresse un état des lieux de l'étude de la langue bretonne au XXI°, sur son pan numérique et participatif. Le breton est la langue celtique parlée dans la partie ouest de la Bretagne. C'est la seule langue celtique actuelle à se développer, et dont l'étude se développe, en contact avec des langues latines et partiellement en dehors du contact de l'anglais. Ses grammaires descriptives remontent au moyen âge, par des grammairiens pratiquant des langues romanes: le bas latin, le moyen français puis le français, et pour quelques rares, le gallo adjacent en Bretagne. L'invention de l'imprimerie a modifié profondément l'accessibilité des oeuvres, leur circulation, et donc la possibilité de leur discussion critique et de leur utilisation par les locuteurs eux-mêmes et apprenants. Il en a résulté une montée sensible de la scientificité des approches, les oeuvres étant dès lors confrontées, même si parfois de façon posthume pour leur auteur, à des désaccords argumentés, eux-mêmes diffusés sous un format perdurant dans le temps. L'usage de l'imprimerie et les changements fondamentaux qu'il a entrainé sont évidemment prolongés par l'avènement du numérique, dans le cas du développement des publications en accès ouvert et des numérisations des archives.

Mon hypothèse est que, pour l'étude des langues minorisées, l'accessibilité du numérique et son développement participatif entrainera aussi des changements fondamentaux dans les pratiques scientifiques, et dans les relations entre le champ scientifique et la société. Consciente de la modestie de notre profondeur historique sur le sujet, et des réalisations émergentes sur le breton, j'étudie spécifiquement dans cette présentation l'impact qu'a le développement du numérique depuis vingt ans sur les pratiques de l'étude des dialectes bretons, et les relations entre ces études et la société. Je montre comment l'accessibilité des ressources, du dépôt et des données enrichies impacte, si même parfois encore légèrement, la scientificité des approches descriptives et même théoriques. J'illustre d'exemples concrets comment les pratiques de science ouverte rapprochent les acteurs de la recherche de la société; locuteurs, apprenants, étudiants et chercheurs, locaux et internationaux. J'identifie des obstacles et des freins au développement de cette recherche. Je prédis en particulier un changement qualitatif lorsque les traducteurs automatiques pourront jetter un pont entre d'une part les réalisations en développement en français sur le breton, et d'autre part le monde anglophone de la recherche fondamentale sur les langues celtiques.

- Desseigne, Adrien, Loïc Cheveau & Pierre-Yves Kersulec. 2013-2018. "Banque Sonore des Dialectes Bretons, projet de documentation multimédia en ligne", http://banque.sonore.breton.free.fr/index.html.
- Cheveau, Loïc & Pierre-Yves Kersulec. 2012-évolutif. *Dictionnaires bretons parlants*, http://dico.parlant.breton.free.fr/, partiellement hors ligne.
- Jouitteau, M. (éd.). 2009-2021. *ARBRES, site de recherche sur la syntaxe formelle de la microvariation syntaxique de la langue bretonne*, IKER, CNRS, http://arbres.iker.cnrs.fr.
- Menard, Martial. 2016. "Devri: Le dictionnaire diachronique du breton", http://devri.bzh/.
- Jouitteau, M. 2013. 'La linguistique comme science ouverte; Une expérience de recherche citoyenne à carnets ouverts sur la grammaire du breton', Charles Videgain (dir.), *Lapurdum* XVI, 93-115. https://journals.openedition.org/lapurdum/2357
- Persée. Collection en ligne *Etudes celtiques*, *https://www.persee.fr/collection/ecelt*.
- Smith, Jennifer & al. 2021. *Scots Syntax Atlas*. Glasgow: University of Glasgow. https://scotssyntaxatlas.ac.uk/linguists-atlas/#6.25/57.929/-4.448.
- Willis, D., 2019. Dialect syntax as a testbed for models of innovation and change: Modals and negative concord in the *Syntactic Atlas of Welsh Dialects*. *Glossa: a journal of general linguistics*, 4(1), https://www.glossa-journal.org/articles/10.5334/gjgl.844/.
- Yekel, Tangi, Riwal Georgelin & Juluan Ar C'hozh. 2016. *Brezhoneg Bro-Vear*, Blog kevredigezh *Hent don*. https://www.brezhonegbrovear.bzh/yezhadur.php?yezh=fr

10h50-11h30, **Delphine Bernhard (Strasbourg Univ., LiLPa, EA 1339)**

Traitement automatique des langues régionales de France : retour d'expérience sur les dialectes alsaciens

Les langues dites « peu dotées » sont souvent négligées dans les travaux de recherche en traitement automatique des langues (TAL), au profit des langues très dotées. Ces dernières disposent de nombreuses ressources linguistiques numériques, et en particulier de données annotées. Par exemple, une recherche avec le terme générique « corpus » sur le portail européen META-SHARE montre que plus du quart des ressources listées sont en anglais[1].
On assiste toutefois à une timide évolution dans la manière de traiter et considérer la diversité linguistique en TAL, grâce à des ateliers dédiés comme CCURL (*Collaboration and Computing for Under-Resourced Languages*), des groupes de travail (SIGUL : *ISCA Special Interest Group on Under-resourced Languages*, LITHME : *Language in the Human-Machine Era*) ou des projets, notamment européens (DLDP : *Digital Language Diversity Project*, ELE : *European Language Equality*).

Dans cette communication, nous nous intéresserons plus particulièrement au cas des dialectes alsaciens et aux progrès réalisés ces dernières années, grâce à divers projets. Nous nous appuierons notamment sur les résultats du projet RESTAURE[2] (Ressources informatisées et Traitement Automatique pour les langues régionales), financé par l'ANR (2015-2018) et dédié à trois langues régionales de France : l'alsacien, l'occitan et le picard. Nous insisterons plus particulièrement sur les défis posés : manque de données numériques, variations dialectales et graphiques, communauté de recherche réduite, difficultés à valoriser le travail de collecte et d'annotation des données. Nous montrerons également quelles solutions méthodologiques peuvent être privilégiées pour faciliter la création de ressources et renforcer la visibilité de ces langues : coopération entre équipes de recherche intéressées par diverses langues régionales, collaboration de chercheurs et chercheuses appartenant à diverses disciplines, utilisation de standards, réutilisation d'outils existant, adoption des principes FAIR[3] pour la diffusion des ressources. Il s'agira ainsi de montrer comment le travail sur des langues régionales peu ou très faiblement dotées peut, au-delà des réalisations et productions concrètes, enrichir et alimenter la réflexion sur les pratiques de recherche en TAL et linguistique outillée.

[1] http://metashare.ilsp.gr:8080/repository/search/?q=corpus Recherche effectuée le 17/05/2021.
[2] https://restaure.unistra.fr
[3] *Findability, Accessibility, Interoperabily and Reuse*. Voir Wilkinson et al. (2016) « The FAIR Guiding Principles for scientific data management and stewarship », Scientific Data, 3, https://www.nature.com/articles/sdata201618

11h30-12h10, **Christophe Rey (Cergy-Paris Univ., EA 7518, LT2D)**

Peut-on revitaliser la langue picarde grâce aux nouvelles technologies?

Au cours des deux dernières décennies, grâce à plusieurs initiatives de recherche universitaire, la langue picarde a pu bénéficier de ce que nous pouvons considérer comme des avancées particulièrement significatives pour sa description et sa diffusion. La constitution de la base de données *PICARTEXT* (2008-2011)[4] - réalisée au sein de l'Université de Picardie Jules Verne -, la conduite du projet RESsources informatisées et Traitement AUtomatique pour les langues REgionales (RESTAURE (2015-2019))[5], et dans une moindre mesure l'élaboration d'un *Atlas pan-picard informatisé* (2018-2020)[6] - à l'Université de Lille -, ont ainsi fait entrer la langue picarde dans le concert des langues régionales de France dotées de ressources numériques.

Dans le cadre de notre intervention, nous souhaitons, à travers une présentation des attentes et réalisations concrètes de ces différents projets, montrer que cette langue bénéficie désormais de ressources électroniques intéressantes pour sa description et sa valorisation.

Au-delà de la constitution même de ces ressources qu'il reste désormais à exploiter et valoriser, nous montrerons que l'un des apports du Traitement Automatique des Langues semble avoir été l'accélération de la prise de conscience, de la part des acteurs de la promotion du picard, de la nécessité de dépasser les multiples phénomènes de variation linguistique sur le vaste territoire picardophone. Aujourd'hui se pose en effet désormais davantage la question de l'existence d'outils permettant à cette langue d'accéder à une grammatisation salutaire. Nous ferons ainsi le point sur quelques avancées très récentes comme la création du *Dictionnaire fondamental français-picard* (2020)[7], celle de la « Commission de néologie et de terminologie pour la langue picarde »[8], ainsi que la conduite du projet *METALPIC* (2017-2022) réalisé dans le cadre d'un financement par l'Institut Universitaire de France.

Bernhard, Delphine, Ligozat, Anne-Laure, Martin, Fanny, Bras, Myriam, Magistry, et al., 2018, « Corpora with Part-of-Speech Annotations for Three Regional Languages of France: Alsatian, Occitan and Picard », *11th edition of the Language Resources and Evaluation Conference*, May 2018, Miyazaki, Japan.

Bernhard, Delphine, Todirascu, Amalia, Martin, Fanny, Erhart, Pascale, Steiblé, Lucie, Huck, Dominique, Rey, Christophe, 2017, « Problèmes de tokénisation pour deux langues régionales de France, l'alsacien et le picard », *Actes de DiLiTAL 2017*, pp. 14 23.

Martin, Fanny, Rey, Christophe, Reynés, Philippe, 2020, « Enseigner le picard au XXIème siècle : pour qui, comment ? », *Variation et enseignement des langues le cas des langues à faible diffusion*, Forlot, G. et Ouvrard, L. (dir.), Presses de l'Inalco, pp. 191-201.

Rey, Christophe, 2021, *La langue picarde et ses dictionnaires*, Collection Lexica, mots et dictionnaires, n°38, Honoré Champion.

[4]https://www.u-picardie.fr/LESCLaP/PICARTEXT/Public/

[5]Projet financé par l'Agence Nationale de la Recherche. https://restaure.unistra.fr/

[6]Projet financé par l'Agence Nationale de la Recherche. https://anr-appi.univ-lille.fr/

[7]https://languepicarde.fr/dictionnaire-fondamental-francais-picard/

[8]https://languepicarde.fr/commission-de-neologie-et-de-terminologie/

12h10-12h50, **Nicolas Quint (CNRS, LLACAN, UMR 8135, Villejuif)**

Les parlers du Croissant : un aperçu des actions actuelles de documentation et de promotion d'un patrimoine linguistique menacé

Traditionnellement pratiqués aux franges Nord du Massif Central, les parlers du Croissant comptent parmi les plus méconnues des variétés gallo-romanes. Située à la limite de trois grandes langues néo-latines (l'occitan ou langue d'oc, les langues d'oïl et le francoprovençal), l'aire du Croissant linguistique (ainsi nommé à cause de sa forme géographique évoquant une demi-lune) présente une multitude de parlers locaux, présentant simultanément des traits considérés comme caractéristiques de ces trois ensembles qu'ils jouxtent et prolongent à la fois. Du fait de leur caractère mixte et de la difficulté à les classifier, les parlers du Croissant ont fait l'objet de moins d'études que la plupart des langues régionales pratiquées aujourd'hui en France.

Or ces parlers, justement du fait de leur caractère mixte et intermédiaire, constituent précisément un patrimoine aussi précieux que riche : la diversité interne du Croissant est impressionnante et, en particulier au niveau de la phonologie et de la morphologie, des variétés pratiquées à seulement quelques kilomètres de distance présentent parfois des différences assez poussées pour justifier des descriptions ou des études séparées. À un peu plus de trois heures de route au Sud de Paris, il existe donc un foisonnement de variétés vernaculaires à la fois originales et quasi-inconnues de la communauté scientifique.

Cependant, le temps presse désormais pour étudier les parlers du Croissant, dont la plupart des locuteurs natifs ont plus de 70 ans. Prenant acte de cet enjeu, des projets successifs, regroupant locuteurs désireux de transmettre et linguistes intéressés, ont pris corps dans la seconde décennie du vingt-et-unième siècle, afin de documenter et de promouvoir ces variétés tant qu'il est encore possible de le faire.

Dans la présente communication, je présenterai tout d'abord les parlers du Croissant et montrerai au moyen de quelques exemples ce qui constitue l'originalité et l'attrait de ces variétés. Dans une seconde partie, après un rapide point sur l'histoire des recherches sur les parlers du Croissant, je donnerai le détail des principaux projets et actions mis en place sur lesdits parlers au cours de la dernière décennie ainsi que des acteurs impliqués dans ces projets. Enfin, dans une troisième partie, je montrerai comment les linguistes et les locuteurs participant à ces entreprises de sauvegarde des parlers du Croissant ont pu bénéficier des technologies récentes (informatique, Internet, enregistreurs numériques, audio-visuel) pour accroître leur efficacité et mieux pérenniser les données recueillies. Je conclurai sur les perspectives qui s'ouvrent quant à l'utilisation de ces données tant par les scientifiques que par les personnes résidant sur les territoires concernés ou qui s'y sentent attachées.

13h50-14h30, **Myriam Bras (Toulouse 2, CLLE, UMR 5263)**

Nouvelles perspectives pour la linguistique occitane à partir de la base textuelle BaTelòc

Je proposerai dans cette communication un bilan des avancées réalisées au cours des quinze dernières années en matière de ressources et outils pour la linguistique occitane, à partir de mon expérience de linguiste sémanticienne, pratiquant une linguistique sur corpus et progressivement initiée à la linguistique outillée et au traitement automatique des langues. Mon premier constat, alors que je souhaitais mener sur l'occitan des études dans le domaine de la sémantique temporelle, a été celui de l'absence de données textuelles facilement accessibles pour la linguistique descriptive de l'occitan (Bras 2005). Ce constat a motivé la création d'une base textuelle pour la langue occitane (Bras 2006), sur le modèle de la base Frantext. Un prototype de base textuelle pour l'occitan a vu le jour en 2008 (Bras et Thomas 2011), suivi d'une base opérationnelle, nommée BaTelÒc (Bras et Vergez-Couret 2016), mise en ligne en juin 2016 avec une centaine de textes couvrant tous les dialectes de l'occitan pour 3,7 millions de mots (redac.univ-tlse2.fr/bateloc/). Nous illustrerons l'utilisation de BaTelÒc pour la description sémantique en nous appuyant sur deux études en sémantique temporelle sur les temps simples et composés du passé et du futur en occitan (Bras et Sibille 2020, 2021).

La perspective d'enrichir les textes de BaTelòc avec des informations linguistiques afin d'améliorer les requêtes et par là-même l'accès aux données, nous a aussi amenés à créer des outils et des ressources pour le traitement automatique de l'occitan, que nous présenterons brièvement. Les premiers maillons de la chaine de traitement – segmenteur des textes en phrases et mots puis lexique de formes fléchies, analyseurs morpho-syntaxique et syntaxique – sont maintenant développés (Vergez-Couret et Urielli 2014, Miletic et al. 2019b), ainsi que les premiers corpus occitans annotés en parties du discours (Bernhard et al. 2018, Miletic et al. 2019a)) et en dépendances syntaxiques (Miletic et al. 2021 a, b).

Nous terminerons par un bilan des collaborations mises en place. La création de ces ressources et outils a bénéficié d'échanges avec des chercheurs travaillant sur d'autres langues peu dotées en France et en Europe (alsacien, picard, poitevin-saintongeais, basque, aragonais, serbe), avec le soutien financier de la Région Midi-Pyrénées et de l'UT2J, de la DGLFLF et de l'ANR dans le cadre du projet RESTAURE (2016-2018), et du fonds européen POCTEFA, dans le cadre du projet interreg LINGUATEC (2018-2021). Elle a également permis d'établir un partenariat crucial pour la linguistique occitane avec le Congrès Permanent de la Lenga Occitana, organisme associatif de régulation de la langue, avec qui nous avons commencé à travailler lorsqu'il a établi la feuille de route pour le développement numérique de la langue occitane en 2014. Des applications orientées vers le grand public ont été conçues par cet organisme dans le cadre du projet européen LINGUATEC sous la coordination du Congrès (clavier prédictif, traduction automatique, synthèse vocale, reconnaissance vocale).

Bernhard, D. et al. (2018). "*Corpora with Part-of-Speech Annotations for Three Regional Languages of France: Alsatian, Occitan and Picard", LREC, 7-12 de mai de 2018,* Miyazaki, Japon.

Bras, M. (2005). « A propos de quelques noms de temps en occitan », in I. Choi-Jonin, M. Bras, M. Rouquier, A. Dagnac (eds.) « Questions de classification en linguistique. Mélanges offerts au Professeur Christian Molinier », Peter Lang, Berne, pp. 55-80.

Bras, M. (2006). « Le projet TELOC : construction d'une base textuelle occitane », Langues et Cité : bulletin de l'observation des pratiques linguistiques, 8, p.9.

Bras, M., Thomas, J. (2011). « Bateloc : cap a una basa informatisada de tèxtes occitans », in A. Rieger & D. Sumien (eds). *L'Occitanie invitée de l'Euregio. Liège 1981 - Aix-la-Chapelle 2008 : Bilan et perspectives / Occitània convidada d'Euregio. Lièja 1981 - Aquisgran 2008 : Bilanç e amiras / Okzitanien zu Gast in der Euregio.Lüttich 1981 - Aachen 2008 : Bilanz und Perspektiven.* Actes du Neuvième Congrès International de l'Association Internationale d'Études Occitanes, Aix-la-Chapelle, 24-31 août 2008, Aache, Shaker, 2011.

Bras, M. & Vergez-Couret, M. (2016). « BaTelÒc: A text base for the Occitan language. », in Vera Ferreira and Peter Bouda (eds.) *Language Documentation and Conservation in Europe*, Honolulu: University of Hawai'i Press, pp. 133-149.

Bras, M., Vergez-Couret, M., Hathout, N., Sibille, J., Séguier, A., Dazéas, B. (2020). Lofloc : Lexic obèrt flechit occitan, *in* Jean-François Courouau / David Fabié (éds), *Fidelitats e dissidéncias. Actes del XIIn Congrès de*

l'Associacion internacionala d'estudis occitans. Actes du XIIe Congrès de l'Association internationales d'études occitanes. Albi 10-15/07/2017, Toulouse, SFAIEO, pp. 141-15.

Bras, M., Sibille, J. (2020). Lo Futur Perifrastic de tipe ANAR + Infinitiu en occitan, *in* Jean-François Courouau / David Fabié (éds), *Fidelitats e dissidéncias. Actes del XIIn Congrès de l'Associacion internacionala d'estudis occitans. Actes du XIIe Congrès de l'Association internationales d'études occitanes. Albi 10-15/07/2017*, Toulouse, SFAIEO, pp. 157-168.

Bras, M., Sibille, J. (2021). Preterit and perfect in Romance: new insights from Occitan, in Louis de Saussure and Laura Baranzini (eds.) *Aspects of tenses, modality and evidentiality* proceedings of the 13rd Chronos conference, Neuchâtel, Suisse, 4-6 juin 2018, Cahiers Chronos, Leiden : Koninklijke Brill NV.

Miletic, A., Bernhard, D., Bras, M., Ligozat, A., Vergez-Couret, M., (2019a). Transformation d'annotations en parties du discours et lemmes vers le format Universal Dependencies : étude de cas pour l'alsacien et l'occitan. In Morin, E., Rosset, S., Zweigenbaum, P., Ligozat, A.L., Ghannay, S. (Eds.) Actes de la Conférence sur le Traitement Automatique des Langues Naturelles (TALN-RECITAL) 2019, Toulouse (427-435).

Miletic, A., Bras, M., Esher, L., Sibille, J., Vergez-Couret, M. (2019b). Building a treebank for Occitan: what use for Romance UD Corpora?, In Gerdes, K., Kahane, S. (Eds.) Proceedings of the International Conference on Dependency Linguistics, SyntaxFest – Depling 2019, Paris, France.

Miletic, A., Bras, M., Vergez-Couret, M., Esher, L., Poujade, C., Sibille, J. (2020a). Building a Universal Dependencies Treebank for Occitan. Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 2932–2939, Marseille, 11–16 May 2020.

Miletic, A., Bras, M., Vergez-Couret, M., Esher, L., Poujade, C., Sibille, J. (2020b). A Four-Dialect Treebank for Occitan: Building Process and Parsing Experiments. In Proceedings of the 7th VarDial Workshop on NLP for Similar Languages, Varieties and Dialects, pages 140–149, Barcelona, Spain (Online), December 13, 2020.

Vergez-Couret, M. and Urieli, A. (2014). Pos-tagging different varieties of Occitan with single-dialect resources. In *Proceedings of the First Workshop on Applying NLP Tools to Similar Languages Varieties and Dialects*. Association for Computational Linguistics and Dublin City University.

14h30-15h10, **Hervé Lieutard (Montpellier 3, ReSo)**

Nouvelles approches linguistiques et lexicographiques de l'occitan médiéval

Ces dernières années de nombreux projets de recherche d'envergure sur l'occitan ont vu le jour grâce au développement du métalangage XML et en particulier à l'apport de la TEI qui a permis de repenser la nature des textes occitans médiévaux et d'en proposer de nouvelles éditions qui permettent d'envisager une interopérabilité de plus en grande pour les éditions de manuscrits médiévaux encodés dans ce format.

À partir d'un exemple concret, l'édition critique numérique du *Petit Thalamus* (http://thalamus.huma-num.fr), qui regroupe les livres du gouvernement du Consulat de Montpellier, nous présenterons les apports des nouvelles technologies dans le domaine de l'édition, notamment le fait qu'elles favorisent des approches transversales novatrices et qu'elles ouvrent de nouvelles perspectives de recherche pour les études linguistiques en diachronie sur l'histoire de la langue occitane. L'édition du *Petit Thalamus* permet d'ores et déjà d'interroger une base considérable de documents encodés en XML-TEI et d'en extraire des données massives et fiables qui permettent de conduire de nouvelles études linguistiques qui améliorent notre compréhension et notre connaissance de la langue occitane médiévale.

Au-delà du travail d'édition, ce projet ANR a aussi été l'occasion de réfléchir concrètement à la nécessaire question de la lemmatisation pour la gestion interne de la variation diachronique inhérente à un ensemble de documents médiévaux écrits sur plusieurs siècles. Grâce à la graphie classique occitane, le travail sur l'index des noms de personnes, réalisé à partir de plus de plusieurs milliers d'entrées, a permis de montrer que la graphie classique ne se bornait pas à proposer un système graphique cohérent, apte à gérer la variation dialectale en synchronie, mais que, du fait même de sa parenté génétique avec les usages médiévaux, il était tout aussi capable de gérer la variation lexicale en diachronie, en permettant de regrouper sous une forme lemmatisée stabilisée les formes orthographiques divergentes de l'occitan médiéval.

À l'issue du projet sur le *Petit Thalamus*, notre prise de conscience des potentialités ouvertes par les nouveaux formats d'édition pour connecter entre elles des masses considérables de données, a fait naître la volonté d'aller plus loin et, notamment, de relier tous les projets d'édition numérique en cours sur l'occitan ancien au sein d'un réseau international qui a vu le jour en 2018 sous le nom d'AcTo (*Aculhir e Tornar, Ressorsas numericas per l'occitan medieval*, https://acto.hypotheses.org). Ce réseau a pour but de faire converger les méthodes d'encodage entre tous les projets d'édition en cours et à venir, condition nécessaire pour envisager la création d'un espace numérique virtuel de recherche. Il a d'ores et déjà permis d'envisager la mise en place d'un premier projet de lemmatisation de l'occitan médiéval en lien avec le DOM, le *Dictionnaire d'occitan médiéval* (http://www.dom-en-ligne.de/) de la Bayerische Akademie der Wissenschaften. Dans ce nouveau projet, le *Petit Thalamus* servira de candidat privilégié pour la création de nouveaux outils de lemmatisation dont les applications devraient par la suite pouvoir s'étendre à d'autres documents médiévaux, mais devrait aussi au-delà pouvoir permettre de lier entre eux les divers projets concernant la lexicographie occitane.

- Hervé Lieutard, « Les apports récents et à venir du numérique pour la recherche en domaine occitan », *TENSO: Bulletin of the Société Guilhem IX*, 36, 2021, pp. 171-176
- Hervé Lieutard, « La grafia classica de l'occitan al servici de l'antroponimia medievala », Amb un fil d'amistat, Mélanges offerts à Philippe Gardy, Centre d'étude de la littérature occitane, Toulouse, 2014, pp. 667-678.
- Gilda CAITI-RUSSO, Jean-Baptiste CAMPS, Gilles COUFFIGNAL, Francesca FRONTINI, Hervé LIEUTARD, Elisabeth REICHLE, and Maria SELIG. 2019. 'AcTo: How to Build a Network of Integrated Projects for Medieval Occitan'. In Proceedings of the CLARIN Annual Conference 2019, edited by Kirl Simov and Maria Eskevich, 134–37. Leipzig.

15h30-16h10, **Patrick Sauzet (Toulouse 2, CLLE, UMR 5263)**

L'occitan « *pro-drop* or not *pro-drop* »: l'éclairage de la base de données en ligne SYMILA (http://symila.univ-tlse2.fr/)

Le projet SYMILA (http://symila.univ-tlse2.fr/) n'a pas pour but direct la revitalisation d'une ou plusieurs « langues régionales de France ». C'est un projet scientifique dont l'objectif est l'observation de la variation syntaxique fine dans des variétés romanes primaires parlées sur le territoire métropolitain continental de la République française. Les linguistes savent que « langue » est un mot très polysémique et ne s'accordent d'ailleurs pas sur les sens qu'on peut lui donner. On peut néanmoins distinguer des « langues-grammaires », systèmes qui réalisent la faculté de langage chez un individu ou dans un groupe dont les productions s'approchent de l'homogénéité (communauté réduite ou usager d'un standard codifié) et les « langues-cultures » qui intègrent dans un système d'échanges des locuteurs de lieux et d'époques différents, dont les production sont intercompréhensibles et interpertinentes. C'est en ce sens que les textes de Michel Houellebecq, Chrétien de Troyes, Paul Valéry et Maître Gims, les productions orales d'un parler cajun ou d'un point d'enquête de l'ALF entre la Loire et la Manche sont du « français » et font le « français ».

SYMILA est l'acronyme de « Syntactic microvariation in the Romance *languages* of France » : « dans les langues romanes de France » et non « dans les parlers » ou « dans les patois (gallo)romans ». Il y a dans ce choix une part d'affichage : il n'est pas sans intérêt qu'un projet consacré aux langues-grammaires soit aussi l'occasion de manifester les espaces respectifs de quelques unes des langues-cultures du territoire de la République et que l'on désigne officiellement comme « langues de France ». Le projet SYMILA documente ainsi la variation géographique primaire du français ou langue d'oïl, de l'occitan, du francoprovençal, et marginalement du catalan. Mais l'interaction entre langues-grammaires qu'il s'agit de décrire et langues-cultures ne s'arrête pas à cette manifestation. Le projet utilise les formes culturellement codifiées des langues (notations orthographiques) pour favoriser la lisibilité des formes phonétiques de l'ALF et comme vecteur de l'association à ces formes de l'information grammaticale explicite absent de l'ALF. À un troisième niveau, le projet permet d'explorer les corrélats grammaticaux présents dans les langues-grammaire abritées sous le toit des langues-cultures dons le domaine est traité par le projet. L'existence de la langue-culture « occitan » et de parlers occitans ne résulte pas de la décision de linguistes. L'occitan (diversement nommé dans son histoire) a émergé et s'est éprouvé culturellement, clairement au moyen-âge puis au XIXe siècle et dans la suite, de manière plus complexe entre ces deux époques. Les linguistes ont cherché à le délimiter parce qu'il existait, et non l'inverse. Une manière de concevoir la contrepartie grammaticale de la langue-culture qu'est l'occitan est (selon le modèle proposé par Jules Ronjat) de caractériser par une liste de propriétés un occitan prototypique, plutôt que de vouloir trouver des propriétés que tous parlers occitans sans exception possèderaient. Le caractère *pro-drop* fait assurément partie des traits qui caractérisent globalement l'occitan qui prototypiquement oppose « Canta » (« Chanta », « Que canta. ») au français « Il chante » (« Il cante », « Al chante », « Le chante »). Toutefois, ce trait prototypique n'est pas présent dans toutes les variétés, ni géographiques, ni historiques de la langue. Géographiquement, le nord ouest de l'occitan (haut limousin et marchois) présente régulièrement des pronoms sujets. Historiquement, l'occitan de la fin du moyen-âge et de l'âge baroque (XVIe-XVIIe) présente un emploi massif des pronoms sujets qui ne se laisse pas réduire à une imitation passagère du français.

La base SYMILA permet d'utiliser les données de l'ALF qui aident à caractériser le phénomène. En zone à sujet nul on voit des pronoms sujets sporadiques. Ils sont typiquement non clitiques : « n(os)autres ». Des parlers occitans extrêmes (aux confins du « Croissant ») présentent des pronoms sujets redoublant un sujet lexical (à fonction affixale) que les parlers d'oïl voisins n'ont pas. Plus au sud, des parlers présentent des formes pronominales clairement faibles (sans voyelle ou avec schwa). Enfin, au contact de la zone régulièrement à sujet nul, on trouve des formes plus pleines. Cela suggère un processus évolutif non déterministe où la grammaire à verbe second (V2) de l'occitan médiéval favorise l'emploi massif de pronoms sujets. Cet emploi se dissocie ensuite de l'ordre V2 et se résout soit en cliticisation de formes pronominales sujets, soit par retour à un fonctionnement à sujet

nul. Un corrélat remarquable de cette évolution divergente est la généralisation des formes pronominales toniques accusatives par les parlers à pronoms sujets clitiques (occitan du nord-ouest : « *me* i chante per *me* », français « *moi* je / i chante pour *moi*» ») et des formes nominatives dans les parlers occitans à sujet nul (occitan commun « *ieu* canti per *ieu*, *io* chante per *io*, *jo* que canti per *jo* »).

Champclaux, Yaël Sauzet, Patrick 2020 SYMILA, un système d'information linguistique pour l'étude des micro-variations syntaxiques dans les langues romanes de France in Sibille éd. 2020
Courouau, Jean-François, Fabié David eds 2020 *Fidelitats e dissidéncias* (Actes del XIIn Congrès de l'Associacion internacionala d'estudis occitans, Albi, 10-15 juillet 2017), 959 p (2 vol.)
Dagnac, Anne 2020 Introduction : vers une microsyntaxe galloromane in Sibille éd. 2020
Maiden, Martin, Ledgeway, Adam eds 2016 *The Oxford guide to the Romance languages,* Oxford ; New York : Oxford University Press, LIV-1193 p.
Michèle Oliviéri, Georg A. Kaiser, Katerina Palasis, Michael Zimmermann, Richard Faure 2020 Quand la dialectologie, la diachronie et l'acquisition se parlent : Étude comparative des pronoms sujets en occitan et en français, in Sibille 2020
Oliviéri, Michèle, Sauzet, Patrick 2016 Southern Gallo-Romance : Occitan in Maiden et Ledgeway eds 2016
Ronjat, Jules 1930-1941 *Grammaire (h)istorique des parlers provençaux* (= occitans) *modernes*, Montpellier : Société des Langues Romanes, 4 vol.
Sauzet, Patrick 2020 Dissidéncia o dòxa dins los estudis occitans *in* Courouau et Fabié 2020, 41-74.
Sauzet, Patrick, Champclaux, Yaël 2020 SYMILA : Una basa de donadas sintaxicas per l'occitan e las autras lengas romanicas de França, *in* Courouau et Fabié 2020, 271-284
Sibille, Jean ed. 2020 *La microvariation syntaxique dans les langues romanes de France* : actes du colloque SYMILA, Toulouse, 11 et 12 juin 2015, Limoges : Lambert Lucas, 194 p.

16h10-16h50, **Jonathan Kasstan (University of Westminster, UK) & Michela Russo (Lyon 3/SFL UMR 7023 CNRS/Paris 8)**

Maintenance *in* shift: on nasalisation in transitional Francoprovençal and Occitan areas

This paper presents a diachronic and synchronic account of nasalisation in two regional minoritised languages in contact: Francoprovençal and Occitan. We marshal datasets from linguistic atlases dating back to the turn of the century (Gilliéron & Edmont 1902-1910; Gardette 1950-1956; Nauton 1957-1963; Tuaillon & Martin 1971-1981) which we compare with more recent radio and interview recordings from speakers in (the former) Rhône-Alpes, Haute-Loire Ardèche, Drôme regions of France, presenting us with over 100 years of time depth.

Nasalisation in Francoprovençal applies to all contexts where French has nasal vowels. However, Francoprovençal has a comparatively more complex vocalic inventory, including high nasal vowels [ũ] and [ĩ], a feature preserved from Old French. Historical atlas data present ample evidence to suggest that in Francoprovençal we find nasalised vowels from a sequence of underlying /i/ + nasal coda. Conversely, in Central Occitan, we find oral vowels; this is also observed in transitional zones, such as Auvergne, where Francoprovençal and Occitan are in contact with one another, and with French.

We can posit two phases for the nasalisation of /i + N/ in Francoprovençal: a first phase with a nasalised [ĩ] quality, which is still preserved along the Francoprovençal/Occitan border, followed by a second phase with the mid-open [ɛ̃] quality, evidence for which can be gleaned from existing ALLy data (e.g., maps n° 438 'pin/pine', n° 923 'matin/morning', n° 837 'chemin/path'). In the Loire, the forms are nasalised in [ĩ] or in [ɛ̃] until data pts. 34 (Saint-Marcel-d'Urfé), 48 (Essertines-en-Chatelneuf), 55 (Sury), 61 (La Valla), 66 (Roizey), indicating an isogloss. At data pts. 62 (Sainte-Croix) and 54 (Saint-Bonnet-les-Oules) we can observe forms such as [madzi] 'maison/house' and [t͡ʃami] 'chemin/path'; pt. 55 (Sury) has [t͡ʃami] without nasalisation as expected in Occitan.

In the northern Occitan speaking region of the Haute-Loire, nasalisation is also present in the historical record (Nauton 1974). All /m/ and /n/ consonants in coda position trigger nasalisation in the preceding vowel: [tsɑ̃]/[tsõ] 'champs/fields' < CAMPUS. However, Latin intervocalic /m/ and /n/ consonants that later became final and lenited do not undergo the same process. While the /m/ nasalised any preceding vowel, as in [fũ] 'fumée/smoke' < FUMU or [fõ] 'feu/fire' < FAME, the final /n/ fell without nasalising the preceding vowel, as in [ple] 'plein/full' < PLENU, [vi] 'vin/wine' < VINU. We therefore observe variation in the system. Furthermore, we also observe that nasalisation is triggered when a lexical item is inserted in a constituent before a consonantal onset, e.g. [ɛ̃ plɛ̃ sa] 'a full bag'.

Owing to methodological issues associated with traditional dialectological (representativity, low numbers of data points etc.), we augment these historical accounts with exploratory acoustic and distributional analyses of Francoprovençal and Occitan from several sources. First, we present an acoustic analysis of nasalisation in recordings obtained from departmental archives, focusing in particular on the Haute-Loire. These materials include recordings of Occitan gathered from the Protestant area of Chambon-sur-Lignon (*commune* of Tence) by the dialectologist Théodore de Félice; these recordings are also supplemented with local radio broadcasts. Second, we present distributional analysis of nasalisation in recordings of semi-structured sociolinguistic interviews and wordlist elicitation tasks (n=16 speakers) gathered in the Lyonnais mountains (*communes* of Saint-Martin-en-Haut and Saint-Symphorien-sur-Coise).

In combining datasets in this way, we are able to consider internal and external constraints operating on nasalisation. We propose an analysis of the sound spectra of nasal vowels in various contexts, the acoustic characteristics of nasal vowels in terms of formant and anti-formant distribution, the articulatory features and the acoustical interpretation, acoustic cues for nasal consonants after a vowel. This analysis will be combined where possible with available social factors (age, sex, region) to provide a more complete picture of nasalisation in this Francoprovençal/Occitan-speaking border region.

Triangulating the data in this way presents some evidence for relative stability in the nasal vowel systems of these varieties, in spite of widespread accounts of phonological levelling in cases of language obsolescence (e.g., Dorian 1989, Jones 2000).

ALF = Gilliéron, J. & E. Edmont. 1902-1910. *Atlas linguistique de la France*. Paris: Champion.
ALJA = Tuaillon, G. & J.-B. Martin. 1971-1981. *Atlas linguistique et ethnographique du Jura et des Alpes du Nord* (*Francoprovençal central*). Paris: Éditions du C.N.R.S.
ALLy = Gardette, Pierre. 1950-1976. *Atlas linguistique et ethnographique du Lyonnais*. 5 vol. Lyon/Paris : Institut de linguistique romane des facultés catholiques/Editions du CNRS.
Dorian, Nancy C. (ed). 1989. *Investigating Obsolescence*. Cambridge: Cambridge University Press.
Jones, Mari C. 2000. *Jersey Norman French: A Linguistic Study of an Obsolescence Dialect*. Oxford: Blackwell.
Nauton, Pierre. 1974. *Geographie phonetique de la Haute-Loire*. Paris : Les Belles Lettres.
Nauton, Pierre. 1957-1963. *Atlas linguistique et ethnographique du Massif Central*. Paris : CNRS.
Gardette, Pierre. 1973. « Frontieres linguistiques et limites interieures en lyonnais d'apres l'Ally ». In Georges Straka & Pierre Gardette *Les dialectes romans de France*. 141-171. Paris : CNRS.

16h50-17h30, **Philippe Boula de Mareüil (CNRS, LISN, Orsay), Marcel Courthiade (INALCO, Paris), Frédéric Vernier (CNRS, LISN, Orsay)**

De la Provence aux Balkans : discours épilinguistiques autour d'un atlas sonore des langues régionales ou minoritaires d'Europe

Souvent, les dialectes et langues minoritaires doivent faire face à une double minoration, de la part de leurs locuteurs mêmes, tenants d'un immobilisme hostile à toute évolution et de la part de locuteurs des langues dominantes. Du côté de ces derniers, en France, on se heurte de plus aux sacrosaints deux premiers articles de la Constitution de la Ve République pour promouvoir la diversité linguistique, au nom de l'indivisibilité de la société française (Viaut & Pascaud, 2017). Pour contourner le problème, on assiste à une déterritorialisation de la question : « le vrai territoire d'une langue est le cerveau de ceux qui la parlent » (Cerquiglini, 1999). Même si telle n'était pas l'intention de l'auteur, cette formule a pu contribuer à l'invisibilisation des langues régionales de France dans l'espace public, laissant le français seule langue de la République.

Pour au contraire rendre visible et valoriser la diversité linguistique, nous avons mis au point un atlas sonore de langues régionales ou minoritaires qui, partant de la France hexagonale (Boula de Mareüil *et al*., 2017), a été étendu aux Outre-mer, aux langues non-territoriales (Cerquiglini, 1999) ou sans territoire compact (Courthiade & Karamagkiola, 2013) comme le rromani, ainsi qu'à d'autres pays dans le voisinage immédiat de la France. Cet atlas en ligne <https://atlas.limsi.fr> permet aux visiteurs d'écouter et de lire la même fable d'Ésope dans plus de 700 versions : ce travail est le fruit de nombreuses enquêtes de terrain et du développement d'une interface attractive. Le confinement ne se prêtant guère à la linguistique de terrain, lors de la quarantaine de 2020 nous avons entrepris de collecter une quarantaine de traductions de cette fable, via Internet, dans les langues/dialectes minoritaires d'Europe — la distinction entre langues et dialectes n'étant pas univoque.

Nous décrirons leur cartographie et nous nous concentrerons sur certaines des langues en danger que nous avons collectées, provenant de diverses aires linguistiques : romane (occitan, aroumain et moéso-roumain), finno-ougrienne (sámi, meänkieli et kven) et slave (ruthène, morave et bunjevac). Nous verrons que ces langues soulèvent des questions communes et controversées, du fait de leur obsolescence. L'hétérogénéité de ces langues, presque consubstantielle à leur état minoritaire, alimente ainsi un discours épilinguistique puriste, fixiste et essentialiste : « on ne dit pas ça comme ça » ou, entre deux variétés de ces langues, « ils [les autres] ne parlent pas comme nous ». Dans ces conditions, le passage à l'écrit des langues minoritaires, cruciale pour leur documentation et leur survie, soulève des questions importantes, qui sont partagées par les langues sélectionnées ici. Différentes solutions proposées seront discutées, continuant à faire débat, notamment dans le domaine d'oc (provençal et languedocien oriental, dont nous partirons et que nous analyserons en détail) ainsi que dans les Balkans — avec .le macédonien de Golo Brdo (Albanie), le goran (Kosovo) et le roumain de Transylvanie.

Boula de Mareüil, P., Vernier, F., Rilliard, A. (2017), « Enregistrements et transcriptions pour un atlas sonore des langues régionales de France », *Géolinguistique* 17 : 23–48.

Cerquiglini, B. (1999), Rapport au Ministre de l'Éducation Nationale, de la Recherche et de la Technologie, et à la Ministre de la Culture et de la Communication. <http ://www. ladocumentationfrancaise.fr/var/storage/rapports-publics/994000719.pdf>.

Courthiade, M. & Karamagkiola, S. (2013), « Attitudes comparées de deux minorités européennes sans territoire compact vis-à-vis de la langue maternelle : les Rroms et les Aroumains », *in* Alén Garabato, C. (Dir.), *Gestion des minorités linguistiques dans l'Europe du XXIe siècle*, Éditions Lambert-Lucas, Limoges (pp. 193–215).

Viaut, A. & Pascaud, A. (2017), « Pour une définition de la notion de "langue régionale" », *Lengas* 82 : en ligne.