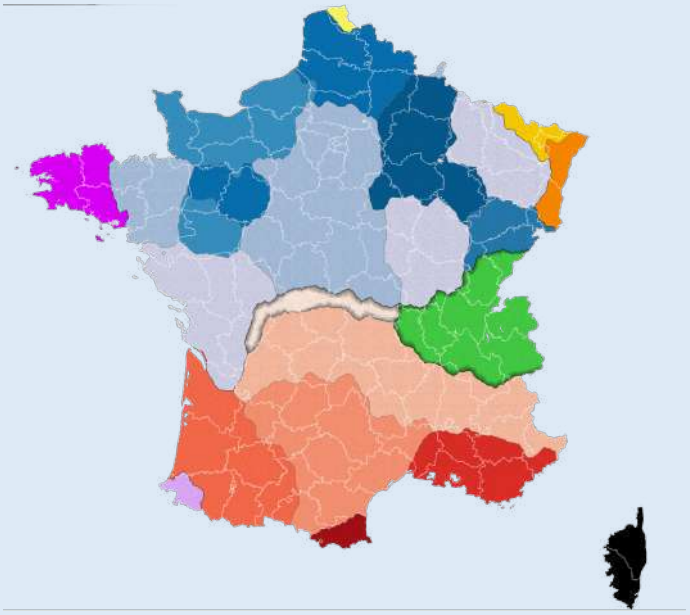
Annie Rialland & Michela Russo (dir.)

Les langues régionales de France

Nouvelles approches, nouvelles méthodologies, revitalisation



Editions de la Société de Linguistique de Paris

Annie Rialland & Michela Russo (dir.)
Les langues régionales de France : nouvelles approches, nouvelles méthodologies, revitalisation

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

sous la direction de

Annie Rialland & Michela Russo

Editions de la Société de Linguistique de Paris

Couverture : fond de la carte de l'Atlas sonore des langues régionales de France, par Frédéric Vernier, Philippe Boula de Mareuïl & Albert Rilliard (https://atlas.limsi.fr/)

EPHE, 4ème section, 45-47 rue des Ecoles, 75005-Paris
Imprimé par BOD GmbH (Allemagne)
ISBN : 978-2-9570894-2-0 (version reliée)
Dépôt légal : Juin 2023
ISBN : 978-2-9570894-3-7 (version digitale)

Introduction

Évolution du statut des langues régionales et de leur environnement numérique

Annie Rialland[1] **& Michela Russo**[2]

[1] Laboratoire de Phonétique et Phonologie, CNRS/Sorbonne-Nouvelle,
[2] Université Jean Moulin, Lyon 3 & Structures Formelles du langage, CNRS/ Paris 8)

Ce volume est issu de la Journée Scientifique de la Société de Linguistique de Paris 'Les langues régionales de France : nouvelles approches, nouvelles méthodologies, revitalisation' qui a eu lieu le 12 juin 2021. Il ne se propose pas de couvrir l'ensemble des langues régionales de France – mais de montrer comment les recherches sur les langues régionales progressent, quels rôles elles peuvent jouer dans le devenir de ces langues.

Les recherches sur les langues régionales, de même que les recherches sur les langues en général, s'inscrivent dans des contextes plus larges, qu'ils soient sociaux ou politiques ou qu'ils relèvent des évolutions technologiques.

Cette introduction aborde en premier lieu le statut des langues régionales dans les institutions et son évolution avec ses reculs et ses avancées jusqu'à la situation actuelle suite à la promulgation de la loi Molac en 2021 (Sections § 1.1 et 1.2). Les réticences institutionnelles face aux langues régionales ont été importantes, comme l'illustrent, par exemple les freins mis à la signature et encore actuellement à la ratification de la Charte Européenne des langues régionales ou minoritaires adoptée en 1992 par le Conseil de

l'Europe.

La Section § 1.3 introduit au contexte actuel en termes de nouvelles technologies. Le développement de ressources et outils numériques pour les langues régionales a fait évoluer leur étude et constitue un enjeu majeur pour leur vitalité et, le cas échéant, leur revitalisation.

Nous présentons ensuite les 7 chapitres de ce livre, qui, portant sur diverses langues régionales, se proposent de répondre aux questions suivantes : en quoi et comment les études des langues régionales évoluent-elles ? En quoi les nouvelles technologies (outils du traitement automatique des langues, développement de bases de données informatisées, technologies du web) 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 et comment la recherche avec ses outils actuels peut-elle l'influencer ?

Nous ajoutons un résumé de trois interventions à la Journée Scientifique du 12 juin 2021 pour inclure dans cet ouvrage l'ensemble des perspectives représentées ce jour-là.

1. Statut des langues régionales

1.1 Évolution du statut des langues régionales à partir de la loi Deixonne (1951)

La reconnaissance et l'enseignement des langues régionales ont connu des avancées et des reculs mais ils se sont souvent heurtés à des réticences et résistances institutionnelles.

La loi Deixonne, promulguée en 1951 est particulièrement importante, apportant pour la première fois une reconnaissance officielle des langues régionales et les faisant entrer dans l'enseignement. Ses deux premiers articles sont les suivants :

Article 1. Le Conseil supérieur de l'Éducation nationale sera chargé, dans le cadre et dès la promulgation de la présente loi, de rechercher les meilleurs moyens de favoriser l'étude des langues et dialectes locaux dans les régions où ils sont en usage.

Article 2. Des instructions pédagogiques seront adressées aux recteurs en vue d'autoriser les maîtres à recourir aux parlers locaux dans les écoles primaires et maternelles chaque fois qu'ils pourront

en tirer profit pour leur enseignement, notamment pour l'étude de la langue française.

Les articles suivants concernent les modalités d'ouverture d'enseignement en langues régionales (en tant que matière facultative dans le cursus des élèves), la mise en place d'une épreuve facultative au baccalauréat, l'ouverture de certificats de licences dans l'enseignement supérieur et la possibilité de créer d' ' instituts d'études régionalistes '.

La loi Deixonne prévoyait une application immédiate à l'occitan, le catalan, le basque et le breton. Elle a été ensuite étendue progressivement à d'autres langues régionales[1]. Elle a été suivie de nombreux textes législatifs plus ou moins favorables aux langues régionales selon la politique des gouvernements successifs.

Sous la présidence de Mitterrand, les circulaires Savary (1982, 1983) ont apporté des améliorations notables pour les langues régionales au sein de l'Éducation nationale, ouvrant, en particulier la possibilité de création de classes bilingues français-langue régionale dans l'enseignement public. Elles s'inscrivaient dans un climat plus favorable aux langues régionales et au minorités. En 1974, François Mitterrand qui avait perdu de peu la présidentielle contre Giscard d'Estaing, avait déjà été qualifié de "président des occitans" (Kremnitz 2020 : 38), ayant été soutenu par une majorité d'électeurs dans les départements occitans et catalan. Dans sa campagne électorale de 1981, François Mitterrand défendait le 'droit à la différence' en Bretagne, concepts abordés également dans le rapport Giordan (1982)[2].

Les ambiguïtés et réticences institutionnelles à propos des langues régionales sont particulièrement bien illustrées par les divers reports de la signature de La *Charte* européenne des langues régionales ou minoritaires du Conseil de l'Europe qui n'a toujours pas été ratifiée.

Cette charte a pour but de promouvoir et protéger les langues régionales dans le respect de la diversité culturelle. Elle est adoptée en 1992 par le Conseil de l'Europe, qui invite les États à ratifier la

[1] Au tahitien en 1981, à 4 langues mélanésiennes en 1992, aux créoles en 2000 (voir Bertile 2020 : 119).

[2] A l' initiative du gouvernement et du ministre de la culture Jack Lang. En 1981 un ouvrage intitulé *La France au pluriel* rédigé par le Parti Socialiste de France (collection minorités) chez Entente

Charte en novembre 1992. En 1994, le parlement européen vote une première résolution sur les minorités linguistiques et culturelles[3] rappelant aux gouvernements qu'il faut promouvoir 'une culture linguistique européenne' en vue de 'la défense du patrimoine linguistique, l'élimination des entraves linguistiques, la promotion des langues de moindre diffusion et la sauvegarde des langues minoritaires'. Elle sera suivie de beaucoup d'autres.

Entre 1997 et 2002, le gouvernement de Lionel Jospin, en co-habitation avec le Président Jacques Chirac, accorde une attention particulière à la politique linguistique. La 'chasse au patois' de la III[e] République n'était pas très loin et la loi Deixonne de 1951 ainsi que les textes légaux subséquents ne sont que partiellement appliqués et ne reconnaissent qu'une partie des langues régionales (Kremnitz 2020 ; Rouquier & Russo 2021).

Le gouvernement de Lionel Jospin, une fois en place en 1997, souhaite faire avancer la question de la signature et de la ratification de la Charte mais rencontre de multiples oppositions (Kremnitz 2020)[4]. Par conséquent, il doit préparer le terrain en commandant plusieurs rapports dont l'un est confié au député breton Bernard Poignant, chargé d'examiner les implications politiques de la ratification de la charte. Un rapport juridique est également confié à un juriste, expert de droit constitutionnel, Guy Carcasonne. Dans ce contexte, deux ministres, Claude Allègre (Éducation nationale) et Catherine Trautmann (Culture et Communication) confient au linguiste Bernard Cerquiglini, alors délégué général à la langue française et directeur de l'Institut national de la langue française, la mission d'établir une liste des langues régionales et minoritaires répondant aux critères de la Charte européenne.

La Charte Européenne entend comme 'langues régionales ou minoritaires', '[les langues] pratiquées sur un territoire d'un État par des ressortissants de cet État qui constituent un groupe numériquement inférieur au reste de la population de cet état et différentes de (la) ou (des) langues officielles de cet État.' 'Elle n'inclut ni les dialectes de (la) ou (les) langues officielles de cet état

[3] JO C 61 du 28.2.1994, p. 110.

[4] Oppositions de la gauche avancées par souci de 'souverainisme' et 'égalité', puis de la droite à laquelle appartient aussi Jacques Chirac, une opposition qui se soucie peu des conséquences sur la politique internationale (ib.).

ni les langues des migrants.' Les 'langues dépourvues de territoire' peuvent être prises en compte.

Le rapport de Bernard Cerquiglini, rendu en 1999 et intitulé 'Les langues de la France'[5], recense 75 'langues parlées par des ressortissants français sur le territoire de la République', 21 pratiquées en 'France métropolitaine', 15 en usage dans les départements d'outre-mer, 39 dans les 'territoires d'outre-mer' (Bertile 2020)[6].

Dans sa définition d'une langue régionale par rapport à un dialecte, le rapport de Bernard Cerquiglini s'appuie sur des notions d'écarts linguistiques et chronologiques. Celles-ci lui permettent de reconnaître comme langues régionales huit langues d'oïl (franc-comtois, wallon, picard, normand, gallo, poitevin-saintongeais, bourguignon-morvandiau, lorrain) à côté du basque, du breton, du catalan, du corse, du flamand occidental, du francoprovençal, de l'occitan parlées en France métropolitaine. Par contre, l'occitan est considéré comme une seule langue régionale avec plusieurs dialectes.

L'ensemble 'Langues de la France' englobe aussi les langues régionales d'outre-mer ainsi qu'un certain nombre de langues minoritaires 'sans territoire'[7].

Parmi les 'langues sans territoire', on compte le yiddish, le judéo-espagnol et le romani. Dans la même catégorie, Cerquiglini a inclus, sur la base de la citoyenneté, quelques langues de l'immigration : l'arabe dialectal, le berbère ou la langue des Hmong installés en Guyane (locuteurs originaires du Laos), toutes sans statut officiel dans leur pays d'origine en 1999[8]. La liste de 1999 inclut donc 75 langues auxquelles on a ajouté ultérieurement : la langue des signes française et le judéo-espagnol en 2002, le ligurien des localités à la frontière avec l'Italie en 2003.

En mai 1999, Pierre Moscovici, qui était ministre des affaires

5 Notons que Cerquiglini utilise le terme ' Les langues de la France ', employé par le ministre de l'éducation et qui rentre ainsi dans la politique linguistique.

6 Bertile (2020: 119) ' les langues autres que le français pratiquées dans les collectivités d'outre-mer relèvent de la catégorie des langues régionales '.

7 A noter que cette conception des 'Langues de la France' a été adoptée par la DGLFLF (Kremnitz 2008 ; Sibille 2013).

8 Exception faite pour l'allemand en Alsace, le néerlandais et l'italo-roman, il n'intègre pas dans la liste des langues qui ont un statut de langues officielles dans un autre état. Cela dit la situation a changé par exemple pour le berbère après 1999, puisque cette langue jouit désormais d'une reconnaissance en tant que langues officielles au Maroc et en Algérie.

européennes signe la *Charte*. Cependant Jacques Chirac, alors Président de la République, s'adresse au Conseil constitutionnel pour demander l'examen de sa compatibilité avec la Constitution, et malgré les rapports favorables préalablement établis, la réponse est négative et la Charte n'est pas ratifiée.

Selon le Conseil Constitutionnel, la *Charte* contredit l'article 1 de la Constitution : 'La France est une République indivisible, laïque, démocratique et sociale'. Une révision constitutionnelle de 1992 avait de plus ajouté l'alinéa 1er de l'article 2 de la Constitution : 'La langue de la République est le français'. Cet alinéa proposé pour défendre la langue française contre l'avancée de la langue anglaise a eu pour conséquence des reculs pour les autres langues parlées sur le territoire de la République. De ce fait, la Charte contredit aussi l'article 2, modifié au profit du français. Elle n'a pas pu être ratifiée et les langues régionales se sont trouvées à nouveau menacées[9].

Sous le gouvernement de Lionel Jospin (1997-2002), d'autres mesures seront prises pour les langues de France ; après la création du CAPES d'occitan en 1992, un CAPES de créole est créé en 2001.

Après le gouvernement de Jospin, la politique pro langues de France devient moins active. Bernard Cerquiglini est resté *Délégué général à la langue française et aux langues de France* de 2001 à 2004, puis il a été remplacé par Xavier North de 2004 à 2014, avec des moyens limités en faveur des langues de France. Sous le président Nicolas Sarkozy, il y a peu d'activité législative concernant les langues régionales et leur enseignement.

Le débat autour de la Charte européenne des langues régionales s'est cependant poursuivi. L'article 75-1 a été ajouté à la Constitution Française : 'Les langues régionales appartiennent au patrimoine de la France ' (révision constitutionnelle du 23 juillet 2008), mais sans que le gouvernement ne revienne sur l'article 2 de la constitution et n'accepte d'inscrire une référence à la Charte européenne. La France défend un monolinguisme officiel depuis au moins la Révolution française et c'est une position encore largement répandue comme l'indique, par exemple, la demande de retrait de cet article par l'Académie Française[10].

La *Charte,* vingt ans après, n'était toujours pas ratifiée. La liste

[9] Voir pour une discussion détaillée Martel & Verny (2020).
[10] https://www.academie-francaise.fr/actualites/la-langue-de-la-republique-est-le-francais

de Cerquiglini a été critiquée, en particulier comme étant trop complexe alors même qu'elle représentait une reconnaissance de la complexité linguistique de la France, en particulier par rapport à l'article 2 de la constitution 'La langue de la République est le français'.

En 2012, la ratification de la *Charte* faisait partie des 60 engagements de François Hollande (Kremnitz 2020 : 43), mais il y renonce un peu avant son élection. En effet, la ratification de la *Charte* implique la modification de la constitution[11].

Depuis, on compte plusieurs propositions de loi sur le statut des langues régionales, enseignement, etc., entre autres celle du député Paul Molac à l'assemblée en 2015[12], ou celle déposée la même année par Philippe Bas au Sénat (qui n'a pas fait l'objet d'un débat en commission)[13], avant la nouvelle déposition de proposition de loi de Paul Molac en 2020, qui a été adoptée par l'Assemblée le 8 Avril 2021, puis promulguée le 21 mai 2021 dépourvue d'une partie de ses articles (voir § Section 1.2)[14].

L'appellation 'Langues de France' est en outre désormais utilisée dans le cadre de l'agrégation, créée en 2017 (qui concerne le breton, le basque, le catalan, l'occitan-langue d'oc, le corse, le créole, le tahitien).

Par ailleurs, dans l'attente de la ratification de la Charte Européenne par l'État, des pouvoirs locaux – régions ou communes d'Alsace et du Pays basque - ont signé depuis 2014 des chartes locales contenant des dispositions choisies parmi celle de la Charte du Conseil de l'Europe[15]. Celles-ci concernent principalement la reconnaissance de la langue, son enseignement, l'élimination des discriminations envers les locuteurs.

[11] La (première) ministre de la Culture et de la Communication Aurélie Filipetti s'intéresse à nouveau à ce sujet, toutefois, elle n'a pas de soutien.
[12] https://www.assemblee-nationale.fr/14/propositions/pion3288.asp
[13] http://www.senat.fr/dossier-legislatif/ppl15-096.html
[14] https://www.assemblee-nationale.fr/dyn/15/dossiers/protection_patrimoniale_langues_regionales
[15] https://www.coe.int/fr/web/european-charter-regional-or-minority-languages/promoting-ratification-in-france

1.2 La loi Molac et les langues régionales dans la politique récente

La Journée Scientifique de la Société de Linguistique de Paris, dont est issu cet ouvrage, a eu lieu le 12 juin 2021 au moment même où les langues régionales se sont trouvées projetées au premier plan de l'actualité politique et médiatique. Rappelons tout d'abord les circonstances.

Le 8 Avril 2021, la proposition de loi relative à 'la protection patrimoniale des langues régionales et à leur promotion', dite 'loi Molac', du nom du député breton Paul Molac, était adoptée par l'Assemblée Nationale à une large majorité par 247 voix pour 76 voix contre. Elle comportait une reconnaissance de l'enseignement immersif des langues régionales, qu'on aurait donc pu croire acquise.

L'enseignement immersif procure un 'bain' linguistique à l'élève qui vient compléter l'exposition qu'il peut avoir à la langue régionale dans son milieu familial. Celle-ci est d'ailleurs très variable en fonction des familles et des régions, ce qui rend l'enseignement immersif d'autant plus important. L'enseignement immersif diversifie l'usage de la langue qui est utilisée non seulement lors de l'apprentissage scolaire mais aussi en dehors de la classe, dans les récréations, les repas ou les activités extra-scolaires. L'enjeu est de garder vivante la langue régionale.

Une pédagogie immersive pour les langues régionales a été mise en œuvre par un ensemble d'établissements associatifs (écoles maternelles, écoles primaires, collèges ou lycées). Ceux-ci se répartissent comme suit à la rentrée 2021 : 56 établissements Diwan pour le breton (4034 élèves), 38 établissements Seaska pour le basque (4064 élèves), 2 établissements Scola Corsa pour le corse (25 élèves), 9 établissements La Bressola pour le catalan (1048 élèves), 71 établissements Calandreta pour l'occitan (3906 élèves), et 12 établissements ABCM-Zweisprachigkeit pour l'alsacien et l'allemand (1230 élèves). À la rentrée 2021, ces établissements regroupaient environ 14120 élèves, ce qui ne représentait qu'une infime proportion des 12 millions d'élèves scolarisés de la maternelle au lycée en France Métropolitaine la même année[16].

[16] Les langues régionales de France reconnues par la circulaire officielle du 16 décembre 2021 sont pour la France Métropolitaine : le basque, le breton, le catalan, le corse, le gallo, l'occitan-langue d'oc, les langues régionales d'Alsace, les langues régionales des pays mosellans, le

Dans ces établissements, l'immersion se pratique surtout aux niveaux pré-élémentaire ou élémentaire et le français n'est pas pour autant négligé comme l'attestent les bons résultats scolaires de leurs élèves.

L'article 4 de la proposition de loi du 8 Avril 2021, en fait central, consistait en l'ajout des termes 'dans le respect des objectifs de maîtrise des deux langues à chaque niveau d'enseignement' à la phrase suivante de l'article L312-10 du code de l'éducation : ' Un enseignement de langues et cultures régionales peut être dispensé tout au long de la scolarité.' L'ajout de ces quelques mots était fondamental puisqu'il permettait 'de poser le principe de la reconnaissance de l'enseignement bilingue français-langues régionales quelle que soit la durée des enseignements dispensés dans ces deux langues, dans le respect des objectifs de maîtrise de la langue française fixés par le code de l'éducation'.

Cette proposition de loi prévoyait également la mise en place d'accords entre communes pour le versement du forfait communal à une école privée d'une autre commune en cas d'absence d'enseignement de langues régionales dans la commune de l'élève. Ces mesures avaient pour but de permettre un meilleur financement des écoles immersives.

Dans un autre article, elle autorisait les signes diacritiques des langues régionales dans les Actes d'État Civil, par exemple, le tilde du prénom breton Fañch.

Cette proposition de loi ayant été votée par une large majorité de députés, on aurait pu s'attendre à ce qu'elle soit entérinée sans beaucoup d'encombres mais ce n'est pas ce qui s'est passé. Un recours auprès du Conseil Constitutionnel concernant la constitutionalité du mode de financement (forfait communal) est porté par une soixantaine de députés le 22 avril 2021.

Le 21 mai 2021, le Conseil Constitutionnel rend ses décisions, en particulier les décisions **20, 22** et **23** reproduites ci-dessous :

---

francoprovençal, le flamand occidental, le picard et pour les départements d'Outre-Mer, le tahitien, les langues mélanésiennes (drehu, nengone, paicî, ajië), le wallisien, le futunien, le kibushi et le shimaoré. Leur statut, leur vitalité ainsi que leur enseignement sont très variables. L'immersion est plus ou moins nécessaire en fonction de la vitalité et l'utilisation des langues et elle est souvent présente plus ou moins informellement et à des degrés divers. Nous ne citons ici que les écoles s'identifiant comme 'immersives'.

20. Par conséquent, en prévoyant que l'enseignement d'une langue régionale peut prendre la forme d'un enseignement immersif, l'article 4 de la loi déférée méconnaît l'article 2 de la Constitution. Il est donc contraire à la Constitution.

22. En prévoyant que des mentions des actes de l'état civil peuvent être rédigées avec des signes diacritiques autres que ceux employés pour l'écriture de la langue française, ces dispositions reconnaissent aux particuliers un droit à l'usage d'une langue autre que le français dans leurs relations avec les administrations et les services publics. Dès lors, elles méconnaissent les exigences précitées de l'article 2 de la Constitution.

23. Par conséquent, l'article 9 de la loi déférée est contraire à la Constitution.

Les décisions du Conseil Constitutionnel n'ont pas porté sur l'objet du recours mais ont posé comme non constitutionnels l'enseignement immersif et les signes diacritiques des langues régionales dans les Actes d'État Civil. La loi relative à 'la protection patrimoniale des langues régionales et à leur promotion' parue le 21 mai 2021 au Journal Officiel a été amputée des 'dispositions déclarées non conformes à la constitution '.

Ce recours et ces décisions du Conseil Constitutionnel ont donné lieu à de fortes réactions. Elles sont venues de diverses directions. Des spécialistes de droit constitutionnel, comme Wanda Mastor, Professeur de Droit Public à Toulouse, ont vivement protesté sur les méthodes de la prise de décision du Conseil Constitutionnel, sans aucun débat et sans tenir compte d'autres possibilités d'interprétation des textes de loi et de la constitution.

De nombreuses manifestations de protestation ont été organisées. Le 20 mai 2021, elles ont rassemblé plusieurs milliers de personnes à travers toute la France et l'évènement a reçu une large couverture médiatique.

Face à ces contestations, le 26 mai 2021, le président Macron a posté un message sur Facebook, dont voici des extraits : 'Les langues de France sont un trésor national. [...] Depuis des décennies, un mouvement majeur de transmission par l'école immersive, au travers d'associations come Diwan, Seaska, les Calendretas, Bressola, AMBC et autres, a fait vivre ces langues et a garanti leur avenir. Rien ne saurait entraver cette action décisive portée par nombre

d'engagés, qui ont tout à la fois, l'amour de leur région, la passion de la France et le goût de l'universel [...]. Le droit doit libérer, jamais étouffer. [...]. Voilà pourquoi j'ai demandé au gouvernement et au Parlement de trouver les moyens de garantir la transmission de cette diversité linguistique dans le respect des cadres pédagogiques largement reconnus depuis un demi-siècle '

Le Premier Ministre, Jean Castex, lui-même locuteur de catalan, s'est ensuite investi dans la défense des langues régionales et a très rapidement promis une circulaire permettant de corriger des effets de la loi du 21 mai 2021. Effectivement, une circulaire intitulée ' Langues et cultures régionales' a été publiée le 16 décembre 2021 au Bulletin Officiel de l'Éducation Nationale, de la Jeunesse et des Sports, permettant dans la pratique 'l'enseignement bilingue par la méthode dite immersive'. Nous reproduisons ci-dessous la partie concernant l'enseignement immersif.

'L'objectif des classes bilingues et des sections bilingues, de la maternelle au lycée, est d'assurer une maîtrise équivalente du français et de la langue régionale, que ce soit par la parité horaire hebdomadaire dans l'usage des deux langues ou par l'enseignement bilingue par la méthode dite immersive. Cet enseignement par immersion est une stratégie possible d'apprentissage de l'enseignement bilingue. S'agissant en particulier des trois cycles d'enseignement primaire considérés dans leur globalité, cet enseignement associe l'utilisation de la langue régionale et celle de la langue française pour parvenir rapidement à une certaine aisance linguistique des élèves dans les deux langues. Le temps de pratique de chacune des deux langues peut varier dans la semaine, l'année scolaire ou encore à l'échelle des cycles, en fonction des besoins effectivement constatés.'

L'enseignement immersif est donc accepté en pratique mais reste dans la précarité juridique, une circulaire pouvant être aisément abrogée. Cette circulaire représente cependant un grand pas en avant, autorisant officiellement l'enseignement immersif dans tous les établissements. Elle ouvre ainsi la porte à sa diffusion là où il n'était pas ou très peu pratiqué, en particulier dans les écoles publiques.

Nous nous trouvons actuellement dans une situation qui évolue de façon continue en ce qui concerne les langues régionales non

seulement dans leur statut, leur enseignement mais aussi dans leur approche scientifique avec l'apport de nouvelles technologies et théories.

1.3 L'étude des langues régionales et les nouvelles technologies

Les langues régionales font progressivement leur entrée dans l'ère numérique. Leur étude, leur archivage, leur diffusion recourent désormais à des outils numériques. L'existence de ces outils influent sur la vie des langues, le regard que les locuteurs et le public portent sur elles, leur vitalité, leur revitalisation, le cas échéant.

Un inventaire des ressources et des technologies numériques disponibles pour les langues régionales a été dressé en 2014 par l'Agence ELDA (*agence pour l'évaluation et la distribution des ressources linguistiques*) en partenariat avec la DGLFLF (Direction Générale à la Langue Française et aux Langues de France). Il donne une idée de la diversité des ressources et des acteurs en jeu.

Pour les ressources numériques, l'agence a recensé les corpus de texte, de parole, les corpus parallèles, les corpus multimédias, les dictionnaires, les grammaires. Pour les technologies numériques, elle a considéré la traduction automatique, la synthèse et reconnaissance vocale, les correcteurs orthographiques, les analyseurs grammaticaux et sémantiques. Ces ressources proviennent de différents acteurs : médias, réseaux associatifs – très actifs dans la transmission du savoir des langues régionales – universités ou centres de recherche. Il ressort de cet inventaire que les langues régionales, quoique toutes fortement sous-dotées comparé à des langues comme le français ou l'anglais, le sont cependant à des degrés divers. Celles qui ont le plus de ressources sont l'occitan, le basque et le breton. Cet inventaire est consultable en ligne[17].

On peut aussi consulter les vidéos des présentations ainsi que les Actes en ligne du colloque ' Les technologies pour les langues de France' (TLRF[18]). Celui-ci a été organisé en 2015 par l'IMMI-CNRS,

[17] http://www.elra.info/media/filer_public/2014/12/17/rapport_dglflf_05112014-1.pdf.

[18] Les conférences en captation vidéo sont disponibles en ligne : https://webcast.in2p3.fr/container/tlrf . Les actes sont parus en 2016, également en accès ouvert à partir de cette adresse: https://www.culture.gouv.fr/Thematiques/Langue-francaise-et-langues-de-France/Agir-pour-les-langues/Innover-dans-le-domaine-des-langues-et-du-numerique/Accompagner-les-technologies-de-la-langue-et-la-normalisation/Actes-du-colloque-sur-les-technologies-pour-les-langues-regionales-de-France-TLRF-2015

la DGLFLF, le LIMSI-CNRS, ORTOLANG et ELDA, avec des représentants des collectivités locales et des offices des langues régionales. Il a, entre autres, permis de formuler des propositions pour que les avancées en Traitement Automatique des Langues puissent être mises au service des régions et collectivités locales.
D'autres sources ne concernant pas spécialement les langues régionales sont aussi disponibles, comme les sites d'archivage de ressources et outils numériques que nous allons présenter plus loin (CoCoON, Ortolang, Nakala, *European language grid*, CLARIN en particulier).

L'archivage de données de parole se développe et se structure institutionnellement depuis le début des années 2000. Il s'est inscrit dans le sillage des actions initiées par le CNRS en 2008 avec un très grand équipement pour l'archivage scientifique, qui est devenu la TGIR (la *Très Grande Infrastructure de Recherche*) des humanités numériques (Huma-Num) en 2012 (voir Baude 2016, TLRF en ligne, note 18).

Des outils ont été déployés pour créer des dépôts et un archivage des données orales. La responsabilité de leur mise en place a été donnée au CRDO (*Centre de Ressources et description de l'ora*l) qui, en 2012, est devenu CoCoON (*Collection de Corpus Oraux Numériques).* La mission qui lui a été confiée était de constituer une plateforme apte à la gestion pérenne des opérations d'archivage et de permettre l'accès libre aux données. Ce travail a aussi impliqué le service des archives sonores de la BNF *(Bibliothèque Nationale de France).* Des corpus de parole sont désormais disponibles sur le site CoCoON[19] et sur le site de la BNF. L'équipex ORTOLANG[20], un service complémentaire d'Huma-Num développé depuis 2014, est un ' réservoir de données et d'outils ' pour la langue française ainsi que pour les langues de France. Cette dernière plateforme a ainsi rendu possible l'accès aux atlas linguistiques des langues de France à travers des cartes numérisées (ORTOLANG héberge par exemple l'ALLy = *L'Atlas linguistique et ethnographique du Lyonnais*).

Citons aussi Nakala[21] qui est un ' entrepôt de données de

[19] https://cocoon.huma-num.fr/exist/crdo/
[20] Il s'agit d'un équipement d'excellence, labellisé dans le cadre du PIA, *Programme d'Investissement d'Avenir*, géré par l'État.
[21] Mot *swahili* qui veut dire 'copie', 'exemplaire'.

recherche pour les Sciences Humaines et Sociales ', créé au sein d'Huma-Num en 2015 pour tout type de document[22].

Ces bases de données recueillent des documents de diverses natures et sont utilisées largement pour l'étude des langues régionales que l'on s'intéresse aux textes ou aux sons.

Dans ce nouveau paysage, en termes d'archivage des données, il existe désormais aussi des très grandes infrastructures de recherche européennes, telles que CLARIN : *Common Language Resources and Technology Infrastructure*, fondé en 2012 ou l'*European language grid*, disponible depuis 2021, qui est une plateforme centralisatrice consacrée aux ressources numériques pour les langues, bénéficiant de la participation de 1774 partenaires publics ou privés.

Les ressources numériques et les politiques linguistiques sont également au cœur de l'observatoire des pratiques linguistiques de la DGLFLF, qui soutient les chercheurs avec des appels à projets nationaux lancés à travers le Ministère de la Culture appelés 'Langues et Numérique', un budget dédié aux projets technologiques ou de recherche sur les langues de France. La DGLFLF a mis en place une plateforme de présentation des outils numériques pour les langues (DémoTal : https://www.demotal.fr/), des projets collaboratifs, tels que Wikimédia France qui inclut une médiathèque linguistique *Lingua Libre* avec de nombreux enregistrements de langues régionales. La DGLFLF a aussi mené une enquête sur les pratiques linguistiques et numériques des langues régionales. Le récent ' Rapport au Parlement sur la langue française ' publié en mars 2023[23], et rédigé par la DGLFLF, fournit des analyses, statistiques, chiffres clés des opérations menées.

Ce rapport indique que le rôle de la DGLFLF est dans une phase évolutive dans le cadre du projet de la Cité internationale de la langue française dans le château restauré de Villers-Cotterêts, symbole d'une ouverture aux pratiques du français 'langue monde' et du plurilinguisme sur tous les continents[24]. Le projet de la Cité

[22] Nakala organise les données en se basant sur une méthodologie documentaire et des modèles accessibles, on peut ajouter des métadonnées normalisées à l'aide d'identifiants pérennes.

[23] Auteur : la DGLFLF. Il est téléchargeable à ce lien https://www.culture.gouv.fr/Presse/Communiques-de-presse/Publication-du-Rapport-au-Parlement-sur-la-langue-francaise-2023

[24] Rappelons qu'en 1539, par l'ordonnance, signée par François Ier au château de Villers-Cotterêts, l'usage du français a été imposé pour les actes de justice et d'état civil, à la place du latin

internationale de la langue française à Villers-Cotterêts inclut la création d'un centre de référence pour les technologies de la langue : LANG :IA (= LANG pour langues et IA pour Intelligence Artificielle) et la constitution d'un Consortium pour une infrastructure numérique européenne (*European Digital Infrastructure Consortium*, EDIC) pour les langues. 'Il devrait permettre aux acteurs français de valoriser leur savoir-faire dans le domaine des technologies du langage, en France et en Europe, de réunir toutes les forces dans un même écosystème (recherche, industrie, tissu associatif) pour mutualiser ressources, données et services en une même structure française et européenne' (voir rapport DGLFLF 2023, p. 76). À présent il n'est pas possible de mesurer l'espace qui sera dédié aux langues régionales et à la diversité linguistique dans ce projet ambitieux. Ce projet participe de la volonté de l'État de renforcer les technologies vocales afin de rendre la France et l'Europe plus compétitives au niveau mondial et de répondre aux besoins numériques et stratégiques de leurs acteurs économiques.

Le ministère de la Culture opère aussi en faveur des langues régionales à travers le Centre international de recherche et de documentation occitanes (CIRDOC) et les quatre offices publics chargés de promouvoir les langues régionales dans l'espace public (en collaboration avec les collectivités territoriales, les associations et les établissements scolaires) : l'Office public de la langue basque, l'Office public de la langue bretonne, l'Office public de la langue catalane, l'Office public de la langue occitane.

On sait aussi que l'UNESCO suit la mise en pratique de la France des recommandations relatives à la promotion des langues minoritaires et du plurilinguisme dans le 'cyberspace'.

Le développement de ressources et outils numériques pour les langues régionales a, ces dernières années, bénéficié du financement de projets par l'ANR dont plusieurs sont présentés dans les chapitres de ce livre. Nous notons, en particulier les projets suivants depuis

(mais implicitement aussi à la place de l'occitan). Par la suite, l'unification du français à partir du 16e siècle, s'accélère avec la Révolution française. On peut donc être divisé par le choix de l'État du château de Villers-Cotterêts, qui en soi semble le symbole du français 'gage' de l'unité nationale et non le symbole d'un espace plurilingue où la langue française cohabite avec les langues pratiquées sur son territoire.

2012 : Thalamus, (*Édition critique du manuscrit AA9 des Archives municipales de Montpellier dit le petit Thalamus*), SYMILA (*Microvariation syntaxique dans les langues romanes de France*), BIM (*La langue basque en devenir : un regard historique à un isolat linguistique*), APPI (*l'Atlas pan picard informatisé*), DADDIPRO (*Données et analyses dialectologiques, acquisitionnelles et diachroniques des pronoms sujets en Gallo Roman*), DIVITAL (*Accroître la vitalité et la visibilité numériques des langues de France : descriptions linguistiques et corpus annotés)* en cours et RESTAURE (*Ressources informatisées et traitement automatique pour les langues régionales*) que nous allons considérer plus en détail.

Le projet de recherche RESTAURE, financé par l'ANR en 2015 pour une durée de 42 mois a concerné trois langues de France, le picard, l'alsacien, l'occitan, représentées via des laboratoires partenaires auxquels s'est adjoint le LIMSI-CNRS pour ce qui est du traitement automatique des langues.

Le projet RESTAURE a permis de mutualiser des compétences et des ressources et de développer des procédures plus spécifiques pour ces langues régionales qui présentent en commun des difficultés liées à leur variabilité tant à l'oral qu'à l'écrit. Le développement de ressources doit, en effet, composer avec de nombreuses variantes dialectales et des graphies diverses non standardisées.

Des outils de reconnaissance de caractères (OCR) ont été adaptés pour chaque langue de façon à prendre en compte les diverses graphies et les transcriptions phonétiques, notamment dans la numérisation des atlas. De plus, des annotations ont pu être enrichies de façon géo-référencée, de façon à avoir la possibilité de générer des cartes qui permettent de visualiser les données ou les catégories grammaticales associées à un lemme dans un dialecte mais aussi d'étudier les zones de transition entre dialectes ou langues de France différentes. Ces traitements linguistiques permettent de gérer la variation dialectale, par exemple entre les dialectes de l'occitan où il existe deux systèmes graphiques différents : la graphie classique (ou albertine) et la graphie mistralienne (du Félibrige) ; certains projets ont comme point de départ les textes écrits, c'est le cas de BaTelÒc, d'autres des données orales.

Par son adaptation à leur spécificité, le projet RESTAURE a permis des progrès importants pour ces langues en termes

d'amélioration des bases de données et de leur moteur de recherche, de tokenisation, de lemmatisation, de constitutions de lexique, d'étiquetage morphosyntaxique, entre autres. La collecte de données, les méthodes d'annotation ainsi que de diffusion des corpus ont pu être aussi améliorées.

Deux chapitres du présent ouvrage s'inscrivent dans le prolongement de ce projet : le chapitre de Christophe Rey sur le picard et celui de Myriam Bras sur la base BaTelÒc en occitan. Chacun de ses chapitres précise les contributions de RESTAURE dans le cadre des langues considérées et les situent par rapport aux recherches qui les concernent. En picard, le projet RESTAURE a permis des avancées en matière lexicale (tokenisation, lemmatisation, d'étiquetage morpho-syntaxique). En occitan, il a permis de constituer le premier corpus annoté en parties du discours, grâce à un module d'analyse morpho-syntaxique.

Tous les chapitres du présent ouvrage se réfèrent à des données numérisées accessibles en ligne, que ce soit des données textuelles (Lieutard et Bras pour l'occitan, Jouitteau sur le breton, Kasstan et Russo pour le nord-occitan et les zones transitionnelles entre occitan et francoprovençal, Rey pour le picard …), des données sonores (Nicolas Quint pour les langues du Croissant, Michela Russo et Jonathan Kasstan pour les zones transitionnelles entre dialectes nord-occitans), des atlas (Michela Russo et Jonathan Kasstan pour les zones transitionnelles entre occitan et franco-provençal, Philippe Boula de Mareüil & al. pour l'ensemble des langues régionales….).

Si tous les chapitres utilisent des données numériques, certains décrivent des ressources ou des outils qui ont été mis en ligne et traitent de leur élaboration : c'est ainsi le cas du Petit Thalamus (Hervé Lieutard), de la base de données BaTelÒc (Myriam Bras), de l'Atlas sonore de Philippe Boula de Mareüil, de la Wikigrammaire ARBRES (Mélanie Jouitteau).

L'accès à des bases de données et le recours à ces nouvelles technologies renouvellent l'étude des langues régionales que ce soit du point de vue synchronique ou diachronique. La possibilité de faire des recherches dans des corpus de parole ou des textes avec des annotations de divers niveaux mais aussi dans des dictionnaires, des atlas permettent d'élargir les données pour les analyses phonétiques, morphologiques, lexicales syntaxiques, sémantiques, modifiant

l'approche des différentes composantes de la grammaire. Cet élargissement considérable des données et les possibilités élargies de traitement ont aussi des répercussions sur la réflexion théorique comme l'illustrent les contributions de Christophe Rey (Rey 2021 ; Rey, ce volume) à la réflexion sur l'organisation des lexiques ou de Mélanie Jouitteau à la morpho-syntaxe et syntaxe du breton (Jouitteau 2020 ; Jouitteau, ce volume).

De plus, l'accès au numérique ne concerne pas seulement les chercheurs mais affecte aussi la vie des langues. On sait qu'il est essentiel pour la survie des langues non majoritaires. Il favorise l'utilisation et la revitalisation des langues, en même temps qu'il modifie le regard porté sur elles. La présence d'une langue sur le web ainsi que de données accessibles sur internet entraînent une implication plus forte des communautés et une augmentation des échanges entre public et chercheurs. Dans ce volume, nous publions une analyse des courriers électroniques reçus en réaction à l'Atlas sonore du LIMSI (voir Philippe Boula de Mareüil & al.). L'accès au numérique a participé aussi au changement de regard porté sur leurs parlers par les locuteurs du Croissant (voir Nicolas Quint, ce volume).

Les études sur les langues régionales progressent donc actuellement dans et du fait de cet environnement numérique. Tous les travaux présentés dans cet ouvrage ont eu recours à des moyens numériques, au minimum à des bases de données numériques. Au centre du questionnement de notre Journée scientifique sur les nouvelles approches pour les langues régionales, se trouvait donc la question de l'apport des outils numériques et des nouvelles possibilités qu'ils offrent pour l'étude des langues régionales, leur documentation, leur archivage, leur analyse linguistique ainsi que pour leur vitalité ou revitalisation.

2. Présentation des chapitres

Ce livre rassemble 7 chapitres répartis en 4 rubriques, en fonction des outils utilisés et des effets produits ou recherchés sur la vie des langues.

Les langues concernées ou principalement concernées sont indiquées entre parenthèses :

2.1. Outils numériques, linguistique et revitalisation des langues (breton, picard)
2.2. Bases de données textuelles, lexique et syntaxe (occitan)
2.3. Bases de données orales, phonétique et diachronie (aires transitionnelles entre occitan et franco-provençal)
2.4. Outils numériques, bases de données orales et implication des locuteurs ou du public (langues du Croissant, occitan, rromani)

2.1 Outils numériques, linguistique et revitalisation des langues (breton et picard)

Le chapitre de Mélanie Jouitteau et Reun Bideault, intitulé ‘ Outils numériques et traitement automatique du breton ’, présente un état de l’art sur l’environnement numérique du breton. Cette langue, dont le nombre de locuteurs est estimé à plus de 150000, est encore faiblement dotée, même s’il existe des outils précieux. Le breton dispose d’un traducteur breton-français Apertium, disponible en ligne, construit à partir de l’analyse morphologique de Tyers (publiée entre 2007-2009) et d’un dictionnaire bilingue. Un conjugateur de verbes a récemment été développé (Morvan, 2021). Plusieurs correcteurs orthographiques sont disponibles. Il existe également plusieurs dictionnaires en ligne de breton contemporain ainsi qu’un agrégateur de dictionnaires Geriafurch développé par Anthony Lannuzel. Le dictionnaire breton-néerlandais de Jan Deloof (2008-2010), ainsi que le dictionnaire sous format pdf de Favereau (1993), atteignent l’un et l’autre 40000 entrées. On compte aussi des traductions de logiciels, une utilisation sur les réseaux sociaux (Facebook, entre autres). Un système de synthèse de la parole a été développé sous l’égide de l’Office Public de la Langue Bretonne. Le breton n’est pas sous-doté en termes de corpus mais ceux-ci sont très divers et souvent peu utilisables pour le TAL.

Les grammaires en ligne – qui nous intéressent particulièrement en tant que linguistes – font l’objet d’un intérêt particulier. Le site ARBRES développé et maintenu par Mélanie Jouitteau offre une Wikigrammaire du breton accessible à tous. Ses buts sont de ‘fournir

une description fine et théoriquement informée de la variation syntaxique en breton et un état des lieux permanent et à jour des différentes recherches en syntaxe formelle.' Il est tout aussi bien consulté par bretonnants que par des étudiants en linguistique et il a aussi une audience internationale, comme l'indique l'analyse des connections. La banque de données d'ARBRES contient 25000 exemples bretons tous étiquetés morphologiquement et syntaxiquement. Les auteurs détaillent le mode d'étiquetage, qui reflète l'analyse grammaticale sous-jacente. Par ailleurs, une banque d'arbres en termes de structures de dépendance a été initiée par Tyers et Ravishankar (2018) sur la base de 888 exemples. Celle-ci utilise les critères et étiquetages de l'UD (*Universal Dependencies*) qui tend à devenir un standard pour les banques d'arbres. Cette base d'arbres pourra être enrichie des 25000 exemples déjà glosés de la Wikigrammaire, des correspondances pouvant être établies entre les catégories des gloses de la Wikigrammaire et celle de la banque en termes de UD. Elle gagnera aussi en adéquation par rapport aux spécificités du breton.

Enfin, le chapitre se tourne vers l'avenir et propose de nombreuses pistes d'améliorations des outils qui prennent en compte les besoins exprimés par les bretonnants. Au premier rang, figure le besoin de ressources sonores pour l'apprentissage de l'accentuation et d'une 'souplesse cross-dialectale'. L'adaptation 'responsive' pour les téléphones portables et aux nouvelles normes apparait aussi comme une nécessité. Enfin, la recherche universitaire et les procédures d'évaluation devraient accompagner ces développements et valoriser le travail effectué sur les bases de données, les sites web, les banques d'arbres…

Le chapitre de Christophe Rey se pose la question : 'Peut-on revitaliser la langue picarde grâce aux nouvelles technologies ? '.

La langue picarde se parle dans une aire qui couvre en France la région des Hauts de France et en Belgique la province du Hainaut. Elle présente de nombreuses variétés et n'a pas actuellement de forme standardisée. C'est aussi une langue écrite, avec une longue histoire littéraire remontant au Moyen-Âge. Malgré le nombre de locuteurs, sa littérature, il n'y a pas d'enseignement bilingue, a fortiori immersif. Alors même qu'en Belgique, elle a le statut de

langue régionale depuis 1992 et peut à ce titre être enseignée, ce n'est qu'en décembre 2021, qu'elle a été reconnue comme matière d'enseignement dans l'éducation nationale.

La langue picarde bénéficie d'une longue tradition dialectologique, avec, entre autres, la réalisation d'atlas (cf. L'atlas linguistique et ethnographique Picard, 1989, 1998, 2004). Les vingt dernières années ont été marquées par le développement d'outils informatisés. On note en particulier la très importante base de données PICARTEXT constituée sur le modèle de FRANTEXT à l'Université Jules Verne de Picardie avec actuellement 138 textes et 3,5 millions de mots. Par ailleurs, le projet RESTAURE – dont le but était d'étendre et adapter des outils du TAL aux langues régionales – a permis d'améliorer les outils d'étiquetage morpho-syntaxique et lexical et de faire avancer la réflexion sur la constitution des lexiques pour les langues avec de nombreuses variétés comme le picard. Très récemment, un Atlas Picard informatisé a été terminé à l'Université de Lille sous la responsabilité scientifique d'Esther Baiwir. METALPIC est en projet en cours de réalisation permettant de 'dresser une cartographie actualisée et critique des ressources lexicographiques très diverses rédigées depuis des siècles en langue picarde'. Ces travaux ont nourri le *Dictionnaire fondamental français-picard* dirigé par A. Dawson et L. Smirnova, publié avec mise en ligne en 2020. Celui-ci fournit les équivalents picards des 1000 mots les plus fréquents du français littéraire contemporain et s'adresse aux locuteurs ou apprenants du picard.

'L'entrée du picard dans la galaxie des langues investies par le TAL' a eu des retombées concrètes, contribuant à sa revitalisation. Sans doute la plus notable est la création de la *Commission de néologie et terminologie pour la langue picarde* qui ' dans la mesure où elle permet d'arrimer la langue dans une modernité lexicale investie par les acteurs institutionnels fait sortir le picard de l'ornière 'patrimonialisante' dans laquelle beaucoup d'observateurs la cantonnent'. Le picard, au même titre que la plupart des langues régionales de France, n'est pas qu'un patrimoine, il est aussi une langue bien vivante, dont il reste encore des locuteurs et des scripteurs.'

2.2 Bases de données textuelles, lexique et syntaxe (occitan)

L'article d'Hervé Lieutard intitulé 'Nouvelles approches linguistiques et lexicographiques de l'occitan médiéval' porte sur l'édition critique numérique du *Petit Thalamus* (http://thalamus.huma-num.fr), une base de données qui héberge de nouvelles éditions de textes occitans médiévaux encodés dans le langage XML/TEI. Cette base a été créée au départ grâce au financement d'une ANR (https://anr.fr/Projet-ANR-10-JCJC-2003) et à l'apport de l'*Association internationale d'études occitanes* (AIEO). Il s'agit d'une version numérisée du manuscrit AA9 des Archives Municipales de Montpellier, le manuscrit du gouvernement du Consulat de la ville médiévale de Montpellier. Ce manuscrit a été écrit en occitan à partir du XIII^{e} siècle, puis en français à partir du XVI^{e}. Il est également corrélé à un ensemble de documents appelés *Thalami*, qui émanaient du Consulat montpelliérain et concernaient la gestion de la ville. L'édition du manuscrit principal est dotée de deux index électroniques d'anthroponymes et de toponymes.

Hervé Lieutard montre les avantages de cette base et présente le questionnement crucial qui a porté à la lemmatisation des formes occitanes en diachronie. Dans le *Petit Thalamus* la lemmatisation est faite à travers la graphie classique, dans laquelle convergent les différentes variantes graphiques. Il est ainsi possible de façon synoptique de visualiser et comparer divers manuscrits, représentés par les lettres A B C D E F G H, ainsi que les choix de récriture.

Cette base permet des études linguistiques sur une diachronie longue, d'observer l'évolution des graphies (p. ex. l'émergence de la nasale ou de la latérale palatale <nh> et <lh>, la diphtongaison conditionnée), de la morphosyntaxe, notamment la disparition progressive de la déclinaison bicasuelle, la naissance d'une variété écrite suprarégionale à la fin du XV^{e} s. Hervé Lieutard montre comment des solutions graphiques se sont diffusées dans l'espace occitan à partir du centre déjà vers le XIVe s. La base électronique du *Petit Thalamus* représente sans nul doute un apport majeur pour les études sur l'occitan, elle permet d'analyser et interpréter les usages graphiques et d'évaluer l'évolution diatopique des usages de la langue. Il est possible d'extraire les données via des requêtes X-Path.

Ce projet est relié à un réseau national et international d'autres projets d'édition numérique en cours sur l'occitan ancien, notamment AcTo (*Aculhir e Tornar, Ressorsas numericas per l'occitan medieval*, https://acto.hypotheses.org), qui partage un forum de discussion sur les méthodes d'encodage et lemmatisation sous le patronage de l'AIEO. Ces derniers ont donné vie à un 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*, développé à Munich.

L'article passe en revue également les autres bases numériques qui hébergent l'édition électronique de textes médiévaux occitans, telles que le *Corpus linguistique du gascon ancien* de Thomas Field (sur le site : http://mllidev.umbc.edu/gascon/French/index.html), le *Trésor manuscrit de l'ancien occitan* (TMAO, accessible via l'AIEO http://tmao.aieo.org).

L'article de Myriam Bras, intitulé 'Nouvelles perspectives pour la linguistique occitane à partir de la base textuelle BaTelÒc' porte sur le TAL multilingue appliqué aux corpus et ressources pour la linguistique occitane. Myriam Bras évalue les progrès réalisés grâce à la base de données textuelles pour l'occitan, BaTelÒc, tout en reconnaissant l'apport des autres sources numériques pour la langue occitane disponibles en France.

Elle dresse un bilan de ces ressources au service de la linguistique outillée et du traitement automatique des langues, appliqués à la langue occitane, à partir de son expérience de sémanticienne et du constat qu'autour des années 2000 nous étions face à une absence de données textuelles numérisées pour l'occitan. Cette situation a poussé à la création d'une base textuelle pour l'occitan, sur le modèle de *Frantext*, un projet qui a débuté en 2006 au sein du laboratoire ERSS (UMR 5610 de l'Université Toulouse Le Mirail et du CNRS), dont elle fait partie.

Le but de cette base est de réunir des textes de genre différents, rédigés en occitan entre le XIX^{e} et le XXI^{e} siècles. Une version expérimentale a été mise en ligne en 2008 avec un accès réservé. Elle a été ensuite enrichie, notamment dans le cadre du projet RESTAURE (*RESsources informatisées et Traitement AUtomatique pour les langues REgionales*), https://restaure.unistra.fr). ANR-14-

CE24-0003, 2015-2018, devenant alors exploitable à travers les concordances, l'extraction de formes, etc.

La base BaTelÒc mise en ligne en 2016 (http://redac.univ-tlse2.fr/bateloc/) contient une centaine de textes d'auteurs différents, elle couvre plusieurs dialectes de l'occitan. Elle contient des éditions des textes qui vont de 1836 à 2014 (comptes littéraires, poésies, chroniques…). Ils sont rédigés en graphies différentes, alibertine, mistralienne, etc.

L'objectif de cette base est de réunir les textes plus représentatifs pour chaque dialecte de l'occitan, avec des genres textuels variés et une dimension diachronique. Une de ses originalités est la prise en compte de la dimension diamésique à travers des textes de théâtre en transcription orale, ce qui permet aux chercheurs de pouvoir effectuer une comparaison entre l'écrit et l'oral. La base a évolué entre 2016-2023, elle permet désormais de mener des analyses morpho-syntaxiques, de créer un lexique des formes fléchies, des sous-corpus, d'effectuer une analyse syntaxique. Myriam Bras nous précise que cette dernière étape a pu être atteinte grâce à un projet européen LINGUATEC (2018-2021), qui a impliqué également des équipes travaillant sur d'autres langues, comme le basque. Le résultat de cette collaboration a été un corpus d'occitan annoté en dépendances syntaxiques.

LINGUATEC est un projet qui a réuni des chercheurs travaillant autour des corpus annotés POS et DEP et qui a fédéré un nombre considérable de linguistes, informaticiens et dialectologues travaillant sur d'autres langues de France et qui, en lien avec le *Congrès Permanent de la Lenga Occitana*, a établi une feuille de route pour le développement numérique de l'occitan.

2.3 Bases de données orales, phonétique et diachronie (aires transitionnelles entre occitan et francoprovençal)

L'article de Michela Russo et Jonathan Kasstan, intitulé ' On vowel nasalisation in transitional Francoprovençal and Occitan areas ' propose une explication diachronique et synchronique de la nasalisation des voyelles dans les zones nord-occitanes dites 'transitionnelles' entre le francoprovençal et l'occitan. Bien que les voyelles nasales soient souvent comptées parmi les traits opposant les langues d'oïl et le francoprovençal à l'occitan central, la nasalité

est bien attestée dans le nord-occitan de l'Auvergne et dans ses sous-dialectes, en Ardèche, puis dans l'occitan de la Drôme et de l'Isère. Dans cet article, les auteurs soutiennent que la nasalité du nord-occitan est un processus autochtone qui n'a été emprunté ni au français ni au francoprovençal. De nombreux indices acoustiques et phonologiques, expliqués en détail, tels que la vélarisation et l'arrondissement notamment dans les séquences /a +N/ sont pour les auteurs des traces de nasalité dans les variétés de l'occitan septentrional. Pour étayer cet argument, ils ont rassemblé un ensemble de données provenant d'enregistrements radiophoniques et d'interviews récents (avec locuteurs de l'occitan septentrional) venant des *Archives départementales de la Haute-Loire*, puis de la région Ardèche. Ces matériaux comprennent des enregistrements recueillis dans la zone protestante du Chambon-sur-Lignon (commune de Tence) par le dialectologue Théodore De Félice, pour une radio locale de la même région (Chambon-sur-Lignon), appelée Cimes du Lizieux (1984-1990) et de l'émission 'Le patois vous parle'. Leur analyse est également corroborée par les enquêtes orales en occitan-vellave déposés sur la plateforme *CoCoON COllections de Corpus Oraux Numériques*, intégrée aux archives de la TGIR Huma-Num : https://cocoon.huma-num.fr/ et par une analyse de textes vellaves du 17e et 18e siècle, récemment édités et mis à disposition de l'audience scientifique.

2.4 Outils numériques, bases de données orales et implication des locuteurs ou du public (langues du Croissant, occitan, rromani)

Le chapitre suivant par Nicolas Quint s'intitule 'Les parlers du Croissant : un aperçu des actions actuelles de documentation et de promotion d'un patrimoine linguistique menacé'. L'aire des parlers du Croissant a la forme d'un croissant aux limites Nord du Massif Central. C'est une zone tampon entre les parlers d'oïl et les parlers d'oc, avec un foisonnement de variétés intermédiaires, souvent quasi-inconnues de la communauté scientifique.

Le chapitre de Nicolas Quint présente d'abord ces parlers fort mal connus, en insistant sur leurs caractéristiques linguistiques. L'histoire des recherches sur les parlers du Croissant jusqu'en 2013, année pivot, est ensuite présentée selon deux phases : une phase

allant des origines à 1913 et une seconde phase de 1913 à 2013. La première phase correspond à la genèse conceptuelle du Croissant à partir du 19ème siècle avec la reconnaissance de la zone du Croissant et quelques travaux pionniers, dont le plus important est la thèse de l'Abbé Rousselot soutenue en 1891 sur le parler de Cellefrouin (Charente) d'où était originaire sa famille. De 1913 à nos jours, les connaissances se sont progressivement accumulées, avec des thèses, des ouvrages, des atlas linguistiques, des monographies produites par des linguistes ou des érudits locaux. À partir de 2013, une nouvelle dynamique est insufflée marquée par un développement des interactions entre chercheurs et locuteurs. Des projets collaboratifs ont vu le jour et ont pu trouver des financements. Des colloques associant les locuteurs sont organisés. Les recherches deviennent aussi pluridisciplinaires associant linguistique, TAL, sociologie.

On note de nombreuses réalisations récentes : des publications collectives, des monographies, des publications de valorisation (en particulier la traduction du Petit Prince dans un grand nombre de variétés croissantines). Il y a eu aussi une importante activité de collecte de matériaux linguistiques qui pour la plupart sont sous forme numérique, avec des possibilités d'accès en ligne. Un documentaire de 40mn a également été tourné (Franck Guillemain, 2020). Il a contribué à faire connaître les travaux des chercheurs, à développer l'intérêt pour ces parlers et à changer leur image dans la région.

L'avenir de ces parlers est plus qu'incertain, la quasi-totalité des locuteurs natifs ayant plus de 75 ans et la transmission familiale s'étant arrêtée il y a 60 à 70 ans. Il est d'autant plus urgent de les étudier et les archiver.

Le chapitre de Nicolas Quint illustre une des évolutions actuelles de la recherche sur les langues régionales qui consiste en l'établissement de collaborations entre chercheurs et locuteurs ainsi qu'une implication croissante des locuteurs et acteurs régionaux. Ces collaborations ne bénéficient pas seulement à la recherche mais elles modifient l'attitude des locuteurs vis-à-vis de leur langue et leur intérêt pour elle. Les locuteurs deviennent des membres actifs de la recherche et du devenir de leur parler.

L'article de Philippe Boula de Mareüil, Marcel Courthiade†, Frédéric Vernier, intitulé ' De la Provence aux Balkans : discours épilinguistiques autour d'un atlas sonore des langues régionales ou minoritaires d'Europe', est un hommage à Marcel Courthiade, spécialiste de la langue romani, décédé le 4 mars 2021, quelques mois avant la journée de la *Société linguistique de Paris* sur les langues régionales.

Marcel Courthiade était un spécialiste du romani et il était titulaire depuis 1997 de la chaire de langue et civilisation romani à l'Inalco. Il avait été élève de Georges Calvet, et avait soutenu en 1995, un doctorat sur la ' Phonologie des parlers rroms et diasystème graphique de la langue rromani ', sous la direction de Claude Hagège.

Marcel Courthiade était commissaire aux droits linguistiques de l'Union Rromani Internationale, consultant auprès de gouvernements pour l'éducation des Rroms (Hongrie, Serbie, Kosovo, Albanie), représentant de l'Union Rromani Internationale près l'Unesco depuis 1990. Il avait aussi co-organisé le colloque international ' Millénaire Rrom : Bilan et perspectives ' en 2020.
Marcel Courthiade était un éminent linguiste, qui laisse un riche héritage en termes de publications et documentation sur le romani, mais il était aussi reconnu pour son humanisme.

Dans ce chapitre, terminé après la disparition de Marcel Courthiade, les auteurs présentent l'Atlas sonore des langues régionales mais aussi les discours épilinguistiques auxquels il a donné lieu tels qu'ils se manifestent dans les centaines de courriers électroniques adressés aux auteurs.

L'Atlas sonore des langues régionales (accessible en ligne à l'adresse https://atlas.limsi.fr), permet d'écouter et de lire une même fable d'Esope dans 800 versions. Partant de la France hexagonale, il a été étendu à l'Outre-Mer, aux langues non territoriales de France ainsi qu'à des langues européennes. Toutes ces langues sont, à divers degrés, en butte à la dévalorisation sociale, à l'invisibilisation et aussi, à la 'patronalisation' qui les muséifie.

Les auteurs se concentrent d'abord sur l'occitan. La constitution même de l'atlas pour cette langue très dialectalisée est d'abord abordée, avec ses points d'enquête, sa méthodologie, le choix des graphies, les difficultés rencontrées dans la cartographie des aires

dialectales. Ensuite, l'analyse des courriers électroniques reçus révèlent des points sensibles, des fractures, des débats dans la communauté, tant sur le statut des langues, leurs graphies que les aires dialectales. Un zoom plus fin est ensuite fait sur quelques points d'enquête situés dans le domaine provençal et le languedocien oriental et sur certaines de leurs propriétés linguistiques (phonologiques, morphologiques ou syntaxiques).

L'Atlas sonore a été récemment étendu à des langues minoritaires d'Europe (5 langues celtiques, 5 langues finno-ougriennes, 20 langues ou dialectes du domaine germaniques, 5 variétés balkano-romanes et 6 langues sans territoire compact, dont le romani). Un zoom est fait sur plusieurs points d'enquête en macédonien de Golo Brdo, en goran et en roumain de Transylvanie.

Si l'objectif premier de 'cet atlas est d'illustrer la diversité linguistique pour la promouvoir à travers une carte parlante et préserver un témoignage sonore de certaines langues menacées d'extinction', il contribue aussi à les rendre plus visibles. Les échanges électroniques indiquent déjà qu'il a su faire parler de lui et par là même des langues régionales.

La journée annuelle de la SLP du 12 juin 2021 avait également accueilli trois autres conférences sur les langues de France, celle de Ricardo Etxepare (CNRS, IKER, UMR 5478, Bayonne) intitulée 'La langue basque : défis sociaux, paris scientifiques'. Cette conférence sur le basque nous a appris que le basque comporte un nombre de variétés locales et régionales peu explorées, ainsi qu'une variété de basque standard (*euskara batua*), utilisée au sein du système scolaire, dans les médias, et dans la production littéraire.

Pendant sa présentation, Ricardo Etxepare a analysé un certain nombre de défis concernant le basque : la codification d'une langue commune, la généralisation du locuteur bilingue L2 à travers le système éducatif, le devenir sociolinguistique de la langue au sein des politiques linguistiques en France, en particulier, avec la décision récente du conseil constitutionnel portant sur l'enseignement immersif en France, le développement des outils de traitement automatique et la création de big data. À titre d'illustration, il présente un outil numérique d'appui aux recherches sur la variation grammaticale, l'atlas syntaxique *Basyque (Basque Basic Syntactic*

Atlas), et une base de données réalisée en collaboration par une équipe de sociolinguistes et une équipe de juristes, appelée *CLME (Catégorisation des Langues Minoritaires en Europe)*, hébergée à la *Maison des Sciences de l'Homme* d'Aquitaine et coordonnée par Antoine Pascaud (U. Bordeaux Montaigne, UMR 5478 Iker) et Alain Viaut (CNRS, UMR 5478 Iker). Il précise en outre que le pays basque espagnol offre un contexte universitaire et de recherche plus riche en termes de ressources, et une politique qui a encouragé le plurilinguisme.

À propos du défi que constitue la codification d'une langue commune et des moyens de mesurer le degré de variation interne de la langue, il mentionne des outils tels que l'approche quantitative basée sur l'*Atlas ethnolinguistique du pays basque* (EAEL), réalisée par la société scientifique Arantzadi entre 1977-1983 avec 80 informateurs distribués sur l'ensemble du Pays basque (selon le découpage dialectal ' Bonaparte '), âgés de 69 en moyenne lors de l'enquête, 570 questions et publiée en deux tomes en 1983 et 1990. Puis Ricardo Etxepare explique les deux plans de normalisation, avant et après la guerre ainsi que la décision de création d'une académie de la langue basque, prise lors du congrès mondial d'Oñate en 1918, en réponse au souhait d'une langue unifiée chez les contemporains. Il nous indique la proposition en 1935 de Resurrección María de Azkue, président de l'académie de se référer à un ' guipuscoan enrichi ' (des apports d'autres dialectes), vu que le guipuscoan est le dialecte qui a à la fois un nombre élevé de locuteurs et qui constitue en pays basque espagnol, le dialecte littéraire de référence. Concernant le processus d'unification du basque dans l'après-guerre (*batua*), Ricardo Etxepare précise qu'il a été tenu compte de deux faits : le poids des variétés dialectales et la tradition littéraire. Il a souligné que le basque n'a pas de capitale linguistique, Bilbao n'étant pas une ville à majorité bascophone depuis la fin du XVII^{e} s. et a ajouté des explications concernant la variation dialectale, une zone dialectale centrale, le guipuscoan et le bas-navarrais. Il y a aussi quatre grands dialectes littéraires depuis le XVIII^{e} siècle (biscayen, guipuscoan, labourdin et souletin). Dans la perspective de l'académie sur l'unification, un axe vertical temporel, associé à la tradition littéraire, est à intégrer.

Delphine Bernhard (Strasbourg Univ., LiLPa, EA 1339) dans sa conférence intitulée 'Traitement automatique des langues régionales de France : retour d'expérience sur les dialectes alsaciens' a pris comme point de départ les recherches en traitement automatique des langues (TAL), au profit des langues très dotées, notamment à partir de l'exemple du portail européen META-SHARE. Puis en s'intéressant au cas des dialectes alsaciens, elle nous a expliqué quels étaient les progrès réalisés dans les dernières années, notamment au travers du projet RESTAURE 2 (*Ressources informatisées et Traitement Automatique pour les langues régionales*), financé par l'ANR (2015-2018 https ://restaure.unistra.fr), consacré à trois langues régionales de France : l'alsacien, l'occitan et le picard.

Elle insiste sur les défis suivants : manque de données numériques, variations dialectales, difficultés à valoriser le travail de collecte et d'annotation des données, l'adoption des principes FAIR 3 dans le cadre de la diffusion des ressources TAL et linguistique outillée.

Patrick Sauzet (Toulouse 2, CLLE, UMR 5263) dans sa communication ' L'occitan pro-drop or not pro-drop : l'éclairage de la base de données en ligne SYMILA ', s'était concentré sur un projet scientifique dont l'objectif est l'observation de la variation syntaxique dans les langues de France, le projet SYMILA (http ://symila.univ-tlse2.fr/), acronyme pour 'Syntactic microvariation in the Romance languages of France'. Ce projet documente la variation géographique du français ou langue d'oïl, de l'occitan, du francoprovençal, et marginalement du catalan. Il utilise les formes codifiées des langues, les notations orthographiques et il permet d'explorer les corrélats grammaticaux des langues de France. La communication porte principalement sur la variation syntaxique de l'occitan. Patrick Sauzet nous explique que l'occitan a émergé clairement au moyen-âge puis au XIX^{e} siècle et de manière plus complexe entre ces deux époques.

Il se concentre sur une propriété prototypique de l'occitan, le caractère pro-drop, un des traits qui caractérisent l'occitan. Ce trait oppose prototypiquement 'canta' (' chanta', 'que canta') au français 'il chante' ('il cante', 'al chante', 'le chante'). Patrick Sauzet précise

que, toutefois, ce trait syntaxique n'est pas présent dans toutes les variétés occitanes, ni au niveau diatopique, ni diachroniquement.

Par exemple, le nord-ouest de l'occitan (haut limousin et marchois) présente régulièrement des pronoms sujets. En outre, l'occitan de la fin du moyen-âge et de l'âge baroque (XVI^{e}-XVII^{e}) emploie aussi massivement des pronoms sujets, ce qui ne peut pas dépendre d'une imitation du français.

La base SYMILA permet également d'utiliser les données de l'ALF (l'*Atlas Linguistique de la France*) qui aident à délimiter le phénomène. Patrick Sauzet nous montre que dans les aires à sujet nul on trouve des pronoms sujets sporadiques, typiquement non clitiques : 'n(os)autres ', alors que des parlers occitans extrêmes, aux confins du 'Croissant', présentent des pronoms sujets redoublant un sujet lexical (ayant une fonction affixale).

Dans des parlers plus au sud, on trouve des formes pronominales faibles, dépourvues de voyelle ou avec schwa. Au contact de la zone géographique, caractérisée régulièrement par le sujet nul, on voit en revanche des formes plus pleines. Selon Patrick Sauzet, cela suggère un processus non déterministe dans lequel la grammaire de l'occitan médiéval à verbe second (V2) favorise l'emploi de pronoms sujets. D'après lui, cet emploi massif se dissocie ensuite de l'ordre V2 et évolue soit en cliticisation de formes pronominales sujets, soit par retour à un sujet nul.

Ces trois auteurs n'ont pas pris part à la constitution du volume, néanmoins leur contribution à la journée annuelle de la Société Linguistique de Paris s'est enrichie d'un débat intense suscité par leurs conférences centrées sur des questions clés liées à la grammaire, aux politiques linguistiques ou à la linguistique TAL des langues de France.

3. Conclusion et perspectives

La vie des langues régionales, leur survie, leur revitalisation sont fortement dépendantes de leur statut et de leurs possibilités d'enseignement. Nous avons vu comment celles-ci ont évolué malgré les réticences rencontrées jusqu'à l'acceptation dans les faits de l'enseignement immersif, lequel est vital pour certaines langues régionales.

Parallèlement, les technologies des langues progressent très vite et vont continuer à évoluer avec la montée en puissance de l'intelligence artificielle.

Les chapitres de ce volume se situent dans ce contexte en évolution et contribuent à cette évolution. Tous les travaux présentés dans ce livre ont utilisé des données et outils numériques et quasiment tous ont contribué à leur développement, que ce soit des dictionnaires, des atlas, des corpus de parole, des analyseurs syntaxiques, etc…

L'accès à des données plus larges a modifié la recherche linguistique sur ces langues, sur leur diachronie, lexique, syntaxe et sur les langues en général. Des faits de langues régionales, comme ceux du breton, du picard, des langues du Croissant ou de l'Occitan se trouvent connus par une plus large communauté.

Les perspectives de la recherche sur les langues régionales s'inscrivent dans le développement général de la linguistique et des technologies des langues au niveau de la France comme au niveau européen. Des structures communes au niveau français ou européen se développent ou se mettent en place pour mutualiser les ressources et les savoir-faires (depuis des réservoirs de données et outils telles que CoCoON, ORTOLANG ou Nakala d'Huma-Num jusqu'aux grandes plateformes ou consortiums européens comme CLARIN, *European Language grid* ou *European Language Equality* qui concerne précisément les langues sous-dotées).

Le principal défi actuel pour les langues régionales reste l'augmentation des ressources disponibles, en particulier des corpus parallèles et annotés. Les performances d'outils d'intelligence artificielle tels que Chat GPT dépendent, en effet, de la quantité de données. Par ailleurs, l'intelligence artificielle et sa capacité d'apprentissage offrent des possibilités intéressantes d'automatisation dans la création des corpus nécessaires en amont de tout développement d'applications.

Nous sommes donc à un tournant où il est crucial d'améliorer l'équipement numérique des langues régionales pour contribuer à leur préservation, à leur vitalité, à leur reconnaissance dans le monde numérique en constante évolution. La recherche en linguistique ne peut que suivre cette évolution qui va lui procurer plus de données, plus d'outils d'analyse, avec, pour conséquence probable, d'infléchir

ses approches et modèles théoriques.

BIBLIOGRAPHIE

CERQUIGLINI Bernard (1999). Les langues de la France : rapport au Ministre de l'Éducation, de la Recherche et de la Technologie et à la Ministre de la Culture et de la Communication, avril 1999, https://medias.vie-publique.fr/data_storage_s3/rapport/pdf/994000719.pdf

(Consulté le 25/02/2023).

JOUITTEAU, Mélanie (2020). Standard Breton, traditional dialects, and how they differ syntactically. *Journal of Celtic Linguistics*, 21 (1),24-74.

KREMNITZ, Georg (éd.) (2013), avec le concours de Fañch Broudic et de Carmen Alén Garabato, Klaus Bochmann, Henri Boyer, Dominique Caubet, Marie-Christine Hazaël-Massieux, François Pic, Jean Sibille. *Histoire sociale des langues de France*. Rennes : Presses Universitaires de Rennes.

KREMNITZ Georg (2020). *La problématique initiale de la liste Cerquiglini et ses effets ultérieurs. Glottopol*, Université de Rouen, Laboratoire Dylis, 34 : 37-45.

MARTEL, Philippe & VERNY, Marie-Jeanne (2020). Les langues régionales au Parlement, ou l'éternel retour. *Glottopol*, Université de Rouen, Laboratoire Dylis, 34 : 69-90.

REY, Christophe (2021). La langue picarde et ses dictionnaires, Collection Lexica, mots et dictionnaires, n°38, Honoré Champion.

ROUQUIER, Jérémy & Michela RUSSO (2021). Représentations et usages de l'occitan dans les *calandretas* du Béarn : une étude de cas. In Simone Pisano *et al.* (éd.), *Plurilinguismo e Pianificazione linguistica*. Alessandria : Edizioni dell'Orso, 155-180.

PARTI SOCIALISTE (1981). La France au pluriel, Paris : Editions Entente.

SIBILLE, Jean (2013). La notion de langues de France, son contenu et ses limites. In Georg Kremnitz, avec le concours de Fañch Broudic (éd.), *Histoire sociale des langues de France*. Rennes : Presses universitaires de Rennes, 45-60.

SIBILLE, Jean (2010). 'Langues de France' et territoires : raison des choix et des dénominations. In Alain Viaut et Joël Pailhé (éd.). *Langue et espace*. Bordeaux : Maison des sciences de l'homme d'Aquitaine, 85-107.

VIAUT, Alain (2020). De 'langue régionale' à 'langue de France' ou les ombres du territoire. *Glottopol*, Université de Rouen, Laboratoire Dylis, 34 : 46-56.

VIAUT, Alain et PASCAUD, Antoine (2017). Pour une définition de la notion de 'langue régionale'. *Lengas revue de sociolinguistique*, 82. Online : https://journals.openedition.org/lengas/1380

WOEHRLING, Jean-Marie (2005). La Charte européenne des langues régionales ou minoritaires - Un commentaire analytique. Strasbourg : Conseil de l'Europe.

Annie Rialland
Laboratoire de Phonétique et Phonologie
CNRS/Sorbonne-Nouvelle
annie.rialland@sorbonne-nouvelle.fr

Michela Russo
UJML 3 & SFL CNRS
Université de Paris 8 (FR)
michela.russo@cnrs.fr

Outils numériques, linguistique et revitalisation des langues (breton, picard)

Chapitre 2

Outils numériques et traitement automatique du breton

Mélanie Jouitteau[1], Reun Bideault[2]

[1]IKER, CNRS, UMR 5478, Université de Pau et des Pays de l'Adour,
[2]Université Bordeaux Montaigne

Abstract

In this article, formal linguist Mélanie Jouitteau and web developer Reun Bideault present a synthesis of the numeric and NLP tools available or in development for Breton. They discuss the resources for its development. NLP of Breton is still objectively poorly developed, but some new tools have just been made available, which opens a real potential for development. We present a state-of-the-art of the field, and we detail how the first tree bank *Universal Dependencies*, created by Tyers & Ravishankar (2018) could be reinforced by 25000 additional glossed sentences in the databank of the wikigrammar ARBRES (Jouitteau (2009-).

Résumé

Dans cet article, la linguiste formelle Mélanie Jouitteau et le développeur web Reun Bideault présentent un état des lieux des outils numériques et des outils pour le traitement automatique du breton, et discutent les ressources à son développement. Le TAL appliqué à la

langue bretonne est encore objectivement peu développé mais de nouveaux outils viennent d'être créés qui ouvrent un potentiel réel. Après un état des lieux de l'existant, nous détaillons comment la première banque d'arbres au format *Universal Dependencies* créée par Tyers & Ravishankar (2018) pourrait être alimentée de 25000 phrases glosées additionnelles provenant de la banque de données de la wikigrammaire ARBRES (Jouitteau 2009-).

1. Introduction

Le chapitre 2 présente un état des lieux des outils numériques disponibles, et est largement redevable aux synthèses pré-existantes de Aubry (2004), Foret & al. (2015), Tyers & Howell (2021:437-438), et pour l'avancement des traductions d'environnement web au rapport de l'Office Public de la Langue Bretonne (OPLB), Kerbrat (2021a,b). Nous avons complété cet état des lieux par une interview de deux chercheurs à l'IRISA, Damien Lolive (3h) pour la synthèse de la voix et Annie Foret (2h) pour la recherche fondamentale, en novembre et décembre 2021. Le chapitre 3 présente les corpus numériques existants, et c'est sous cet angle que nous présentons la wikigrammaire ARBRES (Jouitteau 2009-2021) et évaluons les pas nécessaires à son utilisation pour la construction d'une banque d'arbres en format *Universal Dependencies* pouvant consolider celle de Tyers & Ravishankar (2018). En partie 4, la conclusion synthétise les prospectives et ouvre des pistes de discussion sur les observations des usages, les pratiques de science ouverte et des considérations très concrètes pour leur développement[1].

[1] Nos plus sincères remerciements vont aux chercheurs de l'IRISA Damien Lolive, Gaëlle Vidal et Annie Foret pour le temps qu'ils nous ont consacré, ainsi qu'à Thierry Poibeau (LATTICE, CNRS), Francis Tyers (U. Bloomington, Indiana) et Stefan Moal (U. Rennes II) pour les références fournies, et enfin au centre de formation *Kelenn* à Quimper pour son accueil. L'historique de la genèse de cet article et ses plus récentes mises à jour sont disponibles en ligne dans Jouitteau (2009- : 'Traitement automatique des langues – Breton'), article qui comprend en plus une description des ressources humaines, des pôles de formation et des ressources de financement pour le TAL du breton.

2. État des lieux de l'existant

2.1. Traducteurs et outils pour leur construction

Apertium fournit une interface de traduction à partir de l'analyse morphologique de Tyers (2007-2009) et d'un dictionnaire bilingue (cf. Tyers 2009, 2010a, 2010b, 2015). L'analyseur est sous licence GPL-2 (copyright Francis Tyers 2008-2011, Fulup Jakez 2009-2011, Gwenvael Jekel 2011), et disponible sur le site d'Apertium. Tyers (2010a,b) décrit le système de traduction automatique breton > français basé sur des règles.

Tyers & Howell (2021) ont évalué les résultats de l'analyseur morphologique couplé avec un désambiguïsateur morphologique basé sur une grammaire de contraintes. Ces deux derniers outils sont disponibles en logiciel open-source du projet *Apertium* (GNU GPL 3.0). L'analyseur consiste en un transducteur à états finis qui gère l'interface entre les formes de surface et les formes lexicales (les tags morphosyntaxiques et leurs lemmas). Il permet l'analyse de formes comme leur production. Les homophones sont départagés par un ensemble de règles de désambiguïsations morphologiques basé sur une grammaire de contraintes qui a été développée à partir de corrections des traductions automatisées par un brittophone[2] et Francis Tyers. L'Office Public de la Langue Bretonne diffuse une version en ligne du traducteur d'Apertium, *troer emgefre* dans le sens de traduction br > fr. Le choix de l'Office est de ne pas distribuer le sens inverse de traduction avant une perfection des traducteurs vers le breton, car le risque d'utilisation sans correction par des non-locuteurs est grand, et serait très dommageable.

Il existe d'autres projets de traducteurs. Le site Glosbe propose certaines traductions br <-> fr. Le site *Lexicool.com* regroupe les dictionnaires multilingues breton-autre langue. En utilisant la technologie des réseaux de neurones, l'équipe OPUS-MT de l'université d'Helsinki développe un traducteur automatique multilingue qui comprend un traducteur anglais-breton et breton-anglais.

[2] Ce locuteur n'est pas identifié clairement. Il s'agit peut-être de Fulup Jakez, remercié en note.

2.2. Conjugateur de verbes

Le conjugateur automatique de verbes *DVB, displeger verboù brezhonek* développé par Per Morvan est en ligne depuis juin 2021.

2.3. Détecteur de langue

Foret (2018b) relève une méthode pour les langues celtiques dans Minocha & Tyers (2014) et cite deux détecteurs accessibles qui gèrent le breton: open.xerox.com et G2LI.

2.4. Outils correcteurs

Le compte rendu d'activités de l'IRISA (2001) mentionnait déjà qu'il était "désormais possible d'appeler le dictionnaire [vocal] comme outil de correction orthographique, dans une application de type traitement de texte", et il existe un rapport de projet de l'ENSSAT de 2003 sur le correcteur orthographique breton (Petit 2003). Poibeau (2014) qui fournit une formalisation des mutations consonantiques en utilisant des transducteurs à états finis suggérait leur utilisation pour un correcteur orthographique. Il en existe aujourd'hui plusieurs.

Le correcteur orthographique et grammatical *Microsoft Office 2013*, développé par l'association An Drouizig fonctionne aussi sur *MacOffice 2001*. Le correcteur orthographique *Hunspell*, aussi développé par l'association An Drouizig, fonctionne sur *Adobe Indesign, Firefox, LibreOffice* et *OpenOffice.org* et *MacOSX*. Le correcteur grammatical pour la suite bureautique *LibreOffice* développé par Dominique Pellé avec l'aide le l'OPLB utilise *LanguageTool*, testable en ligne. L'OPLB rapporte une première version de 400 règles, avec repérage des fautes de mutation. Ce correcteur est évalué dans Morvan (2019).

2.5. Dictionnaires en ligne

Il existe de nombreux dictionnaires en ligne du breton contemporain, et un agrégateur de dictionnaires. Certains sont en accès libre, mais peu sont sous licence libre. Menard & Bihan (2016-) et Favereau (1993) comportent des entrées de dialectes traditionnels mais les autres sont plutôt de breton standard.

L'aggrégateur *Geriafurch* développé par Anthony Lannuzel croise les résultats de plusieurs dictionnaires en ligne et en livre un

résultat allégé. Il existe en application téléphone téléchargeable. Sa portée, en 2021, couvre le dictionnaire Brezhoneg21 = KAG (2016), ressource scolaire des sciences et techniques, le dictionnaire *Devri* de Menard & Bihan (2016-), le dictionnaire en ligne de Favereau (1993), celui de Glosbe, *Preder* et finalement *Termofis*, le dictionnaire terminologique de l'OPLB.

Le dictionnaire breton-néerlandais de Jan Deloof (2008-2010) comporte 40,000 entrées. Kevin Donnelly, qui a géré sa mise en interface, considère qu'il s'agit du plus grand dictionnaire libre (GPL) pour une langue celtique (Donnelly 2010).

Le dictionnaire Favereau (1993) comporte 40 000 entrées. La première synthèse de la voix de l'IRISA avait utilisé un algorithme pour en accepter les orthographes multiples. Il n'est pas en licence libre, raison pour laquelle Tyers ne l'utilise pas (Tyers & Howell 2021:440, fn11).

Le dictionnaire Freelang fr <-> br (disponible en ligne ou téléchargeable) de Tomaz Jacquet comporte 37.800 entrées. Tyers (2009) en a importé semi-automatiquement les classes lexicales.

Le dictionnaire br -> fr de Cornillet (2017) est disponible en ligne dans une version corrigée en 2020. Il a été utilisé pour la synthèse de la parole en 2019-2021 à l'IRISA.

Le dictionnaire de l'association *Stur* traduit 22.302 noms du français vers le breton. Il est cherchable en ligne.

Le dictionnaire Favereau (2016-évolutif) est en ligne sous format pdf, avec des dossiers séparés pour chaque lettre initiale. Le copyright propriétaire mentionné sur le site est de 2016, mais l'auteur enrichit l'ouvrage régulièrement et met en ligne les pdfs par lettre du dictionnaire. La date de dernière modification pour chaque dossier est au début de chaque pdf.

La base de données toponymique KerOfis de l'OPLB liste les noms propres noms de lieux.

Le dictionnaire multilingue *Logos* comprend le breton. Il s'agit d'un site collaboratif de traducteurs professionnels sur invitation.

Le dictionnaire multilingue *wiktionnary* comprend le breton avec *wikeriadur*.

2.6. Grammaires en ligne

Le site ARBRES (Jouitteau 2009-) offre une wikigrammaire du breton, que nous prenons le temps de décrire ici brièvement car elle informe la constitution de sa base de données, discutée en partie II.

Il s'agit d'un carnet de recherches rédigé sous forme de grammaire en ligne. Ses buts sont de fournir une description fine et théoriquement informée de la variation syntaxique en breton et un état des lieux permanent et à jour des différentes recherches en syntaxe formelle. Le projet est de créer un pont entre le milieu international de recherches linguistiques, les travailleuses et travailleurs de la langue qui cherchent plutôt une ressource pédagogique, et tout adulte et citoyen curieux de la structure de cette langue parlée par plus de 150 000 locuteurs.

Le site propose plusieurs entrées, une grammaire descriptive standard ainsi qu'une grammaire formelle qui organise une description de leur impact théorique pour la linguistique formelle. Il comporte une bibliographie générale qui se veut exhaustive pour les recherches en syntaxe, et comporte une centrale d'élicitations par laquelle la communauté internationale de recherche peut co-construire des protocoles avec Mélanie Jouitteau qui opère l'élicitation sur le terrain et poste les résultats en ligne. Fin 2021, la wikigrammaire rassemblait plus de 2000 articles thématiques. Le site est ouvert en écriture et pour les commentaires (pour l'aspect science ouverte et science citoyenne de la wikigrammaire, se reporter à Jouitteau, M. 2013b). L'OPLB a été consulté dès 2008 afin de recueillir ses vœux en termes de développements, vœux qui ont influencé la genèse de la wikigrammaire, en particulier la constitution en format récupérable pour une base de données utilisable en TAL.

Il faut ajouter à cette grammaire en ligne les ouvrages dédiés à des parlers locaux particuliers. La partie grammaticale du blog de collecte *Brezhoneg Bro-Vear* (Yekel, Georgelin & Ar C'hozh 2015-2021) est maintenant considérable. Une ressource non négligeable provient aussi des plus récentes thèses et monographies universitaires dont les textes sont disponibles en ligne. Celles-ci sont recensées dans l'inventaire des grammaires de la wikigrammaire ARBRES.

2.7 Traduction de logiciels, réseaux sociaux, jeux, etc.

Diverses applications utilisables sur internet sont traduites, souvent partiellement, en breton mais cela reste insuffisant pour créer un environnement informatique immersif. Pour le web, les interfaces utilisables sont extrêmement limitées en nombre et en pourcentage de traduction. *Wordpress* est le système de gestion de contenu (abrégé CMS pour l'anglais *Content Management System*) le plus utilisé dans le monde (40% des sites). Il n'est traduit, pour la version plus récente fin 2021 (V. 5.8.x), qu'à 18 %. Ce travail est suivi par 7 personnes. Pour comparaison, les versions en basque sont traduites à 96 %, en occitan à 53 %, pour respectivement 80 et 13 participant.e.s. Ce CMS s'appuie sur des plugins indispensables à une utilisation élargie, où le niveau de traduction est encore plus faible lorsqu'il existe. Reun Bideault, développeur web, considère que l'exemple de *Wordpress* est actuellement généralisable à tous les outils web libres et propriétaires, raison pour laquelle les professionnel.le.s du web ne peuvent actuellement fournir un produit fini et surtout évolutif permettant de travailler en breton à un coût supportable.

L'OPLB fournit la traduction en breton des données du CLDR (*Common Locale Data Repository*) d'Unicode, qui regroupe l'ensemble des paramètres régionaux à destination des applications informatiques. Lors de la publication de la version 38 du CLDR fin 2020, Kerbrat (2021a,b) estime que le breton a atteint l'avant dernier niveau de couverture (*Moderate++*). Tomaz Jacquet rend disponible en ligne sous différents formats un dictionnaire trilingue breton, français anglais du vocabulaire utilisé dans les logiciels. Fin 2021, sont disponibles en environnement traduit :

- une suite bureautique (*LibreOffice*) qui est associable aux correcteurs orthographiques et grammaticaux décrits plus haut
- un logiciel pour la navigation web (*Firefox*)
- un logiciel pour l'échange de courriels (*Thunderbird*)
- quelques logiciels multimédia

(*VLC* pour la vidéo, *Clementine* pour la musique)

- quelques logiciels d'édition graphique

(*Inkscape, Gimp, Tuxpaint*)

Pour les réseaux sociaux, *Facebook* est utilisable en breton depuis 2014 (Ar Mogn 2015). *Mastodon*, réseau semblable à *Twitter* mais libre de droits, fait actuellement l'objet d'un projet de traduction participative.

Il existe une version bretonne pour quelques applications smartphone, en plus de l'autocorrection et la prédiction de mot en breton sur le clavier virtuel Microsoft *SwiftKey*:

- *Firefox* (iOS et Android), navigateur web
- *K-9 mai*l (Android), client de messagerie
- *Vanilla Music* (Android), lecteur musical

2.8. Synthèse vocale

La Région Bretagne à travers l'OPLB a financé à hauteur de presque 200.000 euros la construction d'un moteur de synthèse de la parole (breton KLT standard[3], un homme, une femme). Le projet était dirigé en TAL par Damien Lolive et Gwénolé Lecorvé du laboratoire *Expression* de l'ENSSAT à Lannion en collaboration avec la maison d'édition *Skol Vreizh*. Le programme de synthèse de la voix a été livré à l'OPLB en mars 2021.

Les deux locuteurs qui ont prêté leur voix ont été élevés en milieu brittophone trégorrois, à tendance plus standard pour Annaig Kervella (fille de Frañsez Kervella, auteur de la grammaire standard de référence), et plus traditionnelle pour Pascal Lintanf (avec influences léonardes pour ce dernier)[4]. Chacun des deux corpus oraux produits durent un peu plus de 20h, ils ont été constitués par tâche de lecture d'un corpus de breton standard constitué principalement de discours journalistique, et de textes littéraires (environ 10% sont des dialogues, joués avec expressivité modérée). Le corpus écrit correspondant a été normalisé (écriture en lettres des nombres et acronymes, prononciation différenciée des noms propres, etc.) puis, un panel d'experts choisi par *Skol*

[3] Les trois initiales bretonnes KLT réfèrent aux dialectes cornouaillais, léonard et trégorrois, ce qui exclut le dialecte vannetais. Les propriétés originales des trois dialectes majeurs du KLT sont gommées pour obtenir un standard d'usage entre les trois.

[4] Pascal Lintanf est par ailleurs l'auteur d'un mémoire universitaire sur la phonétisation du breton (An Intanv 1994).

Vreizh et principalement le second locuteur Pascal Lintanf ont constitué un répertoire de règles de prononciation. Un lexique donnant une prononciation standard accentuée en API a été constitué par arbitrage entre plusieurs sources de lexiques phonétisés et écrits en orthographe unifiée : le dictionnaire *An Here* de Menard & Kadored (2001), le dictionnaire de Francis Favereau (2015) et sa dernière version consultable en ligne Favereau (2016-évolutif) consultables en ligne, et de Gérard Cornillet (2017). D'autres données y ont été intégrées comme celles des noms propres, fournies par l'OPLB, et celles rencontrées dans les corpus constitués. Pour dix mois, Gaëlle Vidal, ingénieure d'études, a défini et enregistré un corpus de textes, sélectionné les locuteurs, et procédé aux enregistrements et à leur découpage en phrases. Hassan Hajipoor, ingénieur de recherche, a ensuite eu 18 mois (dont un confinement) pour construire un phonétiseur, comprenant un modèle de la syllabe et de l'accentuation qui a pu être paramétrisé pour les exceptions, et entraîner un réseau de neurones sur le corpus oral et le dictionnaire. Le système en end-to-end livre le fichier son à partir de la phrase écrite. La technique ne permet pas de prendre en charge la structure informationnelle et la prosodie associée, mais l'accentuation de mot et les phénomènes de frontière de mot comme la mutation ou le sandhi sont pris en charge[5,6].

L'OPLB a créé un poste de chargé de développement du numérique pour sa diffusion.

3. Corpus numériques existants

La langue bretonne n'est pas une langue minorisée pour laquelle manquent les corpus, mais ils ne sont pas tous immédiatement accessibles pour des traitements automatiques de la langue (copyright restrictif, éditions épuisées, documents non-

[5] Pour un historique détaillé de l'époque pionnière de la synthèse de la voix dans les années 90 avec Favereau, IRISA & TES. 1999, se reporter à Aubry (2004). Il semble aussi avoir existé un correcteur de prosodie (Mocquard 1999, 2001, Guillou 2000) et un entraîneur prosodique (Aubry 2000, 2004).

[6] L'OPLB, pour le projet de synthèse de la voix, n'a pas recouru à son conseil scientifique.

OCR, corpus numériques à URLs non-stables, etc.). Ci-dessous, sont listées les ressources a-priori disponibles au TAL, ou déjà utilisées[7,8].

3.1. Corpus non-glosés

Thierry Poibeau (c.p.) signale 23 Mo de données brutes de texte en breton, sans annotations, dans le corpus Oscar, qui sert actuellement pour mettre au point des modèles pour le TAL par modèles neuronaux (type *Bert*).

Les archives de traduction de l'Office constituent un corpus bilingue qui a déjà été utilisé pour le traducteur automatique (Tyers 2009). Ar Mogn (2015:15m40s), co-directeur de l'OPLB, mentionne un corpus de 43000 phrases bretonnes traduites. Kerbrat (2021a,b) l'estime à "environ 1 million de mots". Le corpus de traductions de l'OPLB, corpus de phrases en breton, et corpus de phrases en français, sont téléchargeables et libres de droit.

L'association *An Drouizig* revendique pour la construction de son correcteur orthographique *Difazier [ver 4.4]* l'analyse d'un corpus linguistique de 20 millions de mots bretons, qui comprend au moins celui de l'OPLB.

Donnelly (2010) mentionne sa création avec l'aide de Rhisiart Hincks à Aberystwyth d'un corpus parallèle de 3500 phrases en

---

[7] Jouitteau (2009-:'corpus') fournit une liste plus exhaustive de corpus de breton, plus tournés vers l'apprentissage humain.

[8] Leixa & al. (2014) ont essayé de recenser les corpus utilisables en TAL pour plusieurs langues minoritaires de l'État français. L'approche est un brin parachutée. Ils comptent pour le breton 420 corpus utilisables, dont 403 corpus oraux et 17 corpus textes. "On trouve parmi ces ressources de petits enregistrements audio de quelques minutes, mais également d'importants corpus alignés pouvant servir de base à des technologies de la langue. Parmi les ressources audio, nous avons par exemple les enregistrements effectués par M. Jean Le Dû lors d'une enquête dialectologique réalisée en Bretagne, en vue de constituer le *Nouvel Atlas Linguistique de la Basse-Bretagne*" (Le Dû 2001). Cependant, à l'écoute, ces enregistrements sont difficilement utilisables car les élicitations sont effectuées à partir de gestes physiques dénotant des mots à trouver, or cette information gestuelle manque évidemment aux enregistrements. La prosodie interrogative des locuteurs est typiquement celle de quelqu'un qui cherche à deviner un mot, et la confirmation que son choix est le bon. L'identification précise de l'ensemble des corpus listés dans Leixa & al. (2014) "est disponible sur le CD qui est joint au rapport" à la DGLFLF.

breton et gallois, organisé dans un *Breizh-Llydaw Sentence Bank* (license GPL), et accompagné d'un dictionnaire de 1200 mots.

Il existe aussi des corpus parallèles multilingues, comme la *Déclaration des Droits Humains* de l'OHCHR et la traduction du *Petit Prince* de Saint-Exupéry.

3.2. Corpus sonores

On a vu que l'IRISA à Lannion a constitué un corpus de plus de 40h de la synthèse de la voix dont les phrases ont été individuées.

Il existe aussi différents sites de collecte de données brutes, par des collectifs associatifs à la durée de vie variable. Ces derniers n'en sont pas pour autant négligeables. Ils constituent des travaux considérables, avec traductions des données dialectales ou explicitation en standard, et des traductions en français. A notre connaissance, il n'existe aucune aide ou soutien organisée à ces travaux pionniers, même pour l'hébergement.

- les *Dictionnaires bretons parlants*

(Cheveau & Kersulec 2012-évolutif)

- la *Banque sonore des dialectes du breton*

(Desseigne & al. 2013-2018)

- *Brezhoneg Bro-Vear*

(Yekel, Georgelin & Ar C'hozh 2015-2021)

Common voice de *Mozilla* a lancé en 2018 un module de collecte de la parole en crowdsourcing, qui permet aux utilisateurs d'enregistrer leur propre parole, ou d'évaluer les enregistrements laissés par d'autres (9h d'enregistrements validés en 2021).

Les enregistrements audio de corpus libres existent dans les différents dialectes du breton, stockés dans les archives des différentes radios bretonnes, sous des formats différents allant de l'analogique au numérique. Les fichiers audio des enquêtes du *Nouvel Atlas Linguistique de la Basse-Bretagne*" (Le Dû 2001) devraient pouvoir être au moins partiellement utilisées, mais cela demanderait un tri méticuleux (voir note de bas de page numéro 8).

3.3. Banque d'arbres Universal Dependencies

Il existe pour le breton des corpus glosés traduits. Ils comportent des phrases en breton traduites mot-à-mot et traduites globalement. La traduction mot-à-mot est une glose, qui contient des informations sur l'élément linguistique en question (catégorie grammaticale, fonction, mutation déclenchée, etc.).

La notation universelle qui émerge en 2021 comme recommandation pour les banques d'arbres est celle de *Universal Dependencies* ("format UD"), qui propose un jeu de 17 parties du discours (*parts of speech*, POS) et deux douzaines de fonctions grammaticales. Certains des choix fondamentaux de ce format, comme de subordonner les catégories fonctionnelles aux catégories lexicales ne sont pas soutenus linguistiquement, mais la conversion de structures UD à des structures syntaxiques en constituants est cependant automatisable en grande partie (voir discussion par Osborne & Gerdes 2019).

Tyers & Ravishankar (2018) ont créé la première banque d'arbres réalisée au format UD. Ce corpus tree-bank breton de 10 000 tokens est annoté manuellement. L'analyseur morphologique de Tyers (2009) pour *Apertium* a été utilisé pour la tokenisation et l'annotation morphologique. Ci-dessous, un exemple de codage de la banque d'arbres de Tyers & Ravishankar (2018). On y trouve la phrase bretonne en entier et sa traduction en français, puis une glose mot-à-mots, répartie en lignes. On trouve le lemma (forme comme donnée comme pour un dictionnaire), les étiquettes de parties du discours (POS tags) catégorielles et leurs sous-spécifications, des informations sur la structure choisie qui fait dominer le verbe lexical (noté *root*), ainsi que les traits de la morphologie flexionnelle.

```
# sent_id = apertium.vislcg.txt:1:0
# text = N'int ket aet war-raok.
# text[fra] = Ils n'ont pas progressé.
# labels = to_check
1 N'   ne    ADV adv Polarity=Neg  4 advmod _ SpaceAfter=No
2 int bezañ AUX vblex Mood=Ind|Number=Plur|Person=3|Tense=Pres|VerbForm=Fin 4 aux _ _
```

```
3 ket ket ADV adv _ 4 advmod _ _
4 aet mont VERB vblex Tense=Past|VerbForm=Part 0 root _ _
5 war-raok war-raok ADV adv _ 4 advmod _ SpaceAfter=No
6 . . PUNCT sent _ 4 punct _ _
```

Le texte du corpus consiste en 888 phrases provenant d'exemples de grammaires, de phrases tirées de *wikipedia* en breton, de textes administratifs de l'OPLB et de textes du journal *Bremaik*, avec deux chansons traditionnelles. La composition est détaillée dans Tyers & Howell (2021:450). Certains codages sont étranges, et pourraient être discutés. Par exemple dans les phrases *En em c'houlenn a ran* /se demander R fais/ 'Je me demande' ou *En em gannet out c'hoazh?* /se battre es encore/ 'Tu t'es encore battu?', le pronom proclitique réfléchi *en em* est noté / det + aux /, alors qu'en format UD les réfléchis et réciproques sont étiquetés comme des pronoms (PRON).

La banque d'arbres a d'ores et déjà servi à une expérimentation pour la construction de grammaires de dépendances afin de construire un outil de lecture augmentée (Martinet 2021), et à évaluer un analyseur morphologique et la grammaire de contraintes dans Tyers & Howell (2021:450).

3.4. Une banque de données en wikigrammaire

La wikigrammaire ARBRES Jouitteau (2009-) est sous licence creative commons CC BY-NC-SA. Elle a été prioritairement développée pour un lectorat humain en ligne, mais comporte une masse importante de données du breton localisées par leur dialecte, traduites et glosées mot à mot, organisées dans un format numérique qui est destiné à terme à alimenter un traitement automatique.

Fin 2021, le site contient plus de 75 000 tableaux de type "prettytable" qui ont servi à aligner chaque mot breton breton avec sa glose en français, et à aligner l'ensemble avec une traduction globale de la donnée en français, ainsi qu'avec une typification dialectale du locuteur source. En moyenne, nous estimons que chaque donnée originale en breton a été employée trois fois dans

des endroits différents de la grammaire, ce qui donne une estimation grossière de 25 000 phrases originales en breton.

La wikigrammaire utilise des exemples tirés de plus de mille ouvrages de recherche scientifique sur le breton, des données de séances d'élicitation avec des locuteurs natifs effectuées par Mélanie Jouitteau, à son initiative ou à la demande d'autre linguistes, et dont les résultats bruts sont disponibles en ligne dans la centrale d'élicitation avant exploitation, ainsi que de 399 sources de corpus écrits différents, du vieux breton aux dialectes bretons modernes, breton standard y compris. Les dialectes y sont mentionnés comme tels, et la typification dialectale est associée à chaque donnée, donc il serait possible pour un traitement automatique de mettre de côté les états anciens de la langue, et les quelques données comparatives tirées de langues autres (hébreu, basque, occitan, etc.). Les données du breton ont servi à l'établissement d'une grammaire donc elles ont été sélectionnées pour représenter la plus grande variété possible de structures. La graphie est riche car l'orthographe des sources diverses a été respectée - les gloses, elles, sont en orthographe unifiée *peurunvan*. Certaines données de ARBRES, plutôt rares à l'échelle du corpus, comportent en plus une ligne de code donnant la tokenisation de la donnée en API, ou dans des orthographes originales renseignant la prononciation. Ces scripts peuvent être mis de côté car ils sont signalés par une balise de mise en couleur verte (< (/) font color=green >). Les traductions en français viennent soit de corpus déjà bilingues, soit sont effectuées par Mélanie Jouitteau (native français, breton L2).

Pour comparaison avec le treebank UD de Tyers & Ravishankar (2018), je code ci-dessous le même exemple donné plus haut, dans sa forme visible aux utilisateurs et le code que cela nécessite. Le codage dans ARBRES obtient une visualisation comme ci-dessous pour les utilisateurs.

(1) N'int ket aet war-raok.
ne[1] sont pas allé sur-avant
'Ils n'ont pas progressé.'
Dialecte, source référencée de la donnée

Chaque exemple est donné en breton, glosé et traduit. La ligne de gloses fournit la traduction littérale, mot-à-mots en français. Elle comprend une mention des mutations consonantiques en superscript sur son élément déclencheur (ici, l'adverbe négatif *ne* qui provoque une lénition dans tous les dialectes, codée 1 en superscript. La mutation est notée même si, en l'occurrence, elle ne peut pas avoir ici d'effet car l'initiale du verbe qui suit n'est de fait pas mutable. Les gloses en français ne montrent d'accord que si l'élément en breton en montre (cf. *allé*). Parfois, le glosage au plus près de la composition bretonne crée en français des approximations (cf. *sur-avant*). La troisième ligne visible du tableau fournit la traduction globale de la phrase en français standard. Pour obtenir une telle visualisation, alignement des gloses compris, le code wiki est comme ci-dessous (abstraction faite de la balise diu superscript de mutation).

```
0 {| class="prettytable"
1 |(1)|| N'int || ket || aet || war-raok.
2 |-
3 ||| [[ne]][[1]] [[COP|sont]] || [[ket|pas]] || [[mont|allé]] || [[war-raok|sur-avant]]
4 |-
5 |||colspan="10" |'Ils n'ont pas progressé.'
6 |-
7 ||||||||colspan="10" |Dialecte, source référencée de la donnée
8 |}
```

Dans le code, les colonnes (||) de la première ligne fournissent un premier découpage de la donnée bretonne. Cette ligne comporte la ponctuation. Le découpage y est inégal, souvent prosodique car les éléments marqués d'une apostrophe ou d'un tiret n'y sont qu'exceptionnellement séparés. Il découpe aussi parfois des blocs de constituants syntaxiques. La seconde ligne visible pour l'utilisateur est la ligne 3. C'est la ligne de gloses qui fournit une tokenisation plus fine et la lemmatisation. Avec l'exemple de la

négation et de sa copule, on voit que le découpage en double crochets dessine alors les sous-parties du découpage de la première ligne. Les tokens atomiques sont séparés, les clitiques y sont ainsi séparés de leur hôte.

Pour que les gloses soient cliquables pour les utilisateurs, le script wiki nécessite que chaque traduction mot-à-mot, la glose, soit associée à une adresse d'article dans la grammaire. Dans la syntaxe wiki, ce script est ordonné comme suit: [[adresse du lien|glose]]. C'est ainsi que grâce à un script [[mont|allé]], l'utilisatrice qui clique sur la glose *allé*, visible pour elle juste sous le mot breton *aet*, ouvre la page du site dédiée au verbe *mont* 'aller'. Ce script, pour un format UD, fournit le lemma. Ce lemma est associé à la traduction française du token aligné en colonne avec lui. Dans le cas de la préposition composée *war-raok* /sur-avant/ 'en avant', un seul lemma lui est associé.

```
0 {| class="prettytable"
1 |(1)|| mot 1' mot 2 || mot 3 || mot 4 || mot 5-mot 6.
2 |-
3 ||| [[lemma breton 1|français pour lemma 1]][[mutation déclenchée]] [[lemma 2|français pour lemma 2]] || [[lemma 3|français pour lemma 3]] || [[lemma 4|français pour lemma 4]] || [[lemma 5|approximation française pour lemma 5]]
4 |-
5 |||colspan="10" |'Traduction de la phrase en français.'
6 |-
7 ||||||||colspan="10" |''Dialecte'', source référencée de la donnée
8 |}
```

Le lemma breton est donné sous sa forme non-dérivée, ce qui signifie dans cette langue celtique que le lemma est donné au singulier pour un nom comptable mais au pluriel pour un nom collectif. Pour la flexion verbale, le lemma donné est, par convention, la forme infinitive dans la wikigrammaire comme dans UD. Il y a une petite divergence avec le format UD pour les formes qui ont des racines supplétives au comparatif de supériorité comme *gwell* ou *gwelloc'h* 'mieux', ou *gwazh* ou *gwashoc'h* 'pire'. UD

recommande de leur assigner le lemma non-comparatif ce qui donnerait *gwelloc'h* 'mieux' > [[mat|bien]].[[-oc'h|plus]] et *gwasoc'h* 'pire' > [[fall|mal]].[[-oc'h|plus]], alors que la wikigrammaire a prévu de dédier un article à chaque racine irrégulière, ce qui est géré pour l'instant par des redirections ([[-oc'h|mieux]]). Ce pourrait être régularisé assez facilement.

UD requiert que les lemmas soient fournis sous la forme de surface canonique, ce qui pose le problème des formes ambiguës, concrètement en breton les verbes infinitifs et les noms déverbaux, ainsi que les noms différenciés par leur genre en situation (*pal, ar pal* 'le but', *pal, ar bal* 'la pelle' ou *taol, an taol* 'le coup', *taol, an daol* 'la table'). Dans le dictionnaire en ligne Menard & Bihan (2016-), ces ambiguïtés sont prises résolues par un système de spécifieurs numériques assez régulier (*pal.1*, *pal.2*) mais le format UD recommande de privilégier les formes de surface comme lemmas. UD propose de classer ces homonymes dans la colonne MISC dans l'attribut optionnel LId (LId=can-1). Le désambiguïsateur morphologique de Tyers & Howell (2021) semble pouvoir se charger des homophones. Ce dernier pourrait peut-être être solidifié par la liste des pages de désambiguïsation qui liste dans la wikigrammaire les suffixes pouvant être ambigus.

Les mots fusionnés sont un ensemble de plusieurs mots syntaxiques qui apparaissent en breton comme un mot opaque. Ils sont traités en ligne de glose comme des tokens distincts reliés par un point. Ainsi, la préposition *e* devant un article défini est notée *en* en ligne 1 et glosée : [[P.e|dans]].[[art|le]]. La plupart des prépositions peuvent recevoir un pronom objet incorporé - on les appelle les prépositions fléchies. La préposition fléchie *ennon* 'en moi' est glosée [[P.e|dans]].[[pronom incorporé|moi]]. La préposition *ganin* 'avec moi' est glosée [[gant|avec]].[[pronom incorporé|moi]], ce qui permet de récupérer deux formes différentes de pronom incorporé 1SG: *-in* et *-on*, et d'associer chacune avec la préposition qui la déclenche.

En ligne de glose, le découpage en tokens descend au niveau morphologique dans la mesure où le permet son lectorat prioritairement humain. Celui-ci a témoigné régulièrement d'une difficulté d'accès à des formes trop décomposées, ou à des abréviations linguistiques pourtant communes de type 3SG, 3PL.

Ces abréviations sont la plupart du temps évitées dans ARBRES, et la dérivation morphologique est inégalement prise en charge dans les gloses de la wikigrammaire. Lorsqu'un seul suffixe est repérable, le découpage donne directement le suffixe en question dans la glose, mais lorsque plusieurs suffixes forment une finale complexe, le lemma donné est directement cette finale complexe. Le nom *distresadur* 'transformation' est glosé [[di-, dis-|trans]].[[tres|form]].[[-adur|ation]]. C'est dans la page de la finale complexe *-adur* que la finale est décomposée dans ses différents suffixes. Le système de catégorisation de pages permet de générer automatiquement la liste des finales complexes et la liste des suffixes répertoriés dans le site.

La dérivation flexionnelle est prise en charge pour les pluriels des noms. Pour les pluriels simples, le morphème pluriel final apparaît séparé d'un point. Ainsi, le nom pluriel *krouadurioù* 'enfants' est glosé [[krouadur|enfant]].[[-ioù (PL.)|s]]. En breton, les pluriels dits "pluriels internes" ont la propriété de modifier leur racine. Le nom pluriel *bugale* 'enfants' est glosé [[bugel|enfant]].[[pluriel interne|s]], avec le lemma qui est la forme de surface au singulier, et le pluriel qui renvoie l'utilisateur à la page sur les pluriels internes. Les morphèmes porte-manteaux de la flexion verbale, les traits de conjugaison, ne sont pas non plus donnés en glose. Ces traits de flexion verbale sont calculables par la traduction française associée, qui, elle, est donnée fléchie dans les gloses. Les traits UD (UD features) sont donc récupérables dans la mesure où la morphologie verbale française est assez riche. La matrice de traits "Mood=Ind|Number=Plur|Person=3|Tense=Pres|VerbForm=Fin" du verbe breton *int* 'sont' peut-être récupérée par la glose en français *sont*. Cette carence dans la glose de ARBRES pour la flexion verbale pourrait en principe aussi être supplée par les données de DVB, *displeger verboù* développé par Per Morvan.

Un cas difficile et intéressant est posé par la tempête de variation morphologique (et syntaxique) dans le verbe et auxiliaire 'avoir'. En (2), ce verbe précédé de la négation *ne* sous sa forme proclitique est orthographié de manière discontinue, *o dez*. Il comporte les traits du sujet interprété sur sa gauche avec un pronom 3PL *o* sous une forme qui semble oblique, puis d'une initiale /d-/ typique des personnes 3 (au singulier comme au pluriel;

en de(v)ez 3SGM, *he de(v)ez* 3SGF, *o de(v)ez* 3PL). La racine marque la trace de la forme dite d'habitude, qui n'est pas interprétée ou produite dans toutes les variétés sur ce verbe. On pourrait, dans le même contexte syntaxique, trouver *n'o deus ket* en breton standard, la notion d'habitude étant convoyée par un présent à lecture générique. Ceci implique qu'un glosage précis nécessite d'être en mesure de vérifier pour chaque variété si le morphème comprend réellement ces traits, en syntaxe comme en sémantique. Enfin, la finale pourrait être, selon les analyses, une racine dénuée à sa droite de morphème d'accord, un accord 3SG réalisé avec un élément qui n'est pas le sujet, ou encore un morphème d'accord par défaut qui ne fait qu'emprunter la morphologie 3SG et qui apparaît lorsque le sujet est exprimé ailleurs (se reporter aux analyses formelles du système d'accord).

(2) Ha forzh boued n'o dez ket...
et beaucoup nourriture ne 3PL 3.a pas
'Et ils n'ont pas beaucoup de nourriture.'
Vannetais, Herrieu (1994:90)

Ce problème n'est pas facilement écartable car certains dialectes centraux ont, de toute façon, pour une sous-partie du paradigme, un morphème d'accord à droite du composé (*memp* 'nous avons'), dialectes dans lesquels peuvent exister en plus des règles d'accord différentes (*ni meump* /1PL 1.racine.1PL/ vs. *ni neus* /1PL 3.racine.3SG/ ou /1PL 3.racine.Ø/, 'nous avons'). Les gloses dans la wikigrammaire reflètent la diversité des données au plus près de ce qu'on en comprend scientifiquement, et cela peut être un frein à la conversion automatique. Les buts d'un traitement automatique peuvent nécessiter de faire abstraction de la variation et de se contenter de stocker les formes diverses en lien avec leur traduction française.

Les traits de tous les types de pronoms sont récupérables en glose. Le pronom fort indépendant (pfi) 1SG *me* 'moi' est noté en glose [[pfi|moi]], Le pronom fort indépendant 2SG *te* 'toi' est noté en glose [[pfi|toi]], etc. De même, le déterminant possessif (POSS) *ma* 'mon, ma', qui déclenche une mutation mixte (codée 2 en

superscript), est glosé [[POSS|mon]]^{[[2]]} dans la plupart de ses occurrences. Cependant, comme le site documente la variation dialectale, les occurrences du cornouaillais de Locronan documentées dans la grammaire, où ce possessif déclenche une lénition (codée 1 en superscript), sont glosées [[POSS|mon]]^{[[1]]}.

La morphologie flexionnelle n'impacte qu'exceptionnellement les adjectifs bretons par suffixation (*mezvez* 'saoule', glosé [[mezv|saoul]].[[-ez (F.)|e]]). Cependant, la qualité, présence ou absence de mutation sur l'adjectif renseigne sur les traits du nom qu'il modifie. En ligne de glose, la traduction de l'adjectif en français révèle les traits obligatoirement interprétables: *an hini vrav* la belle' est glosé :

[[art|un]] [[hini|celui]]^{[[1]]}[[brav|belle]].

Cet exemple permet aussi de noter que les rares éléments qui n'ont pas d'équivalent en français comme la tête nominale sémantiquement générique *hini* sont traduits en glose par une approximation qui a été jugée commode par le lectorat humain.

On a vu que la ligne de gloses comprend, balisées en superscript (<(/)sup>), les mutations morphosyntaxiques associées à chaque élément qui les déclenche. on marque par le chiffre 1 pour la lénition, 2 pour la spirantisation, 3 pour la mutation durcissante, 4 pour la léniprovection et 5 pour la mutation réduite. Les consonnes épenthétiques du breton sont marquées +C en superscript dans la glose. Il arrive que le découpage morphologique d'un mot breton nécessite de mentionner une consonne épenthétique dans la glose en français. Elle est alors écrite, et non-cliquable puisque ne correspondant à rien en breton (*kozhni* 'vieillesse' est glosé [[kozh|vieil]].l.[[-ni, -oni|esse]]).

Le format UD comporte en tout 17 étiquettes de parties du discours (POS tags). Le code de la wikigrammaire ne fournit qu'exceptionnellement la catégorie grammaticale des éléments directement en glose. Les 5 formes du verbe 'être' et la variation dialectale de leur distribution ont nécessité dans la grammaire un glosage hybride, parfois morphologique (*eo*, *a* zo, *emañ*, *ez eus*, *vez*), parfois syntaxique (COP renvoie à l'article sur l'emploi syntaxique de la copule) ou même sémantique (le signe E en adresse renvoie à l'article sur la copule existentielle). La catégorie des éléments est

cependant toujours récupérable automatiquement par les catégorisations de pages (*eo* => auxiliaire, car l'article de la wikigrammaire intitulé *eo* est catégorisé dans le site comme une page concernant un auxiliaire. Tous les éléments sont ainsi catégorisés via la page qui leur est dédiée, par exemple les adjectifs, mais aussi avec une granularité plus fine dans la mesure où ils ont un comportement grammaticalement distinguable, les adjectifs de couleur (voir la liste des catégories).

Ci-dessous, j'inventorie les catégorie UD et je détaille pour chacune les équivalences sur la wikigrammaire, en ajoutant une estimation des nombres de membres de chaque catégorie fin 2021. Ces chiffres vont progresser à l'avenir, surtout pour les catégories lexicales, au fur et à mesure que des exemples nouveaux alimenteront la grammaire.

- ADJ = adjectif. Ils sont listés dans la wikigrammaire dans la (238 membres), auxquels on ajoute les numéraux ordinaux, les participes (une partie sont mentionnés en glose par la dérivation du suffixe *-et*).
- ADP = adposition (préposition et postposition). Ils sont listés dans la wikigrammaire dans la (158 membres) et dans la (11 membres)
- ADV = adverbe. Ils sont listés dans la wikigrammaire dans la (219 membres)
- AUX = auxiliaire. Ils sont listés dans la wikigrammaire dans la (18 membres)
- CCONJ = conjonctions de coordination. Ils sont listés dans la wikigrammaire dans la liste des conjonctions (12 membres)
- DET = déterminants. Les déterminants sont encore à catégoriser dans le corps de la wikigrammaire, qui comprend cependant la liste des quantifieurs (56 membres). Il faut rajouter les deux articles, défini *an, al, ar* et indéfini *un, ul ur*, les déterminants possessifs et le complémenteur *peseurt*. Attention, les pages thématiques de la grammaire ont été catégorisées sous le titre "articles", en opposition aux "fiches" de linguistique formelle.

- NOUN = nom. Ils sont listés dans la wikigrammaire dans la (799 membres)
- VERB = verbe. Ils sont listés dans la wikigrammaire dans la (354 membres), auxquels on peut ajouter la liste des modaux (sauf peut-être *dav*, *ret* et *arabat* qui ont plutôt une distribution adjectivale), et retrancher les verbes légers *-a*, *-at* et *-aat* qui ont une distribution suffixale.
- SCONJ = conjonction de subordination. Dans la wikigrammaire, ils sont compris dans les complémenteurs.
- PART = particule. La particule préverbale (*rannig*) est signalée en glose par la lettre R, suivie lorsque le dialecte le permet de la mutation associée à cette particule. Attention, UD classe les particules Q des questions polaires, de 'est-ce que', dans les particules, qui sont dans la wikigrammaire des complémenteurs.
- NUM = numéral (numéraux cardinaux, car les ordinaux sont classés avec les adjectifs).
- INTJ = interjection. Certaines sont signalées directement en gloses, d'autres ont chacun une page dédiée qui est catégorisée comme interjection (liste des interjections).
- PRON = pronom. Les pronoms ne sont pas identifiés individuellement dans les gloses. Seul le type du pronom y est spécifié (pronom fort indépendant, pronom écho, pronom incorporé, etc.).
- PROPN = nom propre. Quelques noms propres sont mentionnés comme tels en glose, mais cette pratique est récente sur le site. Il est plus sûr de passer par les recensements déjà établis par d'autres programmes (Tyers 2008 les avait extraits de Wikipedia), ou de s'appuyer sur la majuscule en graphie pour les récupérer.
- PUNCT = ponctuation. Cette information est présente en graphie en ligne 1, et devrait avoir un parallèle dans la traduction française.
- SYM = symbole. Il s'agit de symboles écrits ne sont pas codés à ce jour dans la wikigrammaire.
- X = autre. Cette notation n'a pas été nécessaire.

En dehors du système d'annotation des données, le site a nécessité pour son développement interne des outils et listes qui pourraient directement alimenter les entraineurs d'algorithme, comme:

- la liste des pages de désambiguïsation qui liste les suffixes pouvant être ambigus
- la liste des finales de mots qui liste les ensembles de suffixes existants, et les décompose
- le liste des redirections de pages, qui gèrent les différences d'orthographe ou de dialecte. L'exploitation de cette dernière liste nécessiterait cependant de nettoyer les redirections concernant les ouvrages de recherche et les abréviations.
- des inventaires trilingues par catégories grammaticales: inventaire des noms, inventaire des adjectifs, inventaire des adverbes, inventaire des prépositions (très partiel), inventaire des verbes modaux, inventaire des verbes lexicaux, et par sous-catégories, inventaire des verbes inaccusatifs, inventaire des verbes inergatifs.
- trois glossaires (en anglais, breton et français) de plus de 250 termes de grammaire descriptive et formelle, liés chacun à des définitions illustrées par des faits du breton.

4. Conclusion et pistes de discussion

4.1. Prospective et repérage des besoins

Si on synthétise les prospectives dessinées par les différents secteurs, des pôles de demandes émergent assez nettement.

Pour le domaine de l'écrit, Annie Foret (laboratoire LOUSTIC, Rennes I) a mené un repérage des besoins de développement des outils du TAL pour le breton en 2017-2018 (Foret 2018a,b). L'enquête a consisté initialement en huit entretiens libre/semi-orienté d'1h30 d'enseignant.e.s et d'apprenant.e.s, complété par 61 réponses à un questionnaire en ligne comprenant une suggestion ouverte, deux questions sur le profil des répondants (niveau et usage professionnel du breton) et deux autres questions listant des

outils développables en demandant lesquels étaient les plus urgents:

- système de lecture augmentée sur écran/tablette avec des livres enrichis de bulles d'information intégrées
- correcteur orthographique / grammatical
- système d'aide à la recherche/exploration d'information
- plateforme de discussion (exemple : échange de recettes ou autre sujet)
- analyseur (aux niveaux morphologique, syntaxique)
- système de détection d'ambiguïtés pour le breton
- dictionnaire des synonymes et expressions/proverbes
- lien entre un dictionnaire et un réseau sémantique

Mekacher (2018) analyse les résultats des questionnaires: il y a unanimité sur le manque de ressources sonores pour l'apprentissage d'une accentuation correcte et une souplesse cross-dialectale. Les locuteurs souhaitent un correcteur orthographique et grammatical intégré aux outils bureautiques, et sont enthousiastes à l'idée d'un système de lecture augmentée. Les résultats doivent être pondérés car il y a peu de répondants, et la liste proposée dans le questionnaire peine à prendre en compte le manque de familiarité des brittophones, enseignant.e.s ou non, avec des outils que justement, ils utilisent peu, d'autant que certains de ces outils sont des outils de développement d'outils numériques. Foret (2016) a exploré un système d'enrichissement de textes qui fournit des synonymes à partir de *Wordnet* et de la base Apertium. Erwan Hupel de l'Université Rennes II a déposé en 2020 une demande de financement pour une thèse sur ce sujet de l'enrichissement de textes par synonymes.

Dans le domaine de la parole orale, la synthèse de la voix de l'IRISA a été livrée à l'OPLB en septembre 2021. Sa diffusion reste un chantier ouvert. Entre autres, un besoin identifié de longue date est celle d'un système GPS capable de prononcer les noms de lieux en Bretagne. En son absence, ce sont les brittophones natifs qui pour utiliser un GPS apprennent dans les faits à interpréter des

formes produites par des synthèses de la voix opérant sur d'autres langues. La communication à distance entre jeunes brittophones privilégie les sms, or la dictée des sms en français est possible, efficace et rapide alors qu'envoyer un sms en breton demande de taper le message, voire de stopper l'autocorrection à chaque mot, interprété comme du français. Développer cet outil demanderait de progresser sur la reconnaissance vocale, sachant que c'est un défi conséquent: si la synthèse de la voix a pu se concentrer sur le breton standard, la reconnaissance vocale nécessite de pouvoir traiter une source multidialectale. En ce qui concerne les conditions de réalisabilité de ce gros chantier de la reconnaissance vocale, et étant donné les techniques actuelles, Damien Lolive (c.p. 10.2021) estime que la reconnaissance de la voix nécessiterait un corpus d'un millier de locuteurs différents ne parlant pas plus de trois minutes, si l'audio est transcrit et que le son est propre (pas de chevauchements, environnement calme). Cela représenterait 50h en tout. Kerbrat (2021a,b) estime, lui, que le corpus devrait atteindre les 200 heures. Kerbrat (2021a,b) mentionne par ailleurs des essais effectués par Francis Tyers avec les données de *Common Voice*. La prosodie de phrase est mal prise en charge dans la synthèse de la voix bretonne actuelle, mais l'un des coordinateurs de la création de la synthèse de la voix travaille de longue date sur la synthèse de la prosodie (cf. Lolive 2008, 2017). Il reste par ailleurs à faire l'étude formelle de la prosodie des phrases en breton, pour systématiser le lien avec la structure syntaxique et avec la structure informationnelle des phrases (signal de focalisation de l'information nouvelle, de signal de l'information donnée, du topique de phrase, etc.).

Les projets qui comportent la création d'une plate-forme pérenne hébergeant les différents corpus sont récurrents, mais peinent à trouver un financement. Le projet *Tal-Breizh (chaînes de traitement et ressources linguistiques pour le breton)* porté en 2015-2017 par Annie Foret (Rennes 1, IRISA) et Ronan Le Coadic (Rennes 2, CRBC) n'a pas été retenu par la Maison de Science de l'Homme de Bretagne. Foret & al. (2015) ont présenté le projet d'une plate-forme ouverte abritant les ressources disponibles pour le breton. Mélanie Jouitteau et Reun Bideault ont présenté en 2018 à la DGLFLF (Délégation Générale à la Langue Française et aux

Langues de France) un projet de plate-forme numérique pouvant articuler les données enrichies de la wikigrammaire ARBRES avec des données de dépôt libre, dont chaque collecteur pourrait rester indépendamment propriétaire, afin de pouvoir proposer un hébergement pérenne, dans une banque cross-interrogeable et sous forme réutilisable. L'idée était de fonder une interopérabilité entre ARBRES et les différents sites de collecte individuels et collectifs, et d'offrir un hébergement pérenne pouvant accueillir et inciter de futurs projets émergents. Tyers & Howell (2021) mentionnent aussi en prospective la mise à disposition de la banque d'arbres UD dans une interface de corpus cherchable destinée aux linguistes.

Enfin, en ce qui concerne les sites webs de contenu en langue bretonne de manière générale, l'adaptation 'responsive web design' est récemment devenue indispensable à leur lecture sur écran réduit. Les terminaux de consultation d'internet sont de taille de plus en plus petite, ce qui a obligé les services web à s'adapter rapidement. Le smartphone est maintenant le premier terminal web utilisé, avec une démocratisation rapide. Il touche presque toutes les couches sociales et tous les âges en sont largement équipés. Les applications dédiées pour ces terminaux sont normalisées pour offrir une lecture facile et ciblée. Le passage au responsif reste à faire pour la plupart des contenus web en breton. Ces travaux sont prévus sur la grammaire ARBRES en 2022. Enfin, une traduction automatique d'applications déjà adaptées serait envisageable si un balisage adapté est mis en place.

4.2. Sociolinguistique et observation des usages

La recherche sociolinguistique est attentive aux usages émergents, à la façon dont les locuteurs des langues minorisées s'emparent des outils numériques, et dont cela peut transformer l'acte de parole dans ces langues. Ce champ universitaire, dont nous ne pouvons rendre compte ici, est d'une vitalité revigorante (Baxter 2009, Moal 2017:76 et les références incluses, Blanchard 2014, 2015, Hicks 2017, Davies-Deacon 2020, Dauneau 2019, rapports réguliers approfondis de l'Observatoire de l'OPLB, etc.). Ce champ sociolinguistique est en dialogue avec les rapports commandés par des suprastructures (DGLFLF, appareils d'État,

Europe), à qui il fournit des retours d'analyse proches du terrain. Nous recommandons cependant que ces structures relativisent la portée des études sociolinguistiques en termes de prospective car par définition même, les études de sociologie sont intéressées uniquement par l'impact sur la société des outils numériques qui sont déjà finalisés et largement distribuées. Par nature, les études sociolinguistiques sont de prospective limitée puisqu'elles étudient la façon dont les utilisateurs s'accommodent ou se saisissent de l'existant. Elles sont donc justifiées à ignorer parfaitement les savoirs de la recherche universitaire fondamentale, les potentiels et les acteurs de développement, les structures de formation essentielles. La mésinterprétation des états de lieux sociolinguistiques comme synthèse des réalisations du TAL en sa globalité et comme base d'analyse servant à son développement pose un problème réel, non pas pour la sociolinguistique qui ne fait que tenir adéquatement son rôle, mais pour le développement TAL puisque cela impacte la visibilité de ses réalisations et potentiels, et donc ses ressources pour les appels à projets.

Le champ sociolinguistique universitaire pourrait par ailleurs se saisir des données de type nouveau que les nouveaux outils numériques fournissent. Si les questionnaires en ligne sont apparus dans les pratiques, les études à ce jour ignorent complètement l'existence des outils d'analyse de fréquentation et d'usage des outils en ligne. Jouitteau (2009-) a un outil *google analytics* associé qui lui fournit une vision assez détaillée des usages de son lectorat (volume d'utilisateurs, pages d'entrée, durée de consultation, flux d'utilisateurs de page à page, trouvabilité par les moteurs de recherche, présence de site le mettant en lien et générant du trafic, etc.). L'étude en est assez ludique. On peut deviner quelle année la wikigrammaire est utilisée par les cours de breton à Moscou, ou quand sont les examens universitaires de linguistique au Québec car les fiches de linguistique formelle rédigées en français montrent alors un pic de connections. Sur les quatre dernières années, l'ouvrage a été ouvert par 130 000 utilisateurs qui ont visionné 165 468 pages. Parmi les utilisateurs, 285 sont revenus plus de 5 fois et 579 plus de trois fois. La durée moyenne des sessions dépasse légèrement 2 minutes. Ci-dessous, vous pouvez voir la synthèse *google analytics* sur cinq ans (2017-

2021) de la provenance de source de connection des utilisateurs par pays et par régions de l'État français.

Pays	Acquisition			Comportement		
	Utilisateurs	Nouveaux utilisateurs	Sessions	Taux de rebond	Pages/session	Durée moyenne des sessions
	129 924 % du total: 100,00 % (129 924)	133 045 % du total: 100,09 % (132 929)	160 064 % du total: 100,00 % (160 064)	79,03 % Valeur moy. pour la vue: 79,03 % (0,00 %)	1,66 Valeur moy. pour la vue: 1,66 (0,00 %)	00:02:01 Valeur moy. pour la vue: 00:02:01 (0,00 %)
1. France	68 860 (51,99 %)	69 565 (52,29 %)	87 614 (54,74 %)	76,04 %	1,86	00:02:27
2. Morocco	9 395 (7,09 %)	9 365 (7,04 %)	10 440 (6,52 %)	85,00 %	1,31	00:01:19
3. Canada	8 828 (6,67 %)	8 855 (6,66 %)	9 991 (6,24 %)	83,77 %	1,33	00:01:15
4. Algeria	6 415 (4,84 %)	6 455 (4,85 %)	7 305 (4,56 %)	86,32 %	1,23	00:01:30
5. Belgium	5 005 (3,78 %)	4 998 (3,76 %)	5 567 (3,48 %)	87,64 %	1,31	00:01:00
6. Switzerland	2 552 (1,93 %)	2 550 (1,92 %)	2 881 (1,80 %)	84,59 %	1,40	00:01:24
7. Tunisia	2 312 (1,75 %)	2 321 (1,74 %)	2 568 (1,60 %)	85,36 %	1,25	00:01:03
8. United States	2 055 (1,55 %)	2 042 (1,53 %)	2 298 (1,44 %)	80,50 %	1,98	00:02:03
9. Cameroon	1 622 (1,22 %)	1 631 (1,23 %)	1 971 (1,23 %)	77,22 %	1,47	00:02:24
10. Spain	1 359 (1,03 %)	1 350 (1,01 %)	1 613 (1,01 %)	77,93 %	1,76	00:01:43

Figure 1 : provenance de source de connection des utilisateurs par pays (2017-2021)

Région	Acquisition			Comportement		
	Utilisateurs	Nouveaux utilisateurs	Sessions	Taux de rebond	Pages/session	Durée moyenne des sessions
	68 860 % du total: 53,00 % (129 924)	69 565 % du total: 52,33 % (132 929)	87 614 % du total: 54,74 % (160 064)	76,04 % Valeur moy. pour la vue: 79,03 % (-3,78 %)	1,86 Valeur moy. pour la vue: 1,66 (12,20 %)	00:02:27 Valeur moy. pour la vue: 00:02:01 (21,30 %)
1. Ile-de-France	**23 696** (33,59 %)	23 396 (33,63 %)	28 169 (32,15 %)	79,38 %	1,66	00:02:09
2. Brittany	**13 166** (18,66 %)	13 019 (18,71 %)	20 385 (23,27 %)	63,38 %	2,65	00:04:04
3. Auvergne-Rhone-Alpes	**5 861** (8,31 %)	5 789 (8,32 %)	6 771 (7,73 %)	82,41 %	1,51	00:01:43
4. Occitanie	**4 101** (5,81 %)	4 016 (5,77 %)	4 500 (5,14 %)	83,56 %	1,41	00:01:26
5. Pays de la Loire	**4 002** (5,67 %)	3 931 (5,65 %)	5 062 (5,78 %)	70,80 %	2,14	00:03:08
6. Nouvelle-Aquitaine	**3 871** (5,49 %)	3 790 (5,45 %)	4 446 (5,07 %)	81,71 %	1,74	00:02:01
7. Hauts-de-France	**3 500** (4,96 %)	3 435 (4,94 %)	3 968 (4,53 %)	82,11 %	1,44	00:01:20
8. Grand Est	**3 441** (4,88 %)	3 414 (4,91 %)	3 784 (4,32 %)	84,22 %	1,41	00:01:25
9. Provence-Alpes-Cote d'Azur	**3 340** (4,73 %)	3 287 (4,73 %)	3 700 (4,22 %)	83,30 %	1,40	00:01:22
10. Normandy	**1 936** (2,74 %)	1 911 (2,75 %)	2 145 (2,45 %)	81,68 %	1,44	00:01:19

Figure 2 : provenance de source de connection des utilisateurs par régions (2017-2021)

On voit se dessiner un lectorat dans les zones traditionnelles de pratique de la langue et dans les lieux d'immigration des brittophones, ainsi qu'un lectorat plus largement dans les pays les plus riches de la francophonie. Il existe en effet un lectorat distinct du lectorat brittophone ou d'apprenants, dont l'intérêt premier est la linguistique formelle plutôt que la description du breton. En novembre 2021, les requêtes web qui ont le plus amené sur ARBRES sont les mots clef: structure syntaxique (59), morphème libre et lié (20), construction syntaxique (19), verbe factif (17), verbes factifs 16), complémenteur (14), grammaire bretonne (13), morphème zéro (12), déictique[sic] spatiaux (11) et verbe ditransitif (10). Ce lectorat apprend les notions de base de linguistique formelle en français à travers des exemples du breton.

La synthèse *google analytics* ci-dessous montre la prédominance du moteur de recherche *google* dans les sources de connection sur un flux d'utilisateurs de page à page, ainsi que la présence timide d'un lectorat fidélisé, à connections directes. Le groupe le plus important est ensuite celui des sources de connection non-retrouvables.

Figure 3 : flux des utilisateurs connectés à partir de l'État français, sur 2021

M. Jouitteau fournit annuellement à la structure d'évaluation du CNRS une synthèse détaillée du développement de ARBRES à partir de ces données d'utilisation, enrichies et éclairées par des séances régulières de « surf accompagné » où des utilisateurs de ARBRES montrent leurs usages, mésusages ou incompréhensions. Les outils d'analyse automatisés des usages ont révolutionné le développement des outils en ligne, et il serait étonnant que les autres développeurs, ou même les rédacteurs de blogs culturels en breton, n'aient pas des outils similaires qui renseigneraient la sociolinguistique des usages des brittophones 2.0.

4.3. Pragmatique d'un projet de science ouverte

Un projet comme la wikigrammaire nécessite un investissement sur le long terme rendu possible par l'existence de contrats de

recherche sur des périodes longues comme le statut de fonctionnariat au CNRS. Dans ce cadre qui le rend possible, les conséquences pragmatiques d'un tel choix de recherche sont aussi évaluables.

ARBRES est encore souvent amalgamé avec *wikipédia* et appréhendé par ses utilisateurs comme une ressource sans auteur réel, et les bibliographies scientifiques sont encore centrées sur le circuit papier, même alors qu'elles sont directement copiées à partir de formats numériques. Les travaux de science ouverte comme les travaux numériques en général sont encore généralement peu cités, et dans des formats des plus créatifs, irrécupérables automatiquement. Ces facteurs rendraient dangereuse une évaluation exclusive de la recherche par les mentions correctes en bibliographies universitaires.

L'attitude des entités évaluatrices s'est cependant nettement améliorée depuis les débuts de la wikigrammaire en 2009. La Déclaration de San Francisco sur l'évaluation de la recherche (Dora) en 2012 et le Manifeste de Leiden en 2015 ont livré une analyse critique des pratiques d'évaluation de la science restreintes à la citométries de revues et d'articles. Ils ont avancé des recommandations en matière d'utilisation d'indicateurs scientométriques pour l'évaluation des scientifiques (Pourret 2021). Le CNRS s'est engagé fermement ces dernières années en soutien de la science ouverte, et les critères bougent. Il y a une dizaine d'années, dans les évaluations annuelles du CNRS, ARBRES a reçu occasionnellement, au milieu de soutiens déclarés de collègues qui en cernaient le potentiel, des évaluations bonhommes se souciant de ce que l'autrice ne « perde pas trop de temps » sur son site, appréhendé comme un blog de loisirs. Certains avis évaluaient son volume de publications en laissant ouvertement de côté le développement de la base de données. Ce type de retour a disparu et il est chaque année plus facile d'insérer une wikigrammaire dans les critères d'évaluation proposés.

De façon assez ironique mais sans doute transitoire, ce sont les initiatives de diffusion numérique des produits scientifiques qui s'appuient le plus sur la science ouverte qui manquent à valoriser les réalisations numériques qui répondent le mieux à ses critères. En France, les plates-formes de visibilisation comme HAL

valorisent les articles papier rendus disponibles à l'intérieur de la plate-forme, et non pas les banques de données, articles en ligne ou sites de recherche disponibles en ligne sans le recours de HAL. L'interface *google scholar* les rend entièrement invisibles, comme c'est d'ailleurs le cas pour tout système numérique qui requiert un ISSN ou ISBN pour reconnaître un travail.

L'aide au transfert des savoirs vers la société est officiellement souhaitée, mais les conditions de réalisabilité peinent encore à être assurées largement. De nombreux laboratoires de linguistique n'ont pas d'ingénieurs de recherche en informatique, et le développement numérique dépend *in fine* des capacités informatiques en propre de chercheurs en sciences humaines. En pratique, un chercheur doit trouver régulièrement et au coup par coup des ressources pour la gestion des mises à niveau du logiciel. L'accessibilité d'un ouvrage numérique dépend principalement des algorithmes de *google*, avec les risques que cela comporte. Dans le tableau ci-dessous construit par *google analytics*, on voit nettement les effets du changement d'algorithme du moteur de recherche de *google* en janvier 2020, qui a déclassé volontairement les productions numériques non-adaptées à la consultation sur smartphones et tablettes. L'adaptation à la visualisation sur smartphones et tablettes est appelée « passage au responsif ». Il n'est pas effectué sur ARBRES.

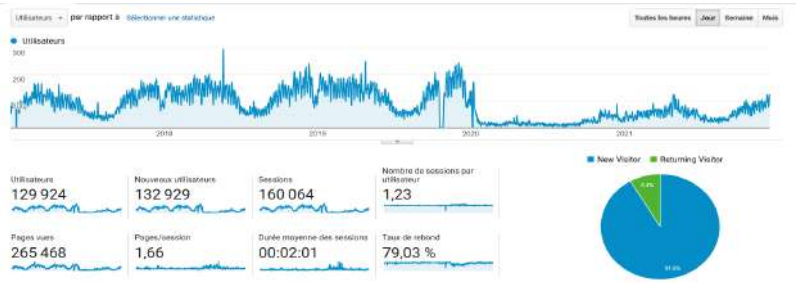


Figure 4 : synthèse google analytics sur 5 ans du nombre d'utilisateurs de la wikigrammaire ARBRES

La moyenne de 130 utilisateurs humains par jour sur 2019 est descendue brutalement en 2020 à 25 par jour, c'est-à-dire une

division par plus de cinq. Le site a été globalement laissé aux utilisateurs déjà fidélisés : entre 2019 et 2020, la moyenne de temps d'une session est passée de 1:50 à 3:15 minutes. Sur la courbe des fréquentations, on repère la chute nette de janvier 2020, et une période de mise hors ligne en 2019, qui est significative. Le laboratoire IKER (CNRS) a financé sur fonds propres la mise à niveau du logiciel wiki et le passage au responsif, confié à une entreprise privée en Bretagne. L'administration a déclenché le paiement avant clôture des travaux, et le passage au responsif n'a finalement jamais été effectué par l'entreprise, qui a même laissé un temps le site hors ligne (la chute exceptionnelle à zéro utilisateurs fin 2019). En bibliothèque papier, dans de telles situations, l'ouvrage serait considéré TDE (Tombé Derrière les Étagères), et des professionnels bibliothécaires et documentalistes s'attaqueraient au problème. Pour le circuit numérique, nous ne repérons pas l'équivalent de ces professionnels chargés de s'assurer de l'accessibilité des ouvrages pour les publics concernés. Ce problème est considérable et touche tous les grands sites numériques scientifiques financés par projet qui sont déposés à clôture dans les grandes infrastructures de type HumaNum. A notre connaissance, après financement de création, les usages et accessibilités web ne sont plus évalués et un changement d'algorithme google peut drastiquement diviser leur lectorat sans que ce soit même repéré. A noter que dans le cas de ARBRES et des autres projets de science ouverte en développement constant, cette question de l'accessibilité ne pourrait de toute façon pas être déléguée car leur dépôt dans les grandes infrastructures n'est pour l'instant pas une option.

Nous pensons avoir concouru ici à documenter et appuyer la conclusion que pour évaluer chercheuses et chercheurs, les critères quantitatifs de publications et de citation, et les critères qualitatifs de gradation des revues et des langues d'expression de la recherche (anglais vs. langues des communautés de locuteurs) devraient être enrichis par une appréhension de la réalité des transferts de savoirs faits vers la société, en facilitant la citabilité des ouvrages numériques libres et en visibilisant les pratiques de science ouverte. Soutenir pragmatiquement le transfert des savoirs vers la société permettrait en sus à plus de chercheurs de se lancer avec

confiance dans des projets d'envergure en science ouverte, à même d'alimenter les traitements automatiques des langues.

BOITE NOIRE

Dans le cadre des recherches pour cet article, Mélanie Jouitteau a contacté et interviewé plusieurs collègues. Annie Foret a été interviewée via le logiciel libre *Jitsi* en décembre 2021. Le résumé de la rencontre lui a été communiqué en ligne. Damien Lolive a été contacté le 15 octobre, puis recontacté le 4 novembre, avec deux collègues hommes en copie du message, dont Reun Bideault, développeur web qui opère régulièrement les mises à jour de la wikigrammaire ARBRES, et qui était alors consultant pour l'article. Ce dernier a accepté d'organiser l'interview de Monsieur Lolive le 24 novembre dans des locaux prêtés par le centre de formation *Kelenn*, et de rejoindre l'article en signature. Le résumé du développement de la synthèse de la voix a été communiqué ensuite à Damien Lolive, et corrigée par ce dernier et Gaëlle Vidal après vérification de quelques points auprès des locuteurs dont la voix a servi à la constitution du corpus oral. Dewi Kerbrat, auteur du rapport 2021 de l'OPLB, a été contacté le 14 octobre, puis le 4 novembre avec deux collègues hommes et les deux co-directeurs de l'OPLB en copie, puis le 6 novembre via Facebook. Il a confirmé par email le 8 novembre que son rapport sur les outils numériques en breton n'incluait pas la wikigrammaire ARBRES, et laissé ouverte la possibilité qu'elle le soit. Recontacté le 9 novembre avec une explication du potentiel pour le TAL de la base de données de ARBRES, et averti que le rapport négligeait des ressources universitaires, il n'a pas donné suite.

BIBLIOGRAPHIE

AUBRY, Yves (2000). *Synthèse vocale en breton*, ms. De mémoire de maîtrise, IUP MIME Le Mans, TES/ENSSAT.

AUBRY, Yves (2004). *Logiciel du traitement de la parole et d'aide à l'enseignement et à l'apprentissage de la prosodie: application au breton*, travaux de D.R.T. d'ingénierie, Université du Maine.

BAXTER, R.N. (2009). 'New technologies and terminological pressure in lesser-used languages. The Breton Wikipedia, from terminology consumer to potential terminology provider', *Language Problems and Language Planning* 33:1, John Benjamins: Amsterdam/Philadelphia, 60-80.

BLANCHARD, Jean-François (2014). « Pratiques langagières et processus dialogique d'identification pour une langue minorée. Le web en langue bretonne », Gaël Hily (dir.), *Expression de l'identité dans le monde celtique*, Rennes : TIR. 9-34.

BLANCHARD, Jean-François (2015). *Pratiques langagières et processus dialogiques d'identification sur les réseaux socionumériques. Le cas de la langue bretonne*, ms. thèse. Université Rennes 2. texte.

CHEVEAU, Loïc & Pierre-Yves KERSULEC (2012-évolutif). *Dictionnaires bretons parlants.*

CORNILLET, Gérard (2017). *Geriadur Brezhoneg-Galleg,* (version corrigée en 2020, texte).

AN DROUIZIG (2021). *Site de ressources numériques et association de création de ressources,* [accédé le 20.11.2021].

DAUNEAU, Goulven (2019). *Brezhoneg, Niverel, Deskadurezh : hiziv ha warc'hoazh*, ms. de mémoire de master, U. Rennes II. texte.

DAVIES-DEACON, Merryn (2020). *New speaker language and identity: Practices and perceptions around Breton as a regional language of France,* ms. de thèse.

DELOOF, Jan (2008-2010). *Bretons-Nederlands Woordenboek*, interface web par Kevin Donnelly.

DESSEIGNE, Adrien, Loïc CHEVEAU & Pierre-Yves KERSULEC (2013-2018). *Banque Sonore des Dialectes Bretons, projet de documentation multimédia en ligne*, site.

DONNELLY, Kevin (2010). 'Jan Deloof Breton-Dutch Dictionary', blog *Me, Myself, Why? Free software and languages, not necessarily in that order*, texte, [consulté le 13.12.2021].

LE DU, Jean (2001). *Nouvel Atlas Linguistique de Basse-Bretagne*, vol. I et II, Centre de Recherche Bretonne et Celtique, Université de Bretagne Occidentale, Brest.

FAVEREAU, Francis (1993). *Dictionnaire du breton contemporain / Geriadur ar brezhoneg a-vremañ*. Morlaix: Skol Vreizh, moulte rééditions, version en ligne.

FAVEREAU, Francis (2015). *Geriadurig ar brezhoneg a-vremañ / Dictionnaire compact du breton contemporain*, Morlaix: Skol Vreizh.

FAVEREAU, Francis. 2016-évolutif. *Grand dictionnaire bilingue breton-français, français-breton*, texte.

FAVEREAU, IRISA & TES (1999). *Ar geriadur a gomz brezhoneg a-vremañ*, Morlaix : Skol Vreizh. CD-ROM.

FORET, Annie, Valérie BELLYNCK & Christian BOITET (2015). 'Akenou-Breizh, un projet de plate-forme valorisant des ressources et outils informatiques et linguistiques pour le breton', présentation à la conférence *TALARE (Traitement Automatique des Langues Régionales de France et d'Europe)*, texte.

FORET, Annie (2016). « Enrichissement de données en breton avec Wordnet », Poibeau, Thierry, Teresa Lynn, Delyth Prys & John Judge

(éds.), *Proceedings of the Second Celtic Language Technology Workshop* (CLTW 2016), 55-61. texte.

FORET, Annie (2018). « Breton-français et numérique, projet LangNum-br-fr (phase conception) ». *Conférence Langues et numérique 2018*, Juillet 2018, Paris, France. texte ou texte.

FORET, Annie (2018b). 'Logiciels et ressources pour le breton', document du projet LangNum-br-fr, ms. 12p.

GUILLOU, A. (2000). *Correcteur de prosodie pour la langue bretonne*, ms. de rapport de projet.

HERRIEU, Loeiz (1994). *Kammdro an ankoù,* Gourhelen, Ronan Huon embanner: Al Liamm.

HICKS, Davyth (2017). « Breton – a digital language ? », *The Digital Language Diversity Project*, Erasmus +. texte.

IRISA (2001). *Rapport d'activité 2001. Projet CORDIAL. Communication multimodale personne-machine à composantes orales : méthodes et modèles*, texte.

AN INTANV, Pascal (1994). *War hent fonetikadur ar Brezhoneg / Sur les chemins de la phonétisation du breton*, ms. de mémoire de maîtrise, Université de Rennes II.

JOUITTEAU, Mélanie (2013b). « La linguistique comme science ouverte; Une expérience de recherche citoyenne à carnets ouverts sur la grammaire du breton », *Lapurdum* XVI, Charles Videgain (dir.), 93-115, texte.

JOUITTEAU, Mélanie (éd.) (2009-2021). « Traitement automatique du langage - Breton », *ARBRES, wikigrammaire des dialectes du breton et centre de ressources pour son étude linguistique formelle*, IKER, CNRS, URL. - génèse du présent article et mises à jour.

KREIZENN AR GERIAOUIÑ (2016). *Geriaoueg yezhadur*, Brezhoneg 21 (éd.), texte.

KERBRAT, Dewi (2021a). *Ar brezhoneg en oadvezh an niverel, diagnostik ha strategiezh diorren*, ms. de rapport pour l'OPLB.

KERBRAT, Dewi (2021b). *La langue bretonne à l'ère du numérique, diagnostic et stratégie de développement*, ms. de rapport pour l'OPLB.

LEIXA, Jérémy, Valérie MAPELLI & Khalid CHOUKRI (2014). *Inventaire des ressources linguistiques de langues de France*, Organisme ELDA, ms. de rapport pour la DGLFLF.

LOLIVE, Damien (2008). *Transformation de l'intonation : application à la synthèse de la parole et à la transformation de voix. Intelligence artificielle [cs.AI]*, ms. de thèse de l'Université Rennes I. texte.

LOLIVE, Damien (2017). *Vers plus de contrôle pour la synthèse de parole expressive. Intelligence artificielle [cs.AI]*, ms. de HDR, Université de Rennes I.

MARTINET, Pierre (2021). *Contributions à l'enrichissement automatisé de langues peu dotées. Cas du breton et des grammaires formelles*, ms. de rapport de stage (6 mois), laboratoire SemLIS (IRISA), Rennes I. texte.

MENARD, Martial et Hervé LE BIHAN (2016-évolutif). *Devri: Le dictionnaire diachronique du breton*, Université Rennes II & Kuzul ar Brezhoneg, en ligne.

MENARD, Martial et Iwan KADORED (dir.) (2001). *Geriadur Brezhoneg*, Embannadurioù An Here.

MEKACHER, Echraf (2018). *Projet LangNum-br-fr*, ms. du laboratoire LOUSTIC, U. Rennes I. texte.

MINOCHA, Akshay et Francis TYERS (2014). « Subsegmental language detection in Celtic language text », *Proceedings of the First Celtic Language Technology Workshop* CLTW1, 76-80, texte.

MOAL, Stefan (2017). *Médiation, transmission, création. La revernacularisation du breton au 21e siècle*, ms. de HDR.

MOCQUARD, Guillaume (1999). *Correcteur de prosodie*, ms. de rapport de stage IFSIC, TES/IRISA, ENSSAT.

MOCQUARD, Guillaume (2001). *Korpus prosodiezh*, ms. de mémoire de maîtrise, Université de Rennes II.

Ar MOGN, Olier (2015). « Langue bretonne et nouvelles technologies : une vitalité à soutenir », présentation à *Technologies pour les Langues Régionales de France*, Meudon. vidéo.

MORVAN, Pierre (2019). *Ha difaziañ a ra LanguageTool ar c'hemmadurioù? Peseurt hentenn sevel evit gellet gouzout peseurt barregezh a zo gant an difazier LanguageTool war ar c'hemmadurioù?*, ms. de mémoire de maîtrise, Université Rennes II.

OPLB (2021a). *Termofis*, dictionnaire terminologique, en ligne.

OPLB. (2021b). *Kerofis*, base de données toponymique, en ligne.

OPLB. (2021c). *Corpus de phrases en breton, ou en français*, accessible.

Petit, M (2003). *Correcteur orthographique de langue bretonne*, ms. rapport de projet, ENSSAT, 1-37.

POIBEAU, Thierry (2014). 'Processing Mutations in Breton with Finite-State Transducers', *Proceedings of the First Celtic Language Technology Workshop*, Dublin, Ireland. texte.

TYERS, Francis Morton (2008). 'Extracting bilingual word pairs from wikipedia', *Proceedings of the SALTMIL Workshop at the Language Resources and Evaluation Conference*, LREC2008, 19–22.

TYERS, Francis Morton (2009). 'Rule-based augmentation of training data for breton–french statistical machine translation', *Proceedings of the 13th Conference of the European Association for Machine Translation*, 213–218. texte.

TYERS, Francis Morton (2007-2009). *Breton morphological analysis,* http://xixona.dlsi.ua.es/~fran/breton/index.php, GNU-GPL.

TYERS, Francis Morton (2010a). 'Rule-based Breton to French machine translation', *Proceedings of the 14th Annual Conference of the European Association of Machine Translation*, 174-181. texte et poster.

TYERS, Francis Morton (2010b). « An treiñ emgefreek diazezet war reolennoù evit treiñ ar brezhoneg e galleg », *Hor Yezh* 262, 27–39. [traduction par Thierry Fohanno]

TYERS, Francis Morton (2015). *Rule-based augmentation of training data in breton–french statistical machine translation*, rapport.

TYERS, Francis Morton et Vinit RAVISHANKAR (2018). « A prototype dependency treebank for Breton », *Actes de la conférence Traitement Automatique de la Langue Naturelle*, TALN 2018, 197-204. texte.

TYERS, Francis Morton et Nicholas HOWELL (2021). « Morphological analysis and disambiguation for Breton », *Language Resources and Evaluation*, 431-473. preview.

POURRET, Olivier (2021). « Comment la science ouverte peut faire évoluer les méthodes d'évaluation de la recherche », *The conversation,* [4 novembre 2021, accédé le 06 novembre].

YEKEL, Tangi, Riwal GEORGELIN et Juluan AR C'HOZH (2015-2021). *Brezhoneg Bro-Vear*, Blog de l'association *Hent don.*

Mélanie Jouitteau
IKER, CNRS, UMR 5478,
Université de Pau et des Pays de l'Adour
Université Bordeaux Montaigne
melanie.jouitteau@iker.cnrs.fr

Reun Bideault
développeur Web Indépendant
lepoleethik2@gmail.com

Chapitre 2

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

Christophe Rey

CY Cergy Paris Université, Institut Universitaire de France

Abstract

For more than ten years now, several research projects have focused on Picard, a regional language of France. These projects have allowed the study of this language to benefit from relevant resources and tools in the field of Automatic Language Processing. This work reviews these major projects and investigates their potential role in a possible revitalization of the language.

1. Introduction

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)[1] – réalisée au sein de l'Université de Picardie Jules Verne –, la conduite du projet RESsources informatisées et Traitement AUtomatique pour

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

les langues REgionales (RESTAURE (2015-2019))[2], et dans une moindre mesure l'élaboration d'un *Atlas pan-picard informatisé* (2018-2020)[3] – à l'Université de Lille –, et la conduite du projet METalexicographie de la Langue PICcarde (METALPIC (2017-2022) ont ainsi fait entrer la langue picarde dans le concert des langues régionales de France significativement dotées en ressources numériques.

Dans le cadre de notre contribution, 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 à valoriser, nous montrerons que l'un des apports du Traitement Automatique des Langues (désormais TAL) 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 (Auroux, 1994) salutaire. Nous ferons ainsi le point sur les deux avancées majeures très récentes que constituent la création de la "Commission de néologie et de terminologie pour la langue picarde"[4] et la publication du *Dictionnaire fondamental français-picard* (2020)[5], tout en insistant sur les carences structurelles qui demeurent tout de même en dépit de ces avancées séduisantes.

[2] Projet financé par l'Agence Nationale de la Recherche. https://restaure.unistra.fr/
[3] Projet financé par l'Agence Nationale de la Recherche. https://anr-appi.univ-lille.fr/
[4] https://languepicarde.fr/commission-de-neologie-et-de-terminologie/
[5] https://languepicarde.fr/dictionnaire-fondamental-francais-picard/

2. Le picard : une “petite” langue collatérale[6] au français

Avant de nous lancer dans la présentation des différents projets mentionnés en introduction, consacrons quelques lignes à la présentation linguistique et sociolinguistique du picard.

Le picard est une langue régionale de France dotée d’environ 700 000 locuteurs, dont l’aire linguistique comprend l’actuelle région des Hauts-de-France (cf. réforme territoriale de 2014) – à savoir l’ancienne région du Nord-Pas-de-Calais, une partie de la région Picardie, à l’exception du sud-est du territoire administratif et d’une partie de moins en moins importante du département de l’Oise – et la province belge du Hainaut (cf. carte ci-dessous). L’aire linguistique dépasse donc les frontières administratives.

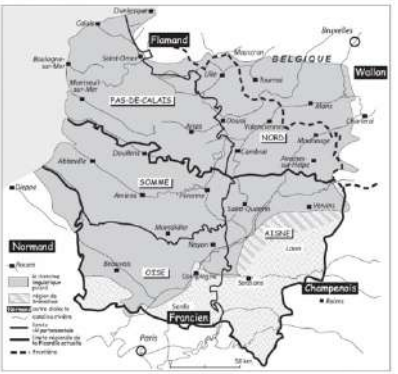


Figure 1. Carte de l’aire linguistique picarde établie par René Debrie selon les données de Raymond Dubois, réalisée par Joëlle Désiré (Amiens, Université Jules Verne, 1985).

La langue picarde dispose d’une longue tradition dialectologique, qui a permis d’établir en 1957 un tracé de l’aire linguistique[7], en lien

[6] Cf. ELOY, 2004.

[7] Des travaux récents (Forlot et Martin, 2014 ; Martin, 2015 et Martin et Forlot 2016) tendent à montrer que l’aire linguistique définie en 1957 par Raymond Dubois semble se réduire principalement au sud-est et au sud-ouest du domaine linguistique décrit par cette carte.

avec les enquêtes préalablement menées et la publication d'un *Atlas linguistique et ethnographique picard*[8]. Au cœur même de cette aire linguistique picarde, nous observons distinctement des variétés de picard, qui correspondent à des pôles de pratique (Forlot & Martin, 2014; Martin, 2015). On dénombre[9] notamment un pôle vimeusien (ouest d'Abbeville), un pôle amiénois (autour d'Amiens), un pôle valenciennois (autour de Valenciennes), un pôle lillois (autour de Lille), un pôle artésien (région d'Arras), un pôle beauvaisien (autour de Beauvais), un pôle thiérachien (contreforts occidentaux du massif des Ardennes) ou encore un pôle tournaisien (dans la région de Tournai dans le Hainaut belge).

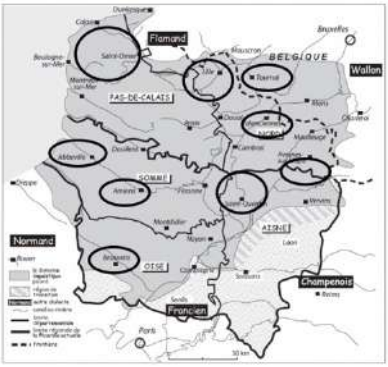


Figure 2. Carte de l'aire linguistique picarde avec localisation des pôles de pratique de la langue, à partir de la carte de l'aire linguistique picarde.

[8] Cette tradition dialectologique fait état des enquêtes dialectologiques réalisées par Fernand Carton et Maurice Lebègue (entre autres) à partir des années 1950 sur le domaine linguistique picard (127 points d'enquête dans 5 départements) qui ont contribué à la réalisation de l'*Atlas linguistique picard* (1989-1998) (soit 660 cartes) et ont permis de définir le tracé de référence de l'aire linguistique publié par Raymond Dubois en 1957, d'après une étude toponymique (communes, hameaux et lieux-dits).

Dubois Raymond, 1957, *Le Domaine picard. Délimitation et carte systématique, dressée pour servir à l'Inventaire général du picard et autres travaux de géographie linguistique*, Arras : Société de dialectologie picarde.

Carton Fernand, Lebègue Maurice, 1989, 1998, 2004, *Atlas Linguistique et Ethnographique picard*, 2 volumes, Paris : Éditions du CNRS.

[9] Précisons que les pôles mentionnés dans cet article s'appuient sur des enquêtes menées récemment et qu'ils ne prétendent pas définir aujourd'hui l'aire linguistique picarde (Martin, Forlot, 2016 ; Martin, 2015). Ils apportent un regard complémentaire sur la vitalité du picard aujourd'hui.

Ces pôles mettent en évidence des variétés du picard qui s'affirment les unes par rapport aux autres et ont du sens pour les locuteurs, mais aussi pour les scripteurs qui cultivent cette singularité dans les confins de l'aire linguistique picarde.

En dépit de son histoire littéraire longue et riche, le picard constitue donc une langue sans standardisation qui s'explique notamment par le fait que cette dernière ne bénéficie d'aucun enseignement généralisé (Forlot & Martin, 2015), d'aucun usage officiel, ni d'une politique linguistique aussi vivace que celle sur laquelle peuvent s'appuyer des langues comme le breton, l'occitan ou encore le corse.

La prédominance de pôles de pratiques linguistiques explique en grande partie le fait que la langue picarde est caractérisée par de nombreux phénomènes de variation, phonétique, morphologique, lexicale, mais aussi et surtout graphique. Sur ce dernier point, la grande variabilité graphique du picard s'explique d'ailleurs par une utilisation très particulière des tirets, des apostrophes et des points :

> « Les cas les plus problématiques sont les suivants :
>
> - le tiret peut être un séparateur (*Est-ce-que*) mais il peut faire partie de certains mots (par exemple dans certains verbes) : *quandis n'mariye-té* [picard de Belgique] / *kan k i s'marite* [picard de l'Amiénois] (*quand elles se marient*) ;
>
> - le point peut être utilisé comme séparateur de phrase, entrer dans la composition d'un sigle ou signaler un allongement de la consonne qui conduit à une nasalisation) *I se proumon.ne* [picard du Cambrésis] (*il se promène*);
>
> - l'apostrophe peut avoir plusieurs interprétations possibles : (a) une marque qui oriente par rapport à la prononciation comme dans les exemples *té mérit'roès* (*tu mérites*), *f'rais* (*ferais*), où une voyelle est supprimée ; (b) une marque de l'élision d'une voyelle *L'aute* (*l'autre*) ; (c) en fin de mot *Dis-l'*.
>
> Il est fréquent d'avoir plusieurs apostrophes dans le même mot avec utilisations différentes *Qu't'os* (*que tu as*), *Coreed'l'histoire* (*encore de l'histoire*);
>
> - l'espace. On peut avoir des séquences espace + lettre + espace. Il s'agit du phénomène d'épenthèse, la lettre marque une liaison entre les deux mots : par exemple la lettre z dans *lé z éfans* (*les enfants*). » (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.)

En pratique, on retiendra qu'il existe quatre grands types de graphies pour la langue picarde, dont trois constituent des substandards –, à savoir l'orthographe proposée par Carton (Carton, 1963), celle de Vasseur (Vasseur, 1968), et enfin celle livrée par Braillon (Braillon, 1991).

3. De nombreux projets valorisant la langue picarde

Après cette présentation succincte de la situation linguistique et sociolinguistique du picard, nous allons à présent lister plusieurs projets de recherche s'étant succédés durant les deux dernières décennies et ayant permis à cette langue d'entrer dans le giron des langues régionales de France outillées par le TAL.

3.1 L'expérience PICARTEXT (2008-2011)

La première expérience que nous souhaitons évoquer ici est celle conduite au sein de l'Université de Picardie Jules Verne de 2008 à 2011 et conçue par le professeur Jean-Michel Eloy : le projet PICARTEXT.

Ayant clairement fait entrer la langue picarde dans l'univers des langues outillées et explorées dans leur structure grâce au TAL, cette base de données a été construite sur le modèle de FRANTEXT, permettant au terme de son développement de pouvoir bénéficier d'environ 3,5 millions de mots informatisés et interrogeables grâce à une interface de consultation performante. Le moteur de recherche mis en place permet en effet, en plus des fonctionnalités classiques de recherche, de pouvoir trouver un mot en fonction de sa possible variation phonétique et graphique sur le domaine ou encore de trouver ses correspondances dans des variétés de picard bien précises. La figure 3 permet de visualiser l'interface de recherche de la base PICARTEXT et de prendre la mesure de la multiplicité des modes de recherche offerts aux utilisateurs.

L'outil constitué s'impose comme une ressource précieuse permettant à tout un chacun d'explorer une multitude de textes de référence de la littérature picarde. Ses développements futurs sont

nombreux et mériteraient d'attirer le regard de financeurs institutionnels et privés.

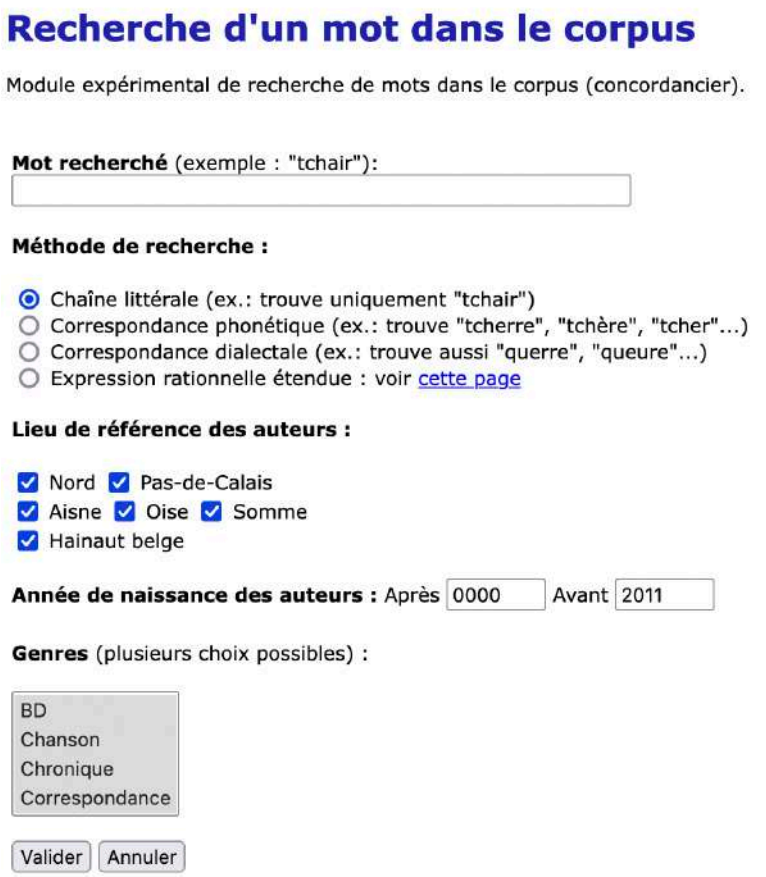


Figure 3. Interface de recherche du projet PICARTEXT

3.2 Le projet RESTAURE (2015-2019)

La seconde expérience que nous évoquerons ici est celle du projet de recherche RESTAURE (RESsources informatisées et Traitement AUtomatique pour les langues Regionales) financé par l'Agence Nationale de la Recherche (ANR) et fédérant différentes équipes

travaillant sur des problématiques liées aux langues régionales de France : les laboratoires LILPa (Linguistique, Langues, Parole, Université de Strasbourg), CLLE-ERSS (Cognition Langues, Langage, Ergonomie – Équipe de Recherche en Syntaxe et Sémantique, Université de Toulouse), HABITER LE MONDE (Université de Picardie Jules Verne, Amiens) et une quatrième équipe d'informaticiens en charge de l'appui au développement de ressources linguistiques informatisées, le LIMSI-CNRS (Laboratoire d'Informatique pour la Mécanique et les Sciences de l'Ingénieur, Université d'Orsay).

Ce projet avait pour ambition de rendre commune une approche visant à éprouver les outils du TAL – et donc à doter en ressources informatisées, mais également en outils de Traitement Automatique – trois langues régionales de France : l'alsacien, l'occitan et le picard.

Ces trois langues ayant déjà fait l'objet de travaux significatifs, leur inscription dans RESTAURE visait à associer et à mutualiser les connaissances et les compétences des divers acteurs du projet dans le but de profiter des expériences, des approches scientifiques et des outils développés antérieurement pour ces différentes langues, ainsi qu'à créer de nouveaux outils.

Les réalisations linguistiques et informatiques de RESTAURE pour la langue picarde sont nombreuses, puisque des avancées considérables ont été réalisées en matière de tokenisation[10], de lemmatisation, d'étiquetage morpho-syntaxique[11], de réflexion sur la constitution de lexiques (de verbes, de noms, de mots grammaticaux, etc.) ou encore de réflexion sur l'élaboration de ressources lexicographiques.

Aujourd'hui achevé, RESTAURE a non seulement permis la réalisation d'avancées significatives pour le traitement automatisé de

[10] Cf. BERNHARD Delphine & TODIRASCU Amalia & MARTIN Fanny & ERHART Pascale & STEIBLE Lucie & HUCK Dominique & REY Christophe (2017) "Problèmes de tokénisation pour deux langues régionales de France, l'alsacien et le picard", *Actes de l'atelier "Diversité Linguistique et TAL" (DiLiTAL 2017)*, article en ligne : https://taln2017.cnrs.fr, pp. 14-23.

[11] BERNHARD Delphine & LIGOZAT Anne-Laure & MARTIN Fanny & BRAS Myriam & MAGISTRY Pierre & al. "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. 2018.

la langue picarde, mais il a aussi et surtout ouvert un champ considérable de possibles, attestant encore davantage l'intérêt de cette langue dont l'extrême variabilité constitue un défi pour le TAL.

Le traitement de données textuelles numériques nous invite désormais à appréhender autrement la langue, tant dans l'acquisition et la normalisation de ressources électroniques écrites, que dans le développement d'outils d'analyse automatique adaptés aux particularités notamment morphosyntaxiques de la langue, que dans la valorisation de ces outils et de ces ressources auprès du grand public.

3.3 L'Atlas pan-picard informatisé (2018-2020)

Le projet APPI (Atlas Pan-Picard Informatisé) est un projet tout récemment achevé et offrant un prolongement aux travaux de constitution d'atlas linguistiques et notamment à l'*Atlas linguistique et ethnographique picard*[12] (ALPic.) et l'*Atlas linguistique de la Wallonie* (ALW).

Croisant avec succès les travaux de dialectologie classiques ayant permis la constitution d'atlas linguistiques pour les langues régionales de France et les travaux relavant du TAL, ce projet, lui aussi financé par l'Agence Nationale de la Recherche (ANR), a été conduit sous la responsabilité scientifique d'Esther Baiwir de l'Université de Lille. Il a mis en lumière l'importance des atlas dans la description linguistique et lexicographique des langues régionales, et donc du picard.

S'appuyant sur les matériaux linguistiques recueillis lors des enquêtes pour l'*Atlas linguistique et ethnographique picard*, l'APPI a notamment permis de mettre à la disposition de tous près de 660 cartes, 652 précisément, en libre consultation et téléchargement[13]. Grâce aux outils du TAL, ce projet propose par ailleurs l'alignement d'une part des matériaux de l'ALPic et de l'ALW avec ceux de l'Atlas Linguistique de France (ALF) sur une base sémantique (environ 200 notions). Les concepts traités sont munis d'une définition qui s'appuie sur le travail lexicographique du *Trésor de la langue*

[12] Cf. CARTON Fernand, LEBÈGUE Maurice, 1989, 1998, 2004, *Atlas Linguistique et Ethnographique picard*, 2 volumes, Paris : Éditions du CNRS.

[13] https://anr-appi.univ-lille.fr/index.php/alpic/?db=xxx

française informatisé (TLFi). Les notices et cartes des trois atlas sont accessibles en mode image.

Par son approche audacieuse, le projet APPI apporte incontestablement sa pierre à l'édifice très complexe qu'est la valorisation de la langue picarde.

3.4 METALPIC

Nourri par plusieurs des projets qui viennent d'être évoqués ci-dessus, le projet METALPIC est un projet de recherche réalisé de 2017 à 2022 et financé grâce au soutien de l'Institut Universitaire de France (IUF). Son approche, mêlant description linguistique et outil du TAL est celle de la métalexicographie, jeune discipline des sciences du langage faisant du dictionnaire un objet d'étude à part entière[14]. En d'autres termes, l'approche de ce projet consistait à fournir une analyse critique du fonctionnement de la lexicographie de langue picarde. Le travail réalisé est ainsi venu combler une importante lacune de la description linguistique du picard puisqu'il a permis de dresser une cartographie actualisée et critique des ressources lexicographiques très diverses rédigées depuis des siècles en langue picarde[15]. Cette historiographie a été accompagnée d'un important travail de description sociolinguistique et technique des multiples productions de cette nature, tantôt nommées “dictionnaires”, “lexiques”, “glossaires”, “vocabulaires” ou encore “parlers”.

En ce qui concerne les apports du TAL à ce projet, ces derniers se manifestent à travers l'établissement d'une base de données informatisée – prochainement disponible – permettant de naviguer à travers les différentes ressources bibliographiques listées et via une schématisation sous la forme d'une cartographie du domaine linguistique picard. Cette schématisation permettra aux utilisateurs de mesurer la répartition des ressources lexicographiques en fonction des grands pôles de pratiques lexicographiques, siècle par siècle et appellation par appellation.

---

[14] Cf. QUEMADA, 1968.

[15] REY Christophe (2021) *La langue picarde et ses dictionnaires*, Collection *Lexica, mots et dictionnaires*, n°38, Honoré Champion.

Plus limité que dans les projets PICARTEXT et RESTAURE, l'apport du TAL au projet METALPIC permettra toutefois de disposer de ressources modernes jusqu'alors inenvisageables il y a peu et sur lesquelles les locuteurs du picard pourront s'appuyer pour mieux appréhender la langue.

4. Vers des avancées salutaires

Les différents projets de recherche que nous avons brièvement présentés ci-dessus ont tous incontestablement contribué à faire de la langue picarde une des langues régionales de France les mieux dotées en termes de ressources linguistiques informatisées. Ces différentes initiatives constituent donc autant de chances pour cette langue de pouvoir s'affirmer – sociolinguistiquement parlant – non seulement comme une langue à part entière, mais aussi comme un système linguistique appréhendé et décrit par les outils du TAL.

Particulièrement précieuse, cette approche informatisée offre des avancées considérables dans la description des particularités linguistiques du picard, accordant à cette langue le statut de langue où la variation constitue davantage un modèle voire une condition d'existence et un facteur important de vitalité des pratiques linguistiques chez les néo-locuteurs notamment.

Toutefois, pour en revenir aux questionnements soulevés par le colloque à l'origine de cette contribution écrite, l'existence de ces différents projets, qui offrent en effet de “nouvelles approches” et de “nouvelles méthodologies” pour l'appréhension du picard, constitue-t-elle pour autant un facteur permettant la “revitalisation” de la langue ? Nous allons tenter d'apporter ci-dessous des éléments de réponse à cette interrogation en adoptant deux points de vue certes opposés mais reflétant néanmoins une réalité qui mérite d'être mise en exergue.

4.1 Un terreau fertile

L'une des particularités essentielles des différents projets de recherche qui ont été évoqués dans cette contribution réside dans le fait qu'un véritable continuum scientifique s'établit entre chacun d'eux. C'est précisément l'existence du précédent qui a permis la conception et la réalisation du suivant, témoignant ainsi qu'ils se sont

en quelque sorte nourris les uns les autres. Il s'agit là selon nous d'une démarche permettant d'illustrer le fait que l'entrée du picard dans l'univers du TAL s'est faite grâce à une prise de conscience progressive du mérite que cette "petite" langue avait d'être décrite selon ces nouvelles technologies. C'est en quelque sorte comme si la langue picarde avait gagné aux yeux des chercheurs qui s'y intéressent une nouvelle légitimité scientifique et le droit d'être investie par ces nouvelles technologies d'investigation.

Un terreau fertile, voilà comment nous percevons avec du recul la succession de ces divers projets de recherche. Avec ces travaux c'est une large porte qui s'est ouverte pour la langue picarde et sans doute un nouvel avenir qui s'est offert à elle. Nous en voulons pour preuve non pas les futurs projets qui seront imaginés pour poursuivre le long chemin qui vient de débuter – il ne s'agit précisément là que d'une question de temps que nous ne maîtrisons pas – mais plutôt les retombées concrètes que l'entrée du picard dans la galaxie des langues investies par le TAL a faites naître. Nous souhaitons plus particulièrement évoquer ici deux initiatives fondamentales pour la langue qui découlent directement de cette période fertile : 1) la création de la *Commission de néologie et de terminologie pour la langue picarde* et 2) la publication du *Dictionnaire fondamental français-picard* (2020) par Liudmila Smirnova et Alain Dawson.

Le gain en légitimité de la langue picarde que nous avons évoqué plus haut semble trouver l'une de ses plus belles illustrations à travers la création, il y a cinq ans déjà, d'une Commission de néologie et de terminologie pour la langue picarde, mise en place par l'Agence Régionale pour la langue picarde à la suite de la proposition de création que nous avions Fanny Martin et nous-même appelée de nos vœux lors de nos discussions informelles dans le cadre du projet RESTAURE.

Se réunissant à un rythme mensuel, cette instance auréolée d'une forme d'officialité en raison de son ancrage au sein de l'Agence régionale pour la langue picarde, a permis la tenue de nombreux débats, notamment orthographiques, parmi les acteurs les plus actifs de la promotion et de la défense de la langue et de la culture picardes. L'une des missions de cette commission étant justement de faire des propositions pour ancrer le picard dans une modernité lexicale, elle

a donc fait paraître ses travaux[16] sous la forme de recueils lexicaux thématiques[17] conçus selon un principe visant à fournir une représentativité des grandes variétés géographiques de picard.

Selon la méthode de travail déjà adoptée dans la traduction des derniers albums de la bande dessinée *Astérix*[18], la commission a ainsi assuré la diffusion de propositions lexicales sous la forme d'une "koïné" graphique homogénéisante, à la fois représentative de la variation de la langue sur son important territoire linguistique mais permettant de dépasser, et ce en dépit de l'existence de travaux antérieurs majeurs[19], l'absence d'une standardisation graphique.

Avec cette commission, la langue picarde et ses locuteurs disposent désormais d'une instance dont un mérite important, voire essentiel, est son existence même. Dans la mesure où elle permet d'ancrer la langue dans une modernité lexicale investie par les acteurs institutionnels, cette commission fait sortir le picard de l'ornière "patrimonialisante" dans laquelle beaucoup d'observateurs la cantonnent. Le picard, au même titre que la plupart des langues régionales de France n'est pas qu'un patrimoine, il est aussi une langue bien vivante dont il reste encore des locuteurs et des scripteurs. La création de cette commission, teintée d'une symbolique forte, est un signe intéressant – nous semble-t-il – de cette vitalité.

Illustrant le "continuum" que nous évoquions plus haut dans cet article, la publication en 2020 du *Dictionnaire fondamental français-picard* se situe dans la droite lignée des efforts entrepris par l'Agence régionale de la langue picarde et sa Commission de néologie et de terminologie pour donner un nouveau souffle au picard, un souffle permettant à la langue d'atténuer quelque peu une variation peut-être trop visible, peut-être excessivement valorisée et du coup sans doute handicapante pour assurer sa pérennité et son acquisition par de néo-locuteurs.

---

16 Disponibles ici : https://languepicarde.fr/artistes-picardisants/bibliotheque/.

17 Les trois premiers thèmes retenus ont été l'école, le jardin et les nouvelles technologies.

18 *Ch'Village copè in II* (2007) ; *Ch'cailleu d'étoéle* (2007) ; *L'Anniversaire d'Astérix et Obélix - Le Livre d'or en picard* (2010) et *El Crape as Pinches d'Or* (2013).

19 Cf. CARTON (1963 et 2004) et DEBRIE (1966).

Sur le modèle de l'expérience linguistique des années 50 envisageant la liste des mots nécessaires à un apprenant étranger pour assurer une communication au quotidien, le programme linguistique de cet ouvrage dirigé par Alain Dawson et Liudmila Smirnova permet de livrer la photographie d'un picard "essentiel", à savoir des équivalents picards des 1000 mots les plus fréquents du français contemporain.

Un des intérêts de cette proposition lexicographique est par ailleurs sa mise en synergie avec PICARTEXT puisque certains exemples proposés dans le corps de l'ouvrage ont été extraits de la base de données constituée à partir du projet éponyme.

Ajoutons également que cet ouvrage s'appuie explicitement sur le choix audacieux de la Commission de néologie et de terminologie en matière d'orthographe puisque comme le rappellent ses auteurs, "les traductions picardes proposées dans le dictionnaire privilégient une forme phonétique et morphologique standardisée, conforme aux recommandations de la Commission de néologie et de terminologie de l'Agence régionale pour la langue picarde"[20].

Afin que le lecteur puisse se rendre compte de la structuration des entrées de cet ouvrage, nous reproduisons en Figure 4 l'article AMUSER justement proposé par ses auteurs pour expliciter la microstructure retenue.

À de nombreux titres, le *Dictionnaire fondamental français-picard* constitue un évènement linguistique important pour la langue. Il ne fait donc nul doute, selon nous, que cet ouvrage non seulement bénéficiera d'un public large et conquis par son projet linguistique, mais permettra également au picard de bénéficier de nouvelles conditions pouvant assurer sa valorisation et sa pérennisation pour les générations futures.

[20] DAWSON, SMIRNOVA, 2020 : 13.

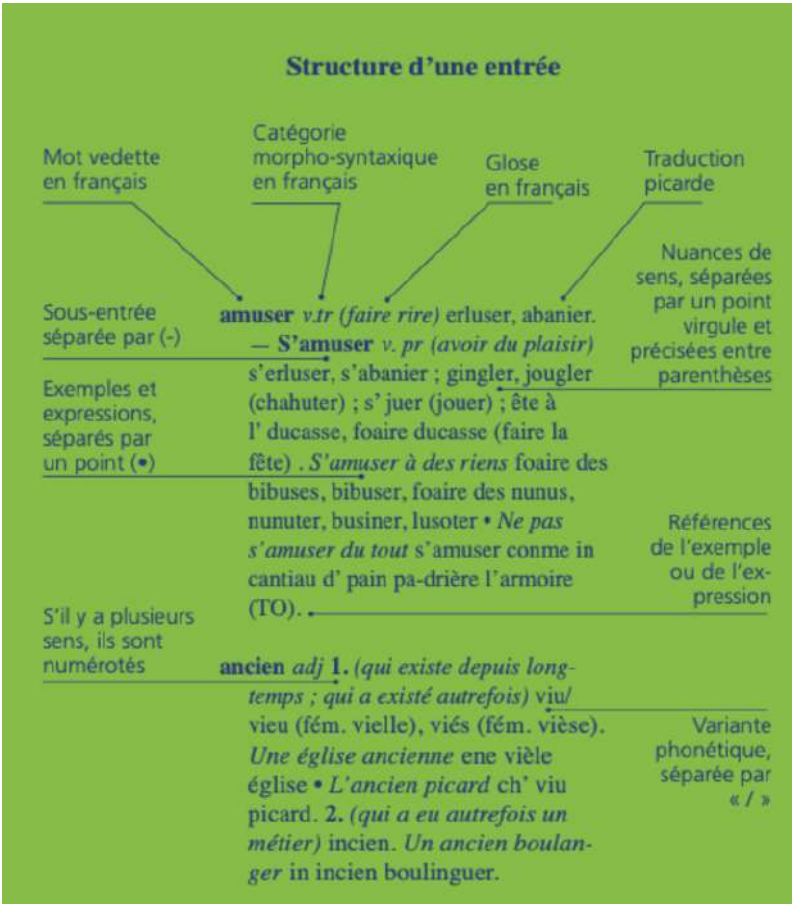


Figure 4. Structuration d'une entrée dans le *Dictionnaire fondamental français-picard*

Les deux initiatives que nous venons de mettre en lumière tendent à souligner que l'une des retombées inattendues de la vague de projets consacrés à la langue picarde et relevant du domaine du TAL a été une prise de conscience de la part des acteurs de sa promotion de la nécessité d'aller plus loin dans la description de la langue, notamment en mettant en place des structures et des ouvrages de référence au service de sa valorisation linguistique. Nous le répétons une nouvelle fois, l'implication de la langue picarde dans la galaxie

des projets relevant du TAL semble avoir fourni ce regain nécessaire de légitimité dont avait besoin cette langue en repli identitaire.

4.2 Une valorisation exigeante et en souffrance

À défaut de vouloir achever notre contribution par une vision plus critique, nous souhaitons clore cette dernière par un souhait : celui de voir mises en place les conditions de la valorisation des ressources générées par l'inclusion du picard dans le rang des langues régionales de France bénéficiant d'outils du TAL.

La mise au point de ressources linguistiques pour une langue donnée sans une exploitation de ces dernières constitue en effet un investissement incomplet. Le projet RESTAURE qui, rappelons-le, avait pour finalité d'éprouver les outils du TAL sur les langues régionales, notamment dans la perspective d'améliorer leur fonctionnement pour des langues comme le français ou l'anglais traditionnellement outillées et décrites par cette approche informatique, n'a fait qu'ouvrir les portes de futures études éventuelles.

En ce qui concerne les langues régionales de France, en tout cas précisément le picard, les outils et corpus constitués ne resteront que des promesses d'investigation tant que les moyens humains et matériels ne seront pas mis à la disposition de la recherche. Trop peu d'enseignants-chercheurs titulaires ont effectivement été impliqués dans la constitution des ressources évoquées et dans la réalisation concrète des projets qui les ont portées. Cette recherche d'excellence repose donc trop majoritairement sur des contrats ponctuels et l'implication de chercheurs au statut institutionnel précaire. Les départs non remplacés des spécialistes de langues régionales titulaires dans les laboratoires de recherche n'augurent pas d'une pérennité des recherches dans ce domaine pourtant essentiel. Dans un monde scientifique en pleine mutation, les travaux sur les langues de France, mais aussi d'ailleurs en Europe et dans le monde, doivent être poursuivis, à la fois pour nourrir une comparaison salutaire avec les langues au statut sociolinguistique plus favorable, mais aussi pour ne pas faire de ces langues les instruments d'une vision patrimonialisante rigide qui les cantonne au statut d'archives témoignant d'un passé linguistique aujourd'hui ou demain disparu.

Pour la langue picarde, l'approche du TAL est une occasion de pouvoir s'appuyer sur des outils conçus et exploités dans la perspective d'une forme de grammatisation de la langue, une grammatisation déjà effective grâce à un vaste corpus de grammaires et de dictionnaires existant[21], mais davantage brandie pour affirmer la grande variabilité de la langue sur son domaine linguistique. Il s'agit certes là d'une réalité linguistique précieuse, mais elle ne constitue pas, selon nous, une réponse appropriée à la nécessité de revitalisation de la langue, une revitalisation nécessaire pour ne pas enfermer encore davantage le picard dans son statut de langue en grand danger.

Les outils du TAL apportent une réponse possible à la revitalisation de la langue, à condition que les moyens financiers et humains soient mis au service de cette approche permettant de décrire et de relativiser la variation de la langue et d'éventuellement proposer des solutions pour à la fois la préserver tout en offrant les conditions d'une proposition homogénéisante salvatrice.

L'arrêté du 16 décembre 2021, permettant au picard d'être reconnu comme langue régionale de France et donc digne de pouvoir être enseigné dans les écoles, constitue une opportunité réelle de valorisation de ces ressources. Ces dernières pourraient même posséder, selon nous, un rôle important dans la réalisation de cette tâche cruciale.

5. Conclusion

À travers les différents projets de recherche que nous avons listés ici, nous espérons avoir montré que la langue picarde, très fortement caractérisée par ses phénomènes de variation linguistique sur l'ensemble de son territoire linguistique et donc particulièrement intéressante du point de vue du TAL, constituait, avec l'occitan, le corse ou encore l'alsacien, l'une des langues régionales de France les

[21] Soulignons le fait que nous en avons relevé l'existence de quatre auteurs picards ayant réalisé une grammatisation "complète" de la langue picarde en proposant à la fois une grammaire et un dictionnaire. Il s'agit des contributions de Daniel Haigneré (1901 et 1903), de René Debrie (1961, 1966, 1975, 1977, 1979, 1983b 1985, 1986, 1987; et 1983a), de Léon Maes (1979 et 1980) et enfin de celle de Gaston Vasseur (1996 et 1998).

mieux dotées en ressources informatisées. Les initiatives successives en faveur de l'élaboration de données numériques ont permis au picard de disposer d'un nouvel éclairage aux yeux de la communauté scientifique.

De l'intérêt des spécialistes de TAL a par ailleurs découlé un mouvement intéressant de "revitalisation" de la langue qui s'est notamment concrétisé par une prise de conscience de la nécessité de faire du picard une langue davantage tournée vers l'extérieur. La constitution d'une commission de néologie et de terminologie pour la langue picarde ainsi que la publication d'un *Dictionnaire fondamental français-picard* constituent ainsi deux propositions visant à impulser un travail de valorisation de la langue, non plus seulement en mettant en lumière les grandes variétés de picard elles-mêmes, mais en proposant une alternative homogénéisante qui peut s'imposer à l'avenir comme une façon de faire de la langue picarde une langue vivante, largement décrite et pouvant être enseignée au plus grand nombre[22].

Soulignons toutefois que le picard est aujourd'hui dans une situation quelque peu paradoxale dans la mesure où il s'agit de l'une des langues régionales de France en partie outillée par le TAL mais toutefois en situation de ne pas pouvoir véritablement bénéficier d'une exploitation effective de ces ressources faute de moyens humains et financiers mobilisés pour continuer à faire de sa description un enjeu scientifique important, un enjeu légitimant notamment l'existence de postes universitaires non précaires.

BIBLIOGRAPHIE

AUROUX Sylvain (1994) *La révolution technologique de la grammatisation*, Paris : Pierre Mardaga Éditeur.

BAIWIR Esther (2020) *Bien dire et bien aprandre* 35, numéro thématique intitulé "Les atlas linguistiques galloromans à l'heure numérique : projets et enjeux", Université de Lille.

BERNHARD Delphine & LIGOZAT Anne-Laure & MARTIN Fanny & BRAS,

[22] Précisons ici qu'enseigner un picard "koinéisé" artificiellement est, selon nous, davantage envisagé comme une inscription dans une démarche de visibilisation du picard en tant que langue de France capable d'être enseignée. Cela ne signifie pas nécessairement d'obliger les picardisants à écrire ce picard.

Myriam & MAGISTRY & al. "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. 2018.

BERNHARD Delphine & TODIRASCU Amalia & MARTIN Fanny & ERHART Pascale & STEIBLE Lucie & HUCK Dominique & REY Christophe (2017) "Problèmes de tokénisation pour deux langues régionales de France, l'alsacien et le picard", *Actes de l'atelier "Diversité Linguistique et TAL" (DiLiTAL 2017),* article en ligne : https://taln2017.cnrs.fr, pp. 14-23.

BRAILLON Jean-Marie (1991) *La graphie FIQP du picard,* Franke Innivierchitèie Pikarte éd Quiérache, Lemé, chez l'auteur.

CARTON Fernand & LEBEGUE Maurice (1989, 1998, 2004) *Atlas Linguistique et Ethnographique picard*, 2 volumes, Paris : Éditions du CNRS.

CARTON Fernand (1963) *Adaptation de l'orthographe Feller à la notation des parlers picards*, Nos Patois du Nord, décembre 1963 (supplément).

CARTON Fernand (2004) "Orthographier le picard : aperçu historique du débat entre "phonétistes" et partisans de graphies "françaises"", *Actes du Colloque international "Des langues collatérales"* (Amiens, 2001), Paris : L'Harmattan, vol. I, pp. 173-186.

DAWSON Alain & SMIRNOVA Liudmila (2020) *Dictionnaire fondamental français-picard*, Agence régionale pour la langue picarde.

DEBRIE René (1961) *Lexique picard des parlers nord-Amiénois*, Société de Dialectologie Picarde V, Archives du Pas-de-Calais, Arras.

DEBRIE René (1966) "Essai d'orthographe picarde", in *Courrier Picard*, n° des 18, 19 et 23 février.

DEBRIE René (1975) *Lexique picard des parlers ouest-amiénois*, Amiens : Centre d'Études Picardes, 424 p.

DEBRIE René & LOUVET Paul (1977) "Lexique picard du parler de Wailly-Beaucamp", in *Eklitra*, Amiens.

DEBRIE René (1979) "Lexique picard des parlers sud-amiénois", in *Eklitra*, Amiens.

DEBRIE René (1981) *Lexique picard des parlers du Vimeu*, Amiens : Centre d'Études Picardes.

DEBRIE René (1983a) *Èche pikar bèl é rade (le picard vite et bien),* livre-cassette - Omnivox, Centre d'Études picardes.

DEBRIE René (1983b) *Lexique picard des parlers est-amiénois*, Amiens : Centre d'Études Picardes, 153 p.

DEBRIE René (1985) *Lexique picard des parlers du Ponthieu*, Amiens : Centre d'Études Picardes, 128 p.

DEBRIE René (1986) *Lexique picard du Santerre*, Amiens : Centre d'Études Picardes, 103 p.

DEBRIE René (1987) *Lexique picard du Vermandois*, Amiens : Centre d'Études Picardes.

DUBOIS René (1957) *Le Domaine picard. Délimitation et carte systématique, dressée pour servir à l'Inventaire général du picard et autres travaux de géographie linguistique*, Arras : Société de dialectologie picarde.

ELOY Jean-Michel (éd.) (2004) *Des langues collatérales. Problèmes linguistiques, sociolinguistiques et glottopolitiques de la proximité linguistique*, 2 volumes, Paris : L'Harmattan et CEP Université d'Amiens, 490 p.

FORLOT Gilles & MARTIN Fanny (2014) "Entre invisibilité et (auto)occultation. Les paradoxes des pratiques langagières minoritaires en Picardie", in K. Djordjevic (éd.), *Les minorités invisibles : diversité et complexité (ethno)sociolinguistiques*, Limoges : Éditions Lambert-Lucas, pp. 77-87.

HAIGNERE, Daniel (1901) *Le patois boulonnais, Tome 1, Introduction, Phonologie et Grammaire*, Boulogne-sur-Mer : Société Académique de Boulogne-sur-Mer.

HAIGNERE, Daniel (1903) *Le patois boulonnais, Tome 2, Vocabulaire*, Boulogne-sur-Mer : Société Académique de Boulogne-sur-Mer.

MAES Léon (1979) *Grammaire mouscronnoise*, Mémoire de la Société d'histoire de Mouscron et de la région, tome I, fascicule I, 1979, 142 p., ill.

MAES, Léon (1979) *Lexique mouscronnois*, Mémoire de la Société d'histoire de Mouscron et de la région, tome II, fascicule I, 1980, 234 p., ill.

MARTIN Fanny (2015) *Espaces & lieux de la langue au XXIème siècle en Picardie. Approche complexe de la structuration des répertoires linguistiques en situations ordinaires - Enquête en Picardie*, sous la direction du Professeur Jean-Michel Éloy, Université de Picardie Jules Verne, Amiens.

MARTIN Fanny & FORLOT Gilles (2016) "Hétérogénéité linguistique et poids des idéologies sur les pratiques linguistiques en Picardie", in A. Boudreau & L. Arrighi (dir.), *Langue et légitimation: la construction discursive du locuteur francophone*, in *Les voies du français*, Laval : Presses de l'Université Laval, pp. 193-210.

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

QUEMADA, Bernard (1968) *Les Dictionnaires du français moderne (1539-1863). Étude sur leur histoire, leurs types et leurs méthodes*, Paris, Didier, in-8°.

VASSEUR Gaston (1996) *Grammaire des parlers picards du Vimeu (Somme)*, Abbeville : Paillart Éditions, 289 p.

VASSEUR Gaston (1998, [1963]) *Dictionnaire des parlers picards du Vimeu (Somme), avec considération spéciale du dialecte de Nibas,* Société de linguistique picarde, Éditions SIDES.

Christophe REY
EA 7518, Lexiques, Textes, Discours, Dictionnaires,
Institut Universitaire de France
christophe.rey@u-cergy.fr

Bases de données textuelles, lexique et syntaxe (occitan)

Chapitre 3

Nouvelles approches linguistiques et lexicographiques de l'occitan médiéval

Hervé Lieutard
ReSO, Université de Montpellier 3

Abstract

In the past few years, numerous large-scale research projects on Occitan have been carried out thanks to the development of the XML metalanguage and, in particular, to the contribution of the TEI (Text Encoding Initiative) format. XML and the TEI initiative have made it possible to shed a new light on medieval Occitan texts and to propose new editions with the potential for greater interoperability.

The digital critical edition of the *Petit Thalamus* (http://thalamus.huma-num.fr), which includes the books of the government of the Consulate of Montpellier, illustrates the advantages of these new technologies in the editing field. In particular, these technologies favour innovative transversal approaches and open up new research perspectives for diachronic linguistic studies on the history of the Occitan language. This edition was performed within the *Petit Thalamus* ANR project.

Beyond this editing work, the *Petit Thalamus* ANR project was also an opportunity to reflect concretely on the question of lemmatisation for the internal management of the diachronic variation inherent in a set of medieval documents written over several centuries.

At the end of the *Petit Thalamus* project, the possibilities opened up by the new publishing formats for connecting together considerable masses of data gave rise to the desire to go further and, in particular, to link all the digital

publishing projects underway on ancient Occitan within an international network. This network was created in 2018 under the name AcTo (Aculhir e Tornar, *Ressorsas numericas per l'occitan medieval*, https://acto.hypotheses.org). The aim of this network is to make encoding methods converge between all current and future publishing projects, which is a necessary condition for considering the creation of a virtual digital research space. It has already made it possible to envisage the setting up of a first project for the lemmatisation of medieval Occitan in conjunction with the DOM, the medieval Occitan dictionary (http://www.dom-en-ligne.de/) of the Bayerische Akademie der Wissenschaften. In this new project, the Petit Thalamus will serve as a privileged candidate for the creation of new lemmatization tools whose applications should subsequently be able to be extended to other medieval documents, and make it possible to link together the various projects concerning Occitan lexicography.

1. Introduction

Ces dernières années de nombreux projets de recherche 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 façon d'éditer les textes occitans médiévaux, notamment grâce à de nouvelles approches qui permettent d'envisager une interopérabilité de plus en grande entre les éditions de manuscrits médiévaux encodés dans ce format.

Un atout majeur de la recherche contemporaine en occitan est lié au fait qu'elle a su s'internationaliser depuis plusieurs décennies, notamment avec la création dès 1981 d'un important réseau de chercheurs, l'Association internationale d'études occitanes (AIEO[1]), qui fédère plusieurs centaines de chercheurs à travers le monde qui s'intéressent à la langue et à la culture occitanes dans toutes ses dimensions. Dans la série des projets pionniers issus de ce réseau, on peut citer celui du *Thesaurus occitan* (THESOC, http://thesaurus.unice.fr) de l'UMR 7320, Bases, Corpus, Langages à Nice, né lors du colloque de l'AIEO de 1989 à Wégimont consacré aux « Outils de la recherche occitane ». Cet outil, crée initialement par Jean-Philippe Dalbera, permet aujourd'hui de compiler et de mettre en ligne les données lexicales et morphosyntaxiques recueillies sur le terrain ou enregistrées

[1] http://www.aieo.org

dans les divers atlas linguistiques de l'occitan au cours du XX^e^ siècle. Le virage de la recherche en occitan vers le numérique a été progressif, mais s'est régulièrement adapté aux innovations technologiques dans ce domaine. Les liens entre chercheurs que l'AIEO a permis de tisser permettent aujourd'hui, grâce au développement du numérique, d'envisager et de concrétiser de nouvelles collaborations scientifiques internationales, comme c'est le cas par exemple pour l'édition de textes médiévaux, en particulier dans les domaines de l'écriture non-littéraire qui n'ont pas fait jusqu'ici l'objet de la même attention que les textes de *trobadors*, mais qui permettent d'enrichir nos connaissances sur le lexique, sur la langue et sur la place sociale de l'occitan au Moyen Âge.

Les nouveaux protocoles d'édition au format TEI, tout comme la mise en place d'espaces virtuels numériques de recherche ouvrent aujourd'hui des perspectives inédites de recherche. En permettant d'interconnecter des données éditées, il devient aujourd'hui possible non seulement d'envisager de nouveaux projets ambitieux de lemmatisation de l'occitan médiéval, mais aussi de relier le lexique médiéval avec la lexicographie occitane conçue dans toute son entendue diachronique.

2. L'édition électronique des textes médiévaux

Pour la langue médiévale, nous disposons déjà de plusieurs bases numériques pour la recherche. Certaines d'entre elles sont déjà anciennes. En raison de leur format, les *Concordanciers de l'occitan médiéval*, COM1 (2001) et COM2 (2005) en cédérom de Peter Ricketts, ne sont malheureusement pas exploitables techniquement aujourd'hui dans le cadre des projets numériques actuels. En revanche, ce n'est pas le cas pour le *Corpus linguistique du gascon ancien* de Thomas Field[2] ou pour le *Trésor manuscrit de l'ancien occitan*[3] (TMAO), sous la responsabilité de Dominique Billy à Toulouse, tous deux encodés en XML-TEI et qui donnent accès à des registres de la langue médiévale encore insuffisamment

[2] http://mllidev.umbc.edu/gascon/French/index.html

[3] http://tmao.aieo.org. Cette base de données textuelles accessible depuis le site de l'AIEO offre un accès limité aux membres de cette association.

étudiés, à savoir ceux que l'on trouve dans les manuscrits non littéraires occitans en prose disponibles des origines jusqu'au XVI[e] siècle. Les éditions en XML-TEI et leurs possibilités d'interopérabilité ouvrent de nouvelles possibilités de recherche pour aborder l'étude de la langue médiévale, mais aussi en permettant d'élargir les perspectives de recherche vers les diverses périodes ultérieures d'écriture, en particulier celles du déclin des formes standard qui permettent de faire le lien avec l'époque moderne, voire contemporaine.

2.1. L'édition numérique du Petit Thalamus de Montpellier

L'édition numérique du *Petit Thalamus* est le fruit d'un projet ANR[4] qui met aujourd'hui à disposition du public et des chercheurs une version numérisée du manuscrit AA9 des Archives Municipales de Montpellier, le manuscrit le plus important du gouvernement du Consulat de la ville médiévale de Montpellier établi dès 1204 (http://thalamus.huma-num.fr/). Ce manuscrit écrit en occitan pour les XIII[e], XIV[e] et XV[e] siècles et en français à partir de 1502, est indissociable d'un ensemble documentaire plus vaste, constitué de plusieurs autres manuscrits ou *Thalami* étroitement liés à l'émergence d'une gestion autonome de la ville par le Consulat montpelliérain.

Cette édition s'accompagne d'une introduction historique et linguistique, de la traduction française du texte occitan, de la contextualisation historique du texte par un riche apparat critique, de commentaires historiques et de deux index électroniques des noms de personnes et des lieux. L'édition du manuscrit principal, est complétée par la transcription et la mise en ligne en parallèle de sept autres *Thalami* qui ont précédé ou accompagné la rédaction de ce manuscrit.

Cette édition en XML-TEI est le fruit d'une collaboration pluridisciplinaire entre historiens médiévistes, juristes, historiens de l'art et occitanistes de l'Université de Montpellier. Si les historiens et les juristes pouvaient trouver un intérêt particulier à l'édition de ces manuscrits pour mieux appréhender comment le

[4] https://anr.fr/Projet-ANR-10-JCJC-2003

manuscrit AA9 procède à partir du XIVe siècle « à une recomposition mémorielle dont les différents manuscrits préservés permettent encore de suivre et de retracer la genèse »[5] ou pour mieux comprendre la capacité des juristes montpelliérains à adapter le droit romain aux nécessités locales, il s'agissait bien évidemment pour les chercheurs occitanistes de proposer l'édition critique d'une documentation fondamentale pour comprendre les conditions d'émergence d'une écriture en occitan dans les domaines pragmatiques et son développement sur plusieurs siècles afin de mieux documenter l'histoire de la langue occitane et en particulier l'émergence d'une forme de langue écrite officielle différente de celle qui caractérise le registre de la lyrique troubadouresque.

Dans le cadre de cette édition, le choix a été fait à travers l'édition en TEI de mettre en place les conditions d'une exploitation linguistique ultérieure de toutes les ressources textuelles numériques éditées. Aujourd'hui l'intégralité du manuscrit AA9 et des autres *Thalami* est éditée, à l'exception des parties juridiques (coutumes, établissements, serments) qui doivent encore être finalisées avant leur mise en ligne. Il est donc possible d'ores et déjà d'utiliser ces fichiers indépendamment de l'édition en ligne qui ne donne à voir pour l'heure qu'une partie du travail de balisage du texte effectué, à savoir celui qui concerne l'apparat critique, les notes historiques, les index de personnes et de lieux.

Ce travail d'édition numérique a été l'occasion de réfléchir aux nouvelles possibilités d'édition qu'offrait le numérique. Il ne s'agissait plus pour les auteurs de cette édition de faire des choix de reconstruction des textes en vue d'une édition papier comme cela avait été fait dans l'édition du XIXe siècle du *Petit Thalamus* (Pégat et al. 1840) réalisée selon des critères d'édition bien éloignés de ceux qui sont en vigueur aujourd'hui. L'édition électronique de ce vaste ensemble documentaire a donné la possibilité de faire émerger les diverses strates d'un texte, ce qui a nécessité de baliser tous les manuscrits du Consulat afin de pouvoir les lier entre eux

[5] http://thalamus.huma-num.fr/introduction/introduction-historique/annales-occitanes-partie-4.html

et permettre dans un second temps de les comparer pour évaluer les choix graphiques, morphologiques et syntaxiques des notaires du consulat sur une longue durée.

Par défaut l'édition électronique permet d'afficher l'image du manuscrit, la transcription, la traduction, les notes historiques, d'accéder aux noms de personnes et de lieux au moyen de liens hypertexte et de naviguer par année de rédaction au moyen d'une ligne de temps présente sur la page (Image 1).

Image 1 : Affichage de lecture par défaut du *Petit Thalamus*, année 1213[6]

La conception synoptique de l'interface permet aussi de comparer les divers manuscrits représentés au-dessus de chaque colonne par les lettres A B C D E F G H[7] , ce qui permet à la fois de mesurer les choix de récriture propres à chaque manuscrit et de concevoir une nouvelle approche de la langue et de la graphie d'un

[6] http://thalamus.huma-num.fr/annales-occitanes/annee-1213.html

[7] Des infobulles liées à chaque lettre indiquent les côtes des divers manuscrits.

point de vue diachronique jusqu'à la version finale représentée par le manuscrit AA9 (lettre H) comme en rend compte la mise en parallèle de quatre manuscrits différents pour l'année 1213 (Image 2).

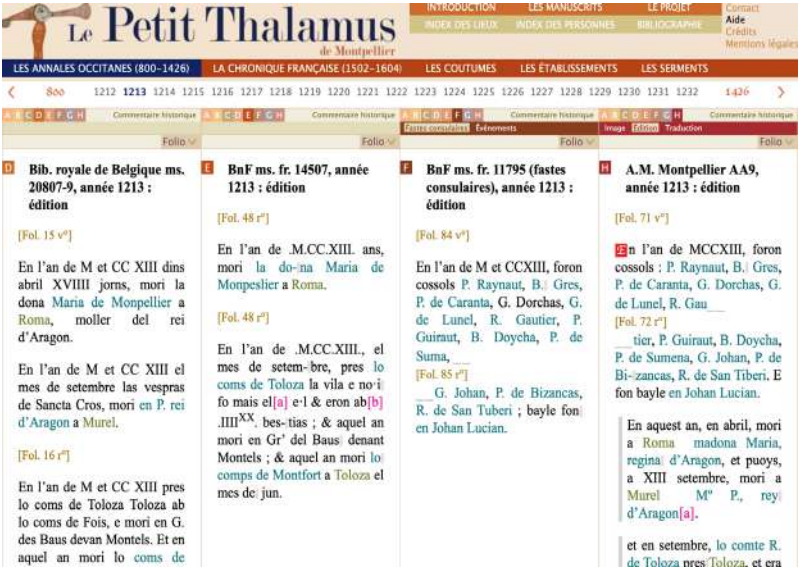


Image 2 : Affichage synoptique des divers manuscrits du Petit Thalamus, Année 1213

2.2. Étude de la langue

Étant donné la longue et complexe genèse du manuscrit AA9, qui commence au début du XIV^{e} siècle par la compilation et la réécriture de textes écrits au XIII^{e} siècle et se poursuit à travers des changements de rédacteurs jusqu'en 1426 en occitan – avant de reprendre en français entre 1502 et 1604 – le *Petit Thalamus* nous permet d'observer une « diachronie longue » et d'étudier une succession de chronolectes qui montrent une graphie, une morphosyntaxe et une syntaxe en nette évolution entre le XIII^{e} et le XV^{e} siècle (Tableau 1)

ms. 20807-9	ms. fr. 14507	ms. fr. 11795	H 119	AA 9
En l'an de M et CC XXII en mai prezeron Boissazon li home de Monpeslier. Et en aquel an mori en R. coms de San Gili e may.	En l'an M CC XXII, el mes de mai, fon pres Boissazon e prezeron lo los homes de Mon-peslier; en aquel an meteis, el mes de setembre, mori en R., coms de San-Gili, a Toloza.	En l'an de M et CC XXII, el mes de mai, fon pres Boichazon e prezeron lo li homes de Monpeslier; et en aquel an mori en R., coms de Toloza.	En l'an de .M. e. CC. e .xxii, el mes de mai, fon pres Boicharo, e prezero lo li home de Mont-p[e]l[ie]r. Et en aquel an mori en R., coms de Tolosa.	A mieg may fo pres lo luoc de Boyssezon per los homes de Montpellier; et aquel an mori lo dich comte de Toloza.

Tableau 1 : L'an 1222 dans cinq des manuscrits

À travers la réécriture du même temps à des époques différentes, il est possible de suivre le lent cheminement qui conduit vers un occitan moderne caractérisé non seulement par la disparition des derniers vestiges de la déclinaison médiévale ou la transformation de la morphologie verbale, mais surtout par une évolution vers une variété écrite suprarégionale, notamment à partir de la seconde moitié du XIV[e] siècle (Lieutard 2014).

En attendant la mise en place d'une lemmatisation plus avancée de l'ensemble des données éditées, la possibilité de faire facilement des requêtes X-Path dans les fichiers XML permet d'en extraire des données particulièrement intéressantes pour étudier la graphie et son évolution, notamment pour évaluer les solutions graphiques pérennes qui se mettent en place pour la notation de tous les phénomènes de palatalisation de l'occitan pour lesquels l'alphabet latin ne propose pas de solutions clés en main.

Pour ne donner que quelques exemples, il est possible sur le plan graphique de relever le décalage qui caractérise l'émergence de la notation de la nasale palatale par <nh> et de la latérale palatale par <lh>, ce qui porte à croire que c'est à partir de l'émergence de <nh> déjà présent dans les plus anciens manuscrits en latin[8] de la première moitié du XIII[e] siècle qu'ont été dérivées

[8] Dans le *Grand Thalamus* (ms AA4), on trouve par exemple les formes *Petrus Gazainnhaire* et *Raimundus de Luganhacco* pour transcrire les anthroponymes occitans *Pèire Gasanhaire* et *Raimond de Luganhac*.

les autres solutions en associant la palatalité au graphème <h> qui n'avait par ailleurs plus de fonction propre[9].

	J339	AA4	20807-809	Naf 4337	Fr. 11795	Fr. 14507	H 119	AA9
<lh>	0	1	0	0	193	2	70	2038
<nh>	10	33	8	24	300	12	64	2770

Tableau 2 : Nombre d'occurrences de <nh> et <lh> dans 8 manuscrits du *Petit Thalamus*

Cette stabilisation progressive d'habitudes graphiques, illustre, au-delà du caractère propre aux *Thalami* de Montpellier, une tendance plus générale de l'occitan écrit à s'orienter vers des variétés suprarégionales, en particulier à partir du XIV^e^ siècle et ouvre aussi la voie au-delà vers des études plus approfondies sur les usages graphiques en lien avec la place sociale de la langue pour reconstruire les processus de standardisation à l'œuvre au Moyen Âge. L'étude de la graphie du *Petit Thalamus* permet ainsi de contribuer à documenter l'existence de réseaux de diffusion des textes administratifs et d'évaluer comment les modèles graphiques se diffusent. Pour <lh> et <nh> par exemple, la carte figurant sur la page suivante (Image 3) permet de voir comment des solutions graphiques issues du centre de l'espace occitan ont continué de se répandre vers la Gascogne et la Provence au XIV^e^ et jusqu'aux confins les plus orientaux de l'espace occitan au XV^e^ siècle (Martel 2020).

L'étude du *Petit Thalamus* montre aussi comment il est possible d'évaluer le rapport entre écrit prestigieux et pratiques orales vernaculaires. La fonctionnarisation des notaires au sein-même du consulat au XIV^e^ siècle entrainera d'ailleurs une homogénéité de plus en plus grande des pratiques graphiques.

On sait d'après les sections juridiques du *Petit Thalamus* (ms. AA9, sections juridiques[10]) que les notaires du consulat devaient être nés montpelliérains ou vivre à Montpellier depuis au moins dix ans, ce qui laisse présumer que leur pratique orale de l'occitan

[9] Le [h] du gascon issu de F latin, bien que présent dès l'émergence de ce dialecte occitan, ne sera pas noté à l'écrit avant le XV^e^ siècle.

[10] Ces sections ne sont pas encore disponibles en ligne dans l'édition électronique.

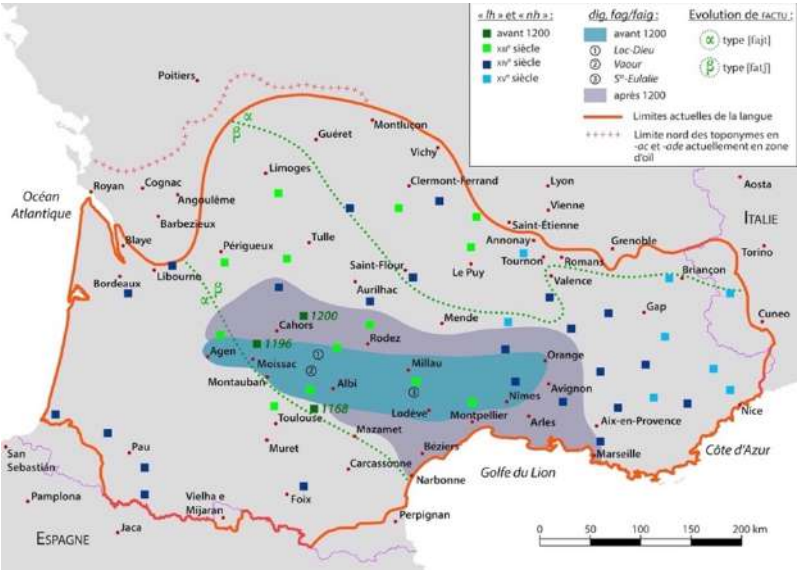


Image 3 : Premières attestations de *dig*, *fag/faig* et des digrammes <lh>, <nh> dans l'espace occitan (Martel 2020)[11]

était sans doute étroitement liée aux caractéristiques du languedocien oriental. Dans ce contexte, il est intéressant de relever la place donnée graphiquement à des traits linguistiques considérés comme typiquement montpelliérains, comme dans le cas de la diphtongaison conditionnée issue de /ɔ/, réalisée [jɔ] à Montpellier, mais de façon majoritaire [ɥɛ] en occitan ('nuit' se dit *nuèit* ou *nuèch* dans une grande partie de l'espace occitan, mais *nuòch* à Montpellier). Les études statistiques montrent que les choix graphiques du *Petit Thalamus* font référence à un espace plus large que celui des pratiques orales locales à travers la notation majoritaire de la diphtongaison en [ɥɛ], transcrite <ue>. On relève par exemple 68 occurrences de *nuèch*[12] contre seulement 8 de *nuòch*. Les formes locales en <uo> finissent d'ailleurs par disparaître complètement dans les dernières années de rédaction du ms AA9 au XV^{e} siècle où plus aucune trace de cette diphtongue

[11] https://journals.openedition.org/lengas/docannexe/image/4827/img-1.jpg
[12] Sous les formes *nueg* ou *nuech*

propre au montpelliérain n'est présente, alors même que c'est cette réalisation seule qui a subsisté jusqu'à aujourd'hui dans la pratique orale en languedocien oriental, après le déclassement social de l'occitan.

Il est d'ailleurs intéressant de relever à ce sujet, comme en témoignent les *Leys d'Amors* (Anglade 1919), un traité de grammaire et de rhétorique du XIV^e^ siècle émanant du *Consistòri del Gai saber*, l'académie poétique fondée à Toulouse en 1323, que c'est la même tendance à adopter des pratiques suprarégionales que l'on voit se dessiner à Toulouse à la même époque où tend à se généraliser le choix de *fach* 'fait', *dich* 'dit', *drech* 'droit', autrement dit des formes telles qu'elles apparaissent à Montpellier et qui ne correspondent pas aux réalisations dialectales toulousaines actuelles pour lesquelles l'évolution du groupe consonantique CT latin s'arrête au stade [jt] (*fait, dit, dreit*) sans connaître l'évolution ultérieure jusqu'à [ʧ] ([kt] > [jt] > [ʧ]) caractéristique des variétés orientales et septentrionales (Lieutard 2017).

Il est bien sûr difficile de dire si ces choix suprarégionaux, qui renvoient à des usages prestigieux de l'occitan, renvoient également à un sociolecte ou à des variations diastratiques qui auraient disparu au seul profit des formes populaires orales dans les siècles suivants. C'est en tout cas ce que pourrait porter à croire certains textes toulousains du XVI^e^ siècle pour la réalisation du bétacisme qui, initialement réservée au gascon, puis aux classes populaires, semble se développer dans l'ensemble du dialecte languedocien avec le déclassement social de l'occitan (Lieutard & Sauzet 2010).

Quoiqu'il en soit, l'édition électronique du *Petit Thalamus* représente un apport majeur pour l'étude de l'histoire de la langue occitane dans la mesure où elle permet d'étudier l'élaboration des usages graphiques en vigueur à Montpellier, de leur naissance jusqu'à la période de dé-standardisation, et permet de ce point de vue de décrire et d'étudier en détail un modèle d'évolution d'une variété locale vers une variété écrite suprarégionale qui gomme en partie la variation diatopique orale, en lien avec des usages socialement valorisés de la langue à cette époque.

3. Les projets de lemmatisation de l'occitan médiéval

Le projet ANR autour de l'édition du *Petit Thalamus* a également conduit ses concepteurs à s'interroger sur la question plus spécifique de la lemmatisation, notamment pour gérer l'ensemble des noms des 2400 personnes mentionnées dans les divers manuscrits du Consulat montpelliérain. Ces noms de personnes, majoritairement de consuls, qui apparaissent dans de nombreuses variantes graphiques au fil des manuscrits ont été associés à des entrées de l'index par des identifiants uniques dont la forme a été créée selon les règles de la norme graphique classique de l'occitan utilisée pour les usages contemporains. C'est une pratique jusqu'ici assez peu répandue dans l'édition de textes médiévaux. Il s'est toutefois avéré que ce choix d'un système graphique stable pour la lemmatisation permettait de rendre compte de la variation propre à l'écrit médiéval sans avoir à privilégier un des divers chronolectes enregistrés par les manuscrits : ainsi, par exemple, toutes les occurrences du nom de famille *Puèg* ou *Puòg*, dont les formes apparaissent dans diverses variantes graphiques, conduisent à une seule entrée PUÈG. La mise en relation des variantes à l'aide d'une graphie normalisée assure d'une part l'identification sans ambiguïté des entrées, d'autre part la variation diachronique et diatopique reflétée par le manuscrit médiéval peut être conservée sans aucune intervention normative dans la transcription des manuscrits (Image 4).

Dans la mesure où la graphie classique contemporaine trouve son inspiration dans les pratiques médiévales, elle se montre apte à englober la variation, que ce soit en synchronie ou en diachronie.

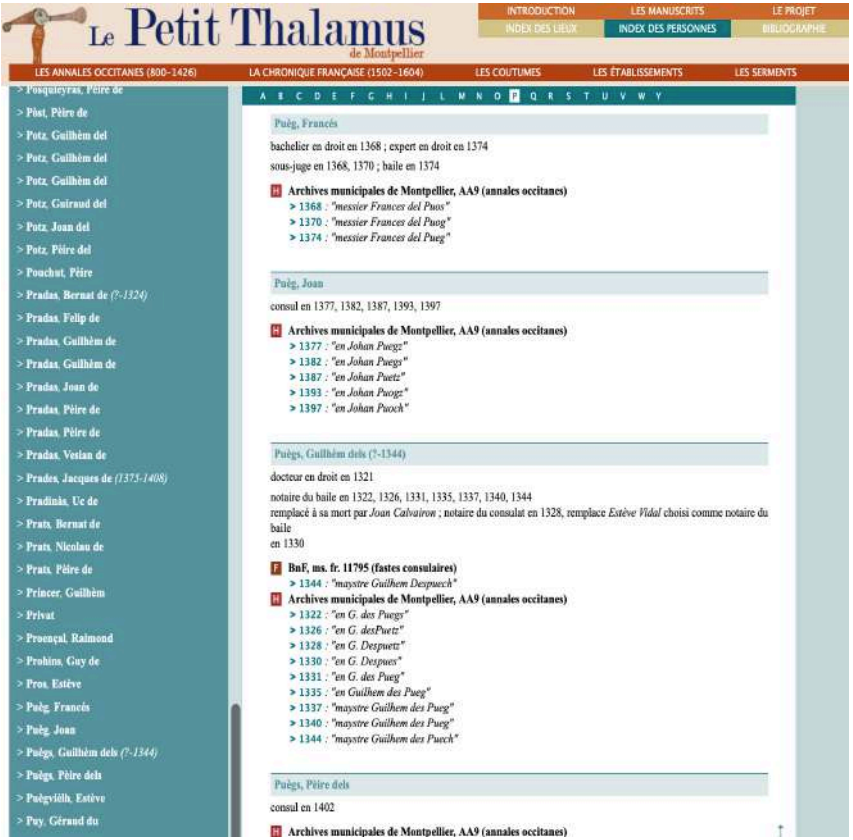


Image 4 : Capture d'écran de l'index des noms du *Petit Thalamus* : le patronyme *Puèg*

Au bout du compte, ce choix rejoint la façon dont il est possible de rendre compte de la variation diatopique moderne de l'occitan (Image 5), en associant la variation phonétique enregistrée à l'oral à une forme lemmatisée, comme c'est déjà le cas dans le THESOC (Brun & Sauzet 2013).

Localisation (code INSEE)	Tr. API	Tr. Graphiq.	Lemme
Gaillac (81099) *(ALLOc 81.05 • THESOC L382)*	pˈɛts		puèg
Cadalen (81046) *(ALLOc 81.07 • THESOC L384)*	pjˈɔts		puèg
Pampelonne (81201) *(ALLOc 81.10 • THESOC L385)*	puˈɛts		puèg
Saint-Julien-Gaulène (81259) *(ALLOc 81.11 • THESOC L386)*	puˈɛts		puèg
Fauch (81088) *(ALLOc 81.12 • THESOC L387)*	pjˈɔts		puèg

Image 5 : Exemple de variation phonétique diachronique de *puèg* dans le Thesoc

En faisant de la graphie classique un outil scientifique susceptible de faire converger toute la variation graphique observable à travers les siècles vers des formes lemmatisées clairement identifiables (et ce même pour les graphies dites oralisantes à partir des XV^{e} et XVI^{e} siècles), il devient possible d'aligner et de comparer des résultats de recherche sur la langue ancienne et sur ses usages plus contemporains. C'est en tout cas-là un enjeu essentiel dans le cadre de projets de lemmatisation ambitieux qui se proposent de lier entre elles les données sur la langue médiévale et sur la langue moderne.

À partir de cette édition du *Petit Thalamus*, le pari a été fait qu'à l'avenir la recherche en occitan pourrait dépasser en partie le clivage qui tend encore parfois à persister aujourd'hui entre études sur la langue médiévale et études sur la langue moderne et contemporaine pour dessiner une histoire complète de la langue occitane en lien avec sa place sociale, et ce grâce aux réseaux numériques pilotés par des chercheurs travaillant avec des protocoles de recherche communs.

3.1. Vers un Espace numérique virtuel de recherche : le projet AcTo

Suite aux premier résultats prometteurs fournis par l'édition électronique du *Petit Thalamus* un groupe de chercheurs montpelliérains occitanistes et linguistiques a mis en place en 2018 le projet d'un espace virtuel de recherche appelé AcTo (Acolhir e Tornar)[13] depuis 2018. L'objectif de ce projet, financé par l'université Paul-Valéry Montpellier 3, est de créer un réseau de projets de numérisation en ancien occitan rassemblant divers pays européens (pour l'instant Allemagne, Espagne, États-Unis, France, Italie, Royaume Uni). Le but de ce consortium est de promouvoir la visibilité des divers projets d'édition de textes médiévaux occitans, leur libre accès, l'interopérabilité des ressources et la durabilité des réalisations numériques. Le réseau AcTo est placé sous le patronage de l'AIEO qui, comme déjà évoqué, favorise les liens entre chercheurs occitanistes au niveau international. Il s'est donc fixé comme but initial de faire converger les méthodes d'encodage entre tous les projets d'édition en cours et à venir afin de pouvoir les relier entre eux dans un second temps et réaliser sur le long terme un espace numérique virtuel (ENV) de recherche spécifique.

Le réseau AcTo doit aussi à terme permettre de développer la relation avec le CIRDOC (Centre international de recèrca e documentacion occitanas) ou encore avec les grandes infrastructures numériques nationales et internationales (Huma-Num, CLARIN...) pour augmenter la visibilité des projets occitans (Caiti-Russo et al. 2019).

3.2. Le projet de lemmatisation de l'occitan médiéval

Suite à plusieurs séminaires de travail, les discussions au sein du réseau AcTo ont permis d'identifier plusieurs champs de travail en commun : l'objectif central du projet reste bien sûr d'unifier les critères de normalisation que les projets numériques en occitan utilisent dans l'annotation, mais grâce à ses données lexicographiques exhaustives, il est surtout apparu que le DOM[14], le Dictionnaire d'occitan médiéval développé depuis plusieurs

[13] https://acto.hypotheses.org/
[14] http://www.dom-en-ligne.de/

années à Munich pouvait jouer le rôle central d'un système de référence largement reconnu pour la lemmatisation et pour accélérer le processus d'unification au sein de ce réseau. Dans le domaine de la lexicographie occitane médiévale, ce dictionnaire de la Bayerische Akademie der Wissenschaften est devenu une référence numérique incontournable. Livré autrefois sous forme de fascicules, de 1993 à 2013, il s'est transformé aujourd'hui en une véritable base de connaissances. Il permet bien sûr l'accès au lexique et aux bases de lemmes, mais sa conception modulaire actuelle offre aussi la possibilité d'accéder aux dictionnaires de Raynouard ou de Levy, sous forme d'images numérisées, voire même d'être dirigé directement vers les pages du FEW.

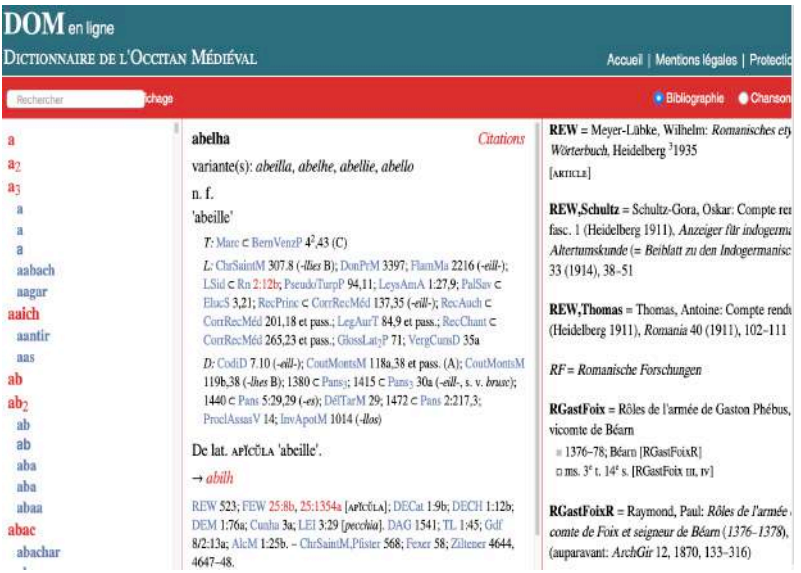


Image 6 : Interface de consultation du DOM en ligne[15]

Partant du constat que les langues anciennes ou médiévales, ou les langues riches en variation, graphique ou morphologique, sont encore peu dotées d'instruments de lemmatisation performants, et

[15] http://www.dom-en-ligne.de/dom.php?lhid=7KWaZNj4P6QW1TVlAEVC4g

qu'aucun jeu de balises morphosyntaxiques ni aucun système d'annotation syntaxique n'est encore réellement disponible pour l'occitan, les travaux relatifs à l'élaboration de stratégies communes de lemmatisation et d'annotation ont fait l'objet de séminaires entre le groupe de recherche de Montpellier, le groupe de recherche allemand du DOM et des collègues de l'École des Chartes (Paris).

Des algorithmes permettant d'améliorer l'annotation automatique des textes occitans ont déjà été testés, notamment ceux qui ont été utilisés pour le projet OMELIE (Outils et Méthodes pour l'Édition Linguistique Enrichie), développé par Jean-Baptiste Camps et Frédéric Duval dans le cadre de la cellule humanité numérique de l'École nationale des Chartes, le but étant de disposer d'un environnement de travail relativement simple dans lequel les éditions en TEI peuvent être téléchargées, lemmatisées automatiquement et annotées avec des étiquettes morphosyntaxiques, avant d'être corrigés manuellement pour améliorer les modèles.

Le projet OMELIE utilise Pyrrha, une application de post-correction collaborative de lemmatisation et d'étiquetage morpho-syntaxique (POS-tagging) qui rend notamment possible le traitement des langues qui présentent une grande variation graphique ou dont les règles morphosyntaxiques ne sont pas homogènes (Clérice et al. 2021). Le DOM y est utilisé comme référentiel de lemmes. Comme pour l'annotation, la lemmatisation est semi-automatique et peut-être continuellement améliorée grâce aux routines habituelles d'entraînement et de révision.

La dernière génération de lemmatiseurs permet donc de pallier en partie le manque d'outils spécifiques, mais le passage d'outils à base de règles à des approches fondées sur l'apprentissage profond implique l'utilisation de grandes quantités de données d'entraînement, nécessitant un important travail de reprise et correction. Il a donc été décidé que du fait même de sa conception, de son intérêt et de son ampleur documentaire, l'édition électronique du *Petit Thalamus* servirait de candidat privilégié pour l'élaboration de nouveaux outils de lemmatisation de l'occitan ancien et que les applications mises au point devraient pouvoir s'étendre par la suite à d'autres documents édités en TEI.

3.3. Les recherches à venir sur l'histoire de la langue

Ce travail de lemmatisation du *Petit Thalamus* n'est qu'un premier pas qui doit conduire à terme à constituer et à publier un corpus linguistiquement annoté de l'ensemble des textes de l'ancien occitan. À partir de ce corpus, il sera alors possible de constituer un thesaurus qui répertorie sous une entrée unique toutes les variations graphiques d'un même lemme ou d'analyser un corpus étendu pour observer, par exemple, l'état du système ou encore étudier certains traits dialectaux

La collaboration entre les chercheurs français et les chercheurs allemands qui est engagée vise à terme à contribuer à une meilleure compréhension de l'histoire de l'occitan entre la continuité linguistique (orale) et la discontinuité (écrite). Les données obtenues seront utiles pour reconstruire le processus de standardisation et de dé-standardisation de l'occitan médiéval dans les domaines graphiques et lexicaux. Mais cette collaboration et cette volonté de travailler en réseau vise aussi à assurer une plus grande prise en compte d'une langue qui n'a pas jusqu'ici été suffisamment étudiée sur le plan linguistique et pour laquelle il n'existe pas suffisamment de ressources numériques.

Un dernier objectif d'ores et déjà envisagé, mais il n'interviendra pas dans l'immédiat, consistera à relier le thesaurus historique aux dictionnaires numérisés des dialectes occitans modernes pour constituer des thésaurus numériques complets du lexique occitan. En reliant les données médiévales à des bases de données lexicales synchroniques de la langue occitane moderne en cours de constitution, comme par exemple celles du *Congrès*[16], également en XML/TEI, ou du *Thesoc*[17], il sera sans doute bientôt possible d'ouvrir de nouveaux champs d'investigation et notamment de mieux documenter et évaluer la continuité du lexique à travers toute l'histoire de la langue occitane, du Moyen Âge jusqu'aux usages les plus contemporains, et, par extension, de mieux comprendre l'histoire d'une langue caractérisée à travers toute son évolution par une discontinuité des pratiques écrites,

[16] https://locongres.org/oc/
[17] http://thesaurus.unice.fr

entre des périodes de standardisation et de dé-standardisation. Ces avancées dans le domaine de la lexicographie devraient pouvoir permettre également de mieux évaluer les évolutions de la graphie au fil des siècles, de la constitution de formes suprarégionales standardisées aux diverses formes locales écrites proches des variétés parlées en fonction des époques. En outre, l'alignement pourrait servir à des comparaisons quantitatives et qualitatives des deux inventaires, médiévaux et contemporains, pour déterminer plus précisément la continuité lexicale dans des domaines sémantiques variés ou encore le rôle joué par la perte des domaines lexicaux relatifs à l'écrit.

4. Conclusion

Les nouveaux protocoles d'édition des textes médiévaux ouvrent de nouvelles perspectives de recherche à une échelle de grandeur qu'il était difficile de concevoir il y a quelques années encore mais aussi en donnant aux chercheurs les moyens de dépasser les obstacles qui semblaient encore infranchissables jusqu'ici entre les études sur la langue médiévale et sur la langue contemporaine.

Il devient aujourd'hui possible de concevoir que les nouvelles technologies utilisées dans la recherche nous permettront d'ici quelques années, grâce à la mise en relation des données numériques sur l'occitan, de donner une image plus complète de l'histoire de la langue occitane que celle dont nous disposons jusqu'ici.

L'exploitation linguistique des résultats obtenus pourrait permettre de mieux évaluer les aspects de la tradition graphique dont le développement reflète directement la formation de réseaux de communication suprarégionaux et l'institutionnalisation de la culture écrite, deux éléments centraux du processus de standardisation. Le lien entre la variation écrit/oral, la variation entre langue de proximité et langue de distance et les processus de standardisation et de dé-standardisation représentent depuis longtemps un aspect central de la recherche sur l'histoire des langues romanes dans laquelle celle de l'occitan n'occupe pas encore toute la place qu'elle devrait occuper.

BIBLIOGRAPHIE

ANGLADE Joseph (ed.) (1919-1920), *Las Leys d'amors: manuscrit de l'Académie des Jeux floraux*. 4 vol. Tolosa : Privat.

BRUN, Guylaine, SAUZET, Patric, (2013), « Le Thesaurus Occitan : entre atlas et dictionnaires », *Corpus* 12 . 105-140.

CAITI-RUSSO Gilda, CAMPS Jean-Baptiste, COUFFIGNAL Gilles, FRONTINI Francesca, LIEUTARD Hervé, REICHLE Elisabeth, SELIG Maria (2019), "AcTo: How to Build a Network of Integrated Projects for Medieval Occitan". In Kirl SIMOV ; Maria ESKEVICH, Maria (eds), *Proceedings of the CLARIN Annual Conference 2019*. Leipzig. 134–137.

CLERICE Thibault, LEVENSON Matthias Gille, ING Lucence, PINCHE Ariane, GABAY Simon, CAMPS Jean-Baptispte (2021), « Lemmatiser des textes et corriger l'annotation grâce à l'apprentissage profond avec Pyrrha », Congrès Colloque Humanistica.

https://hal.archives-ouvertes.fr/hal-03224112.

LIEUTARD Hervé (2014), « La grafia classica de l'occitan al servici de l'antroponimia medievala ». *Amb un fil d'amistat, Mélanges offerts à Philippe Gardy*. Toulouse : Centre d'étude de la littérature occitane. 667-678.

LIEUTARD Hervé (2017), « Emergéncia e elaboracion d'un occitan oficial escrich prediglossic: l'exemple del Pichòt Talamus ». *Occitània en Catalonha : de tempses novèls, de novèlas perspectivas*. CARRERA, Aitor et Isaval GRIFOLL, Isabel (eds), Lheida :BPLT. 203-214

LIEUTARD Hervé (2021), « Les apports récents et à venir du numérique pour la recherche en domaine occitan ». *TENSO: Bulletin of the Société Guilhem IX* 36. 171-176.

LIEUTARD Hervé & SAUZET Patrick (2010), « D'une diglossie à l'autre: observations linguistiques et sociolinguistiques sur deux textes toulousains de 1555 : *Las Ordenansas e coustumas del libre blanc* et *Las nonpareilhas receptas*». In Jean-François COUROUAU, Philippe GARDY, Jelle KOOPMANS (eds.), *Autour des quenouilles, la parole des femmes* (1450-1600). Turnhout : Brepols. 109-145.

MARTEL Philippe (2020) , « À la naissance de deux graphèmes-symboles : LH et NH », *Lengas* [En ligne], 88, consulté le 18 janvier 2022. https://doi.org/10.4000/lengas.4827.

PEGAT Ferdinand, THOMAS Eugène et DESMAZES Casimir (1840), *Le petit thalamus de Montpellier* [Texte imprimé] : *Thalamus parvus* / publié pour la première fois, d'après les manuscrits originaux, par la Société archéologique de Montpellier. Montpellier : J. Martel ainé.

Hervé Lieutard
ResO, Université de Montpellier 3
herve.lieutard@univ-montp3.fr

Chapitre 5

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

Myriam Bras
Université de Toulouse
CLLE, UMR 5263, CNRS & Université Toulouse Jean Jaurès

Abstract

This chapter is about resources and tools for Occitan linguistics. It assesses the progress made in the last two decades by the creation of the first textual database for Occitan, BaTelÒc, and the design of a Natural Language Processing pipeline including a tagger and a parser. These advances enable improvements of BaTelÒc with the implementation of linguistic annotation of texts. We also explain which strategies were used in these developments, focusing on cooperation and re-use of existing tools and resources. The Occitan language can now benefit from corpus linguistic methods and multilingual NLP, and contrastive linguistics can now benefit from Occitan corpora and resources.

1. Contexte et enjeux pour la linguistique occitane

Cet article est consacré aux données et aux ressources dont disposent – ou pourraient disposer – les linguistes qui travaillent sur l'occitan. Nous allons dans un premier temps situer brièvement la langue sur les plans linguistique et sociolinguistique avant d'entrer dans le vif du sujet.

L'aire linguistique de l'occitan s'étend sur le tiers sud du territoire français et déborde dans le Val d'Aran en Espagne et plusieurs vallées en Italie (voir Figure 1). C'est une langue romane occupant une position médiane dans la Romania. Elle est classée dans le groupe des langues gallo-romanes avec le français, les langues d'oïl, le franco-provençal et le catalan. Dans la classification de Bec (1970), elle forme avec le catalan, dont elle est proche, le sous-groupe occitano-roman. Elle est organisée en six grands dialectes, le languedocien, le gascon, le limousin, l'auvergnat, le provençal, le vivaro-alpin, auxquels s'ajoute une zone d'interface avec les variétés d'Oïl, le croissant (Figure 1).

Figure 1. Aires linguistiques des dialectes occitans d'après (Bec 1995)[1]

[1] Carte réalisée par Jean Sibille pour (Bernhard et al. 2021, p.291), les noms des dialectes sont en occitan.

L'occitan est une langue vivante encore parlée par des centaines de milliers de locuteurs[2]. Elle est enseignée en immersion dans le réseau des établissements associatifs Calandreta[3] et dans la modalité bilingue à parité horaire dans l'enseignement public, les effectifs de ces deux types d'enseignement « intensif » ne représentant qu'une très faible proportion de la population scolaire de l'école au collège[4]. Son usage est soutenu par différents organismes associatifs ou publics soutenus par les collectivités comme l'Office Public de la Langue Occitane[5], Le Congrès Permanent de La Lenga Occitana[6], Le CIRDOC- Institut Occitan de Cultura[7].

Ces signes apparents de vitalité ne doivent pas masquer que la langue occitane est une langue en danger, en particulier à cause de la disparition progressive des locuteurs traditionnels et de l'absence quasi totale de la langue dans la société en dehors des lieux dédiés à sa pratique.

Dans ce contexte, l'existence d'une communauté de recherche en sociolinguistique et en linguistique occitane représente un enjeu crucial, d'une part pour documenter la langue et ses usages, d'autre part pour soutenir sa vitalité en contribuant à la construction de ressources pour ses locuteurs (Bras et al. 2021).

Dans cet article, nous proposons de dresser un bilan des avancées réalisées depuis le début des années 2000 en matière de ressources et d'outils pour la linguistique occitane, à partir de notre expérience de linguiste sémanticienne, pratiquant une linguistique sur corpus et

[2] Leur nombre est difficile à évaluer car il existe beaucoup de locuteurs passifs et des niveaux de compétences très variés. On peut néanmoins estimer cet effectif entre 500 000 et un million en s'appuyant sur plusieurs enquêtes socio-linguistiques récentes (Bernhard et al. 2021).

[3] http://calandreta.org

[4] Environ 13 000 élèves en 2021-2022.

[5] L'OPLO est un Groupement d'Intérêt Public Etat-Région Nouvelle Aquitaine, Région Occitanie, crée en 2016, voir https://www.ofici-occitan.eu

[6] Le Congrès est un organisme associatif interrégional de régulation de la langue occitane, créé en 2011, voir https://locongres.org

[7] Le CIRDOC – Institut Occitan de Cultura existe depuis 2019 sous la forme d'un établissement public, il rassemble des organismes fondés en 1975 pour le CIDO devenu CIRDOC et en 1996 pour l'Institut Occitan.
voir https://www.oc-cultura.eu/decouvrir/histoire-et-actualite/

progressivement initiée à la linguistique outillée et au traitement automatique des langues.

Notre premier constat, alors que nous souhaitions mener sur l'occitan des études dans le domaine de la sémantique temporelle (Bras 2005), a été celui de l'absence de données textuelles facilement accessibles pour la linguistique descriptive de l'occitan contemporain : ce travail avait pu en effet être mené grâce à une exploitation manuelle de textes au format papier et d'un petit corpus de textes numérisés.

Cette expérience laborieuse, en comparaison avec les explorations des données du français auxquelles nous étions habituée, a motivé la création d'une base textuelle pour la langue occitane, sur le modèle de la base Frantext, au service de la recherche en linguistique, puis la création d'outils et de ressources dans l'objectif d'enrichir les textes de la base en annotations linguistiques afin d'améliorer les requêtes.

Nous présenterons cette base textuelle dans la section 2, puis nous illustrerons son utilisation en linguistique occitane dans la section 3, avant de présenter, en section 4 les outils et ressources créés par la suite. Nous terminerons par un bilan des avancées de la linguistique occitane suivi de perspectives en section 5.

2. Première étape : construction d'une base textuelle pour la langue occitane (2006-2016)

Le projet de construction d'une base textuelle pour la langue occitane a commencé en 2006 dans le cadre du laboratoire ERSS (UMR 5610 de l'Université Toulouse Le Mirail et du CNRS). L'ambition était de réunir des textes de genres variés (roman, théâtre, poésie, conte, presse, chroniques, etc.) écrits entre le XIX^{ième} et le XXI^{ème} siècles et d'accueillir la variation dialectale et graphique (Bras 2006). Nous avons suivi le modèle de la base textuelle Frantext développée à l'ATILF à Nancy pour la linguistique française en structurant la base et en encodant les textes dans les formats standards de constitution et de diffusion de corpus (format xml, norme TEI P5).

Une première base expérimentale a pu être mise en ligne en 2008 avec 15 textes confiés par un éditeur associatif partenaire[8]. L'accès était réservé aux membres du projet, il s'agissait de mettre à l'épreuve la faisabilité du projet[9] (Bras et Thomas 2011). Puis la base s'est enrichie de nouveaux textes grâce à de nouveaux partenaires[10].

Entre 2012 et 2014, nous avons constitué deux corpus spécifiques dans le cadre d'un projet conjoint[11] avec un laboratoire de littérature (PLH) et un laboratoire d'anthropologie (LISST-CAS) de notre université : un corpus de textes d'auteurs du Rouergue et un corpus de contes littéraires. Puis nous avons continué à enrichir la base dans le cadre d'un projet ANR consacré à la création de ressources pour plusieurs langues de France, l'alsacien, le picard et l'occitan, le projet RESTAURE[12]. Nous avons développé un moteur de recherche opérationnel[13] permettant deux modes de recherche : la recherche simple permettant de visualiser les contextes d'emploi d'un mot (concordances) ; la recherche avancée permettant d'extraire les contextes d'emploi de formes (mots, parties de mots et séquences de mots). Dans les deux cas, la taille des contextes permet de respecter le droit de citation pour les œuvres sous droits.

En juin 2016, nous avons mis en ligne la base opérationnelle, nommée BaTelÒc[14], contenant 95 œuvres ou textes de 49 auteurs différents couvrant les principaux dialectes de l'occitan (languedocien, gascon, provençal, auvergnat, limousin, vivaro-alpin) dans des proportions variées en termes de nombre de mots (58% de languedocien, 17% de gascon, 16% de provençal, 9% des autres dialectes). Les dates de création ou d'édition des textes se répartissent de 1836 à 2014. Les textes relèvent de genres variés (40 classés dans

[8] IDECO / IEO Edicions : https://ideco-dif.com/

[9] Avec le soutien du CNRTL, de la DGLFLF et du CROM.

[10] Editeurs Reclams, Lo Clusèl, ADEO, Editions des Régionalismes et organismes associatifs oeuvrant pour la diffusion de l'occitan : Le Congrès permanent de la lenga occitana, le CIRDOC, les associations GIDILOC et CIEL d'OC.

[11] Projet cofinancé par l'Université Toulouse Jean-Jaurès et la Région Midi-Pyrénées.

[12] Projet RESTAURE (RESsources informatisées et Traitement AUtomatique pour les langues REgionales), ANR-14-CE24-0003, 2015-2018 : https://restaure.unistra.fr

[13] Avec l'aide de Franck Sajous, ingénieur au laboratoire CLLE.

[14] redac.univ-tlse2.fr/bateloc/

le genre ‘roman’, 10 recueils de textes classés dans le genre ‘conte littéraire’, 18 recueils dans le genre ‘poésie’, 12 recueils dans le genre ‘nouvelles’, 8 textes dans ‘mémoires et chroniques’, le reste se répartissant dans les genres ‘chroniques’, ‘traités’ et ‘essais’). Ils sont écrits dans plusieurs graphies (graphie classique ou alibertine, graphie mistralienne ou autres graphies).

La base BaTelÒc, les étapes de sa mise au point, et ses possibilités d’utilisation sont décrites en détail dans (Bras et Vergez-Couret 2016). Nous soulignerons ici que nous avons voulu une base ouverte, qui puisse accueillir toutes les œuvres, quelle que soit leur graphie ou leur variété de langue. En d’autres termes, nous avons évité de proposer un « corpus de référence » pour proposer aux utilisateurs un ensemble de textes assez large pour que chacun puisse y puiser les textes pertinents pour son étude, sous la forme d’un « corpus de travail ». Des corpus prédéfinis peuvent cependant aider les utilisateurs qui ne sont pas encore assez expérimentés pour se construire un tel corpus. Un « còrpus descobèrta » (corpus découverte) permet en particulier de faire ses premiers pas dans BaTelÒc.

La base est de taille modeste, actuellement 3,37 millions de mots, mais le nombre de textes écrits en occitan pour les périodes contemporaine et moderne rend possible un accroissement significatif. L’objectif est de constituer le réservoir de textes le plus représentatif possible de la variation diatopique (dialectes) et diachronique (dates d’écriture des textes), dans une variété de genres. Nous souhaitons également intégrer la variation diamésique en intégrant des transcriptions de textes oraux et des pièces de théâtre (genres non encore abondés) pour permettre des études comparatives entre l’écrit et l’oral.

Nous présentons dans la section suivante quelques illustrations de BaTelÒc pour la linguistique occitane.

3. Premières avancées pour linguistique occitane outillée grâce à BaTelÒc

Les requêtes possibles dans les corpus de textes constituables dans BaTelÒc permettent d’extraire les contextes d’emploi de formes (mots, parties de mots et séquences de mots).

Nous allons prendre comme premiers exemples d'utilisation, des requêtes que nous aurions souhaité faire pour la description sémantique des noms de temps évoquée plus haut (Bras 2005). Nous nous intéressions à la série de noms de temps suivante :

ora (heure), orada («heurée»), jorn (jour), jornada (journée), setmana (semaine), setmanada («semainée»), mes (mois), mesada («moisée»), an, annada (année), matin (matin), matinada (matinée)

La requête présentée en Figure 2 permet d'obtenir les concordances des occurrences du nom *mesada*, en mode « recherche simple ». La base n'étant pas lemmatisée, si l'on souhaite obtenir les occurrences des formes fléchies d'un lexème, il est nécessaire de passer en mode « recherche avancée » et d'utiliser une expression régulière. Par exemple, la requête présentée en Figure 3 permet d'obtenir toutes les occurrences du nom *orada* au singulier et au pluriel.

C'est précisément l'utilisation des expressions régulières qui nous a permis de réaliser à partir des données de BaTelÒc deux études en sémantique temporelle sur les temps simples et composés du passé et du futur en occitan languedocien (Bras et Sibille 2020, 2021). Dans ces travaux, nous cherchions à décrire les valeurs du passé composé et du passé simple (1) ainsi que celles du futur périphrastique et du futur simple (2) :

(1) *L'ostal, l'ai pas mai, l'***ai vendut.** *Lo* **vendèri** *l'an passat.*
La maison, je ne l'ai plus, je l'ai vendue. Je la vendis l'an dernier.

(2) *Vèni pas,* **vau tornar** *a l'ostal.* **Vendrai** *deman.*
Je ne viens pas, je vais rentrer à la maison. Je viendrai demain.

Figure 2. Exemple de requête dans BaTelÒc en mode « recherche simple »

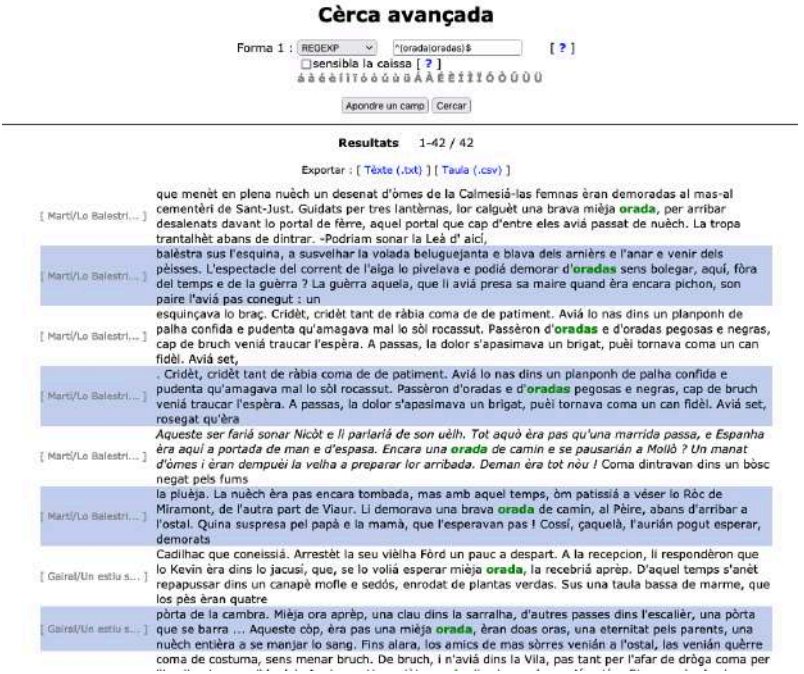


Figure 3. Exemple de requête dans BaTelÒc en mode « recherche avancée »

Les formes fléchies d'un verbe donné à un temps verbal donné sont facilement récupérables avec des expressions régulières décrivant la structure des formes recherchées, potentiellement dans une séquence de formes (Tableau 1). Avec ce type de requête, nous avons pu comparer, pour une liste déterminée de verbes, la fréquence des emplois de chaque verbe au Futur Simple et au Futur Périphrastique.

La possibilité de construire un corpus de travail propre à chaque recherche permet de créer des sous-corpus en fonction du nom de l'auteur ou de sa date de naissance, ouvrant la voie à des hypothèses sur le mode de transmission dont celui-ci a pu bénéficier : par exemple, la sélection de corpus présentée en Figure 4 permet de réunir 59 textes d'auteurs nés avant 1940 pour lesquels on peut émettre l'hypothèse qu'il s'agit de locuteurs natifs. Il est aussi

possible d'utiliser ce type de partition du corpus pour comparer les emplois de certaines expressions linguistiques en fonction des générations d'auteurs, ou de comparer les usages d'un auteur à l'autre. Nous avons pu constater, par exemple, que Joan Bodon employait moins le futur périphrastique que les autres auteurs du corpus languedocien écrit en graphie classique (Bras et Sibille à par.).

Pour obtenir toutes les formes fléchies du verbe *cantar* (chanter)	Expression régulière F1 : forme 1 F2 : forme 2
Au Futur Simple	F1:^(cantarai|cantaràs|cantarà|cantarem|cantaretz|cantaràn|cantarèi)$
Au Futur Périphrastique	F1:^(vau|vas|va|vai|anam|anatz|van|vam|vatz|anem)$ F2: cantar
Au Passé Simple	F1:^(cantèri|cantère|cantères|cantèt|cantèrem|cantèretz|cantèron)$
Au Passé Composé	F1:^(ai|as|a|avèm|avètz|an|èi)$ F2:(cantat|cantada|cantats|cantadas)

Tableau 1. Exemples d'expressions régulières permettant de ramener toutes les formes fléchies d'un verbe

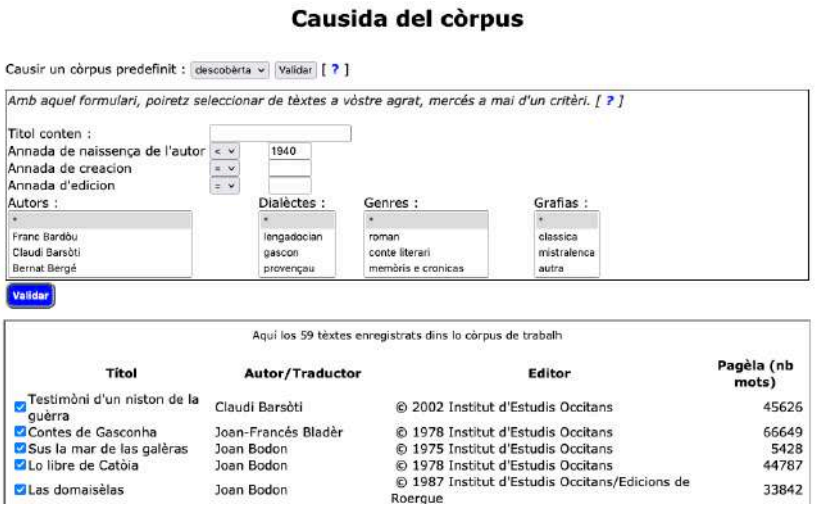


Figure 4. Sélection d'un corpus d'auteurs nés avant 1940 dans BaTelÒc

Il est aussi possible de généraliser les requêtes du tableau 1 pour extraire toutes les formes des verbes réguliers d'un même groupe pour un temps verbal donné (Tableau 2).

Pour obtenir toutes les formes fléchies de tous les verbes du premier groupe	Expression régulière F1 : forme 1 F2 : forme 2
Au Futur Simple	F1:(arai|aràs|arà|arem|aretz|aràn|arèi)$
Au Futur Périphrastique	F1:^(vau|vas|va|vai|anam|anatz|van|vam|vatz|anem)$ F2:ar$
Au Passé Simple	F1:(èri|ère|ères|èt|èrem|èretz|èron)$
Au Passé Composé	F1:^(ai|as|a|avèm|avètz|an|èi)$ F2:(at|ada|ats|adas)$

Tableau 2. Exemples d'expressions régulières permettant de ramener toutes les formes fléchies des verbes du premier groupe

Mais la limite de l'emploi des expressions régulières est vite atteinte, en particulier à cause de la lourdeur des recherches et l'obtention de formes non pertinentes (bruits). La possibilité d'interroger la base à partir d'un lemme sans avoir à indiquer toutes ses formes fléchies, et d'avoir également accès aux catégories grammaticales (ou parties du discours) nous est rapidement apparue comme l'étape suivante à franchir.

4. Deuxième étape : construction d'outils et de ressources pour une linguistique occitane outillée (2016-2022)

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 amenés à créer des outils et des

ressources pour le traitement automatique de l'occitan, en développant une chaine de traitement automatique de l'occitan[15].

En construisant BaTelÒc, nous avions en réalité réalisé les deux premières étapes de la chaine de traitement avec la collecte de textes et la segmentation automatique des textes en mots et en phrases, grâce au segmenteur permettant l'obtention de corpus segmentés, dits « nus » parce que ne contenant encore aucune annotation (Figure 5).

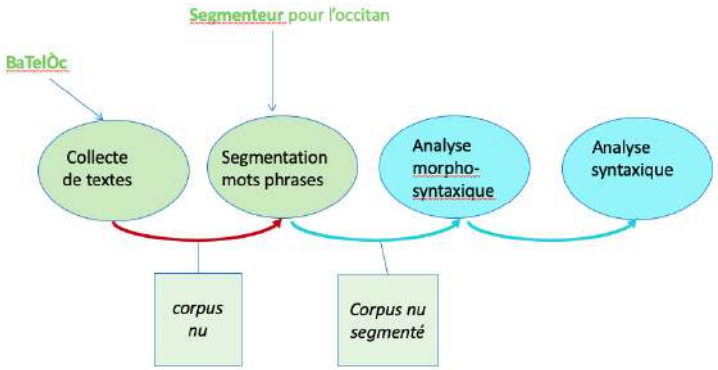


Figure 5. Premières étapes de la chaine de traitement de l'occitan

L'étape suivante a été celle de l'analyse morpho-syntaxique (voir Figure 6). Elle a été réalisée dans le cadre du projet ANR RESTAURE (2016-2018), où nous avons mis en commun notre problématique d'outillage d'une langue peu dotée avec des équipes travaillant sur l'alsacien et le picard (Bernhard et al. 2021). Le travail a d'abord consisté en la création d'un lexique des formes fléchies, réalisé en partenariat avec le Congrès Permanent de la Lenga Occitana (Bras et al. 2020) et en la réalisation d'un premier modèle pour le module d'analyse morpho-syntaxique de la boite à outils générique *Talismane* conçue par Assaf Urieli (2013) pour le français et l'anglais. Ce premier analyseur morpho-syntaxique de l'occitan (Vergez-

[15] Nous renvoyons le lecteur à (Bras et Vergez-Couret, à par. b) pour une description plus détaillée de cette chaine de traitement.

Couret et Urieli 2014), obtenu par entrainements successifs à partir de corpus annotés manuellement, nous a permis de mettre à disposition de la communauté le premier corpus occitan annoté en parties du discours (Bernhard et al. 2018, Miletic et al. 2019a) [16].

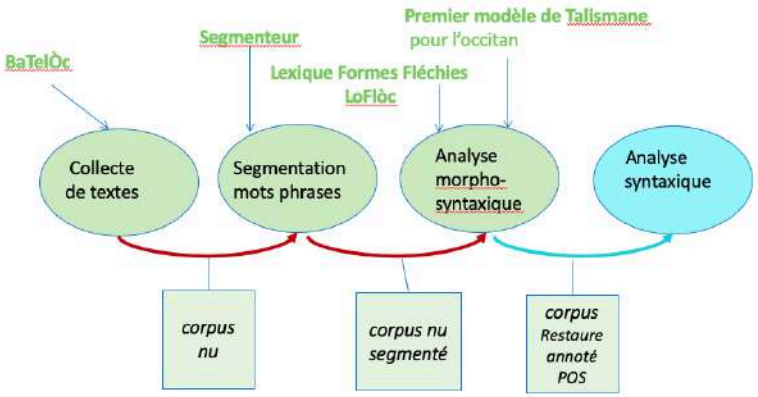


Figure 6. Etape de l'analyse morpho-syntaxique dans la chaine de traitement de l'occitan

Enfin, l'étape de l'analyse syntaxique (voir Figure 7) a été réalisée dans le cadre d'un projet européen transfrontalier, le projet interreg LINGUATEC (2018-2021) financé par le fonds POCTEFA, où nous avons pu collaborer avec des équipes travaillant sur le basque et l'aragonais. Nous avons réalisé un deuxième modèle de la boite à outils générique *Talismane*, cette fois pour les modules d'analyse morpho-syntaxique et syntaxique, et produit le premier corpus annoté en parties du discours et en dépendances syntaxiques pour l'occitan (Miletic et al. 2020 a,b)[17].

[16] Corpus de 12 000 mots partitionné en dialectes (4200 mots de languedocien, 4200 mots de gascon, 1800 mots de provençal, 600 mots d'auvergnat, de limousin et de vivaro-alpin).
Disponible sur https://zenodo.org/record/1182949

[17] Corpus de 25 000 mots partitionné en dialectes (20000 mots de languedocien, 5000 mots de gascon, provençal, limousin).
Disponible sur https://zenodo.org/record/3708268

Nous avons choisi d'adopter les formats définis par le projet international Universal Dependencies (Nivre et al. 2020)[18] pour les étiquettes de Parties du Discours (POS) et celles des Dépendances syntaxiques (DEP). Ce projet réunit une large communauté scientifique autour des corpus annotés en POS et DEP pour une grande variété de langues, ce qui présente l'avantage de bénéficier d'avancées déjà réalisées pour de nombreuses langues. Nous avons ainsi pu utiliser des corpus d'autres langues romanes pour entrainer notre analyseur syntaxique (Miletic et al. 2019b). Pour la suite, l'utilisation de ce format permettra à nos corpus pour l'occitan d'être comparés avec les corpus d'autres langues du monde.

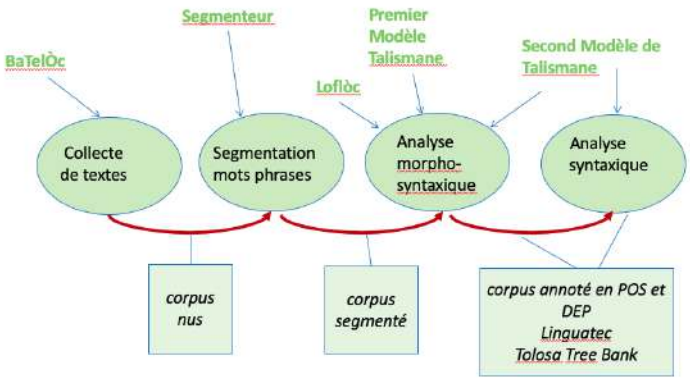


Figure 7. Etape de l'analyse syntaxique dans la chaine de traitement de l'occitan

5. Avancées pour la linguistique occitane

5.1. Bilan

Nous avons dressé dans cet article un bilan des avancées réalisées sur la période 2005-2022 en matière de ressources et outils pour la linguistique occitane. Elles ont été rendues possibles grâce au travail d'une équipe de linguistes, dialectologues et de linguistes informaticiens constituée progressivement – Marianne Vergez

[18] https://universaldependencies.org/

Couret, Jean Thomas, Jean Sibille, Louise Esher, Aleksandra Miletic, Clamença Poujade[19].

Nous avons montré comment l'absence de données textuelles facilement accessibles pour la linguistique descriptive de l'occitan avait motivé la création d'une base textuelle pour cette langue, sur le modèle de la base Frantext. Nous avons décrit la première phase de ce parcours de la création à la mise en ligne de la base BaTelÒc entre 2006 et 2016 et illustré son utilisation pour des études de sémantique temporelle. Nous avons ensuite détaillé la deuxième phase, qui s'est déroulée de 2016 à 2022, à savoir les maillons de la chaine de traitement automatique de l'occitan – segmenteur des textes en phrases et mots puis lexique de formes fléchies, analyseurs morpho-syntaxique et syntaxique – ainsi que les premiers corpus occitans annotés en parties du discours et en dépendances syntaxiques.

La création de ces ressources et outils a suivi les trois grands principes vertueux mis au jour par Soria et al. (2013) pour les langues peu dotées :

1) **Coopérer** : outre la coopération interne entre linguistes, dialectologues et linguistes informaticiens évoquée plus haut, la création des outils et ressources pour l'occitan a bénéficié de coopérations et de collaborations 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, 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. Enfin, la confiance et

[19] Joyeuse et talentueuse équipe que je remercie ici très chaleureusement.

le soutien des éditeurs occitans qui ont accepté de nous confier leurs textes ont été et restent déterminants pour l'existence de BaTelÒc et la poursuite du projet dans son ensemble.

2) **Ré-utiliser :** nous avons montré au fil des sections 2 et 4 comment nous avons pu bénéficier du travail fait pour d'autres langues pour construire les outils et les ressources pour l'occitan : Frantext nous a servi de modèle pour BaTelÒc, Talismane a été utilisé pour l'analyse automatique en POS et DEP. Sur un plan plus général, nous avons utilisé des formats standards déjà mis au point pour les langues mieux dotées (TEIP5 pour l'encodage des textes, formats GRACE puis UD pour l'annotation) et nous avons bénéficié de l'avènement des systèmes par apprentissage automatique ces dernières années permettant de créer facilement des analyseurs automatiques à partir de corpus annotés. Enfin, nous avons pu bénéficier de l'existence de corpus annotés dans des langues romanes mieux dotées (catalan, français, italien, portugais) pour initier l'annotation des corpus occitans. En conclusion, nous pouvons dire que le retard de l'outillage de la linguistique occitane a été finalement un atout pour compenser nos moyens modestes.

3) **Diffuser :** comme nous l'avons indiqué tout au long de l'article, les ressources construites sont largement diffusées. BaTelÒc est en accès libre (voir note 14) et les corpus sont diffusés via la plateforme zenodo (voir notes 16 et 17).

Les avancées réalisées selon ces trois principes sont significatives : elles ont permis à la linguistique occitane de passer du stade d'une linguistique « manuelle » à celui d'une linguistique « outillée sur corpus » avec, d'une part une quantité de données et un accès aux données décuplés, et, d'autre part, la possibilité pour chaque chercheur de construire son corpus de textes au sein d'une base de textes respectant la variété de l'occitan de l'occitan (graphies, dialectes, genres).

Les études en linguistique occitane voient ainsi s'accroitre les données pour des analyses en morphologie et en syntaxe, mais aussi,

grâce aux données textuelles de BaTelÒc, pour des analyses sémantiques en discours.

Les corpus occitans annotés seront bientôt accessibles sur le site de Universal Dependencies[20]. Quelques textes occitans ont déjà été intégrés dans le corpus multilingue parallèle ParCoLab[21], réunissant des textes en serbe, français, anglais, espagnol dans le cadre du projet ParCoLaf[22] soutenu par la DGLFLF (voir Figure 8). Au-delà de la linguistique occitane, les ressources construites sont donc maintenant disponibles pour le traitement automatique de l'occitan, pour le traitement automatique multilingue, pour la linguistique contrastive et pour la linguistique romane.

Figure 8. Exemple de requête dans ParCoLab

Dans la dynamique des premières étapes décrites dans cet article, une autre ressource a été créée récemment : il s'agit d'un corpus de contes occitans incluant des textes oraux transcrits, annotés en POS et relations sémantiques pour l'analyse de la structure narrative, le

[20] https://universaldependencies.org/#possible-future-extensions
[21] http://parcolab.univ-tlse2.fr/
[22] http://parcolab.univ-tlse2.fr/parcolaf/

corpus Corpus OcOR (Caruthers et Vergez-Couret 2018, 2021). Ce corpus vient compléter très utilement les données décrites plus haut pour les études en sémantique discursive à la fois parce qu'il permet de travailler sur des textes oraux, pas encore présents dans BaTelÒc et parce qu'il contient un nouveau type d'annotations, au niveau discursif.

Par ailleurs, le Congrès Permanent de la lenga occitana publie depuis sa création en 2011 des ressources pour les locuteurs et les apprenants de l'occitan – dictionnaires en ligne, conjugueurs, lexiques techniques, etc. Il a été notre partenaire dans le projet ANR RESTAURE et le projet DGLFLF ROLF qui ont permis de développer des claviers prédictifs en occitan et en alsacien, en s'appuyant sur les ressources lexicales comme Loflòc. Dans le cadre du projet LINGUATEC, le Congrès Permanent de la lenga occitana s'est lancé dans le développement d'applications de synthèse et de reconnaissance vocales, ainsi que dans des outils de traduction automatique, faisant ainsi entrer un peu plus l'occitan dans le monde numérique.

5.2. Perspectives

Le chemin ne s'arrête pas là, et se profile maintenant la troisième étape dans laquelle nous projetons d'améliorer encore l'accès aux données de BaTelÒc, en annotant automatiquement une partie des textes de la base en parties du discours, lemmes et en dépendances syntaxiques. Les outils d'annotation déjà construits sont relativement robustes puisqu'ils ont pu être utilisés sur plusieurs dialectes de l'occitan utilisant la graphie classique. Des travaux sur la prise en compte de la variation graphique sont en cours (Poujade, en préparation).

Les deux bases textuelles BaTelÒc et ParCoLab seront enrichies en nouveaux textes dans le cadre du nouveau projet ANR DiViTal[23] dans lequel un groupe de chercheurs du laboratoire CLLE[24] coopère

[23] Projet DiViTal (Accroître la vitalité et la visibilité numérique des langues de France, descriptions linguistiques et corpus annotés), ANR-21-CE27-0004, 2022-2026 : https://divital.gitpages.huma-num.fr/fr/

[24] Il s'agit de chercheuses et de chercheurs du groupe thématique OCRE (Occitan, langues Romanes, langues d'Europe : décrire, outiller, formaliser, comparer) au sein de l'équipe Langues et langage de CLLE : https://clle.univ-

avec des chercheurs travaillant sur l'alsacien, le corse et le poitevin-saintongeais.

Nous visons ainsi un enrichissement quantitatif de BaTelÒc en augmentant le nombre de textes et un enrichissement qualitatif à la fois en équilibrant la répartition en genres textuels, graphies et dialectes, et en enrichissant les textes avec des informations linguistiques, pour finir d'accomplir le programme de recherche défini dans (Bras 2006).

BIBLIOGRAPHIE

BEC Pierre (1970). *Manuel pratique de philologie romane*. Vol. 1. Paris : Picard.

BEC Pierre (1995). *La langue occitane (6ème édition)*. Paris : PUF.

BERNHARD Delphine & Anne-Laure LIGOZAT & Fanny MARTIN & Myriam BRAS & Pierre MAGISTRY & Marianne VERGEZ-COURET & Lucie STEIBLE & Pascale ERHART & Nabil HATHOUT & Daniel HUCK & Christophe REY & Philippe REYNES & Sophie ROSSET & Jean SIBILLE & Thomas LAVERGNE (2018). "*Corpora with Part-of-Speech Annotations for Three Regional Languages of France: Alsatian, Occitan and Picard", Language Resources and Evaluation Conference LREC 2018, 7-12 mai 2018,* Miyazaki, Japon.

BERNHARD Delphine & Anne-Laure LIGOZAT & Myriam BRAS & Fanny MARTIN & Marianne VERGEZ-COURET & Pascale ERHART & Jean SIBILLE & Amalia TODIRASCU & Philippe BOULA DE MAREÜIL & Daniel HUCK (2021). "Collecting and annotating corpora for three under-resourced languages of France: Methodological issues", Language Documentation & Conservation, 15, 316-357.

BRAS Myriam (2005). « A propos de quelques noms de temps en occitan », in I. Choi-Jonin, M. Bras, Myriam Rouquier, A. Dagnac (eds.) *Questions de classification en linguistique. Mélanges offerts au Professeur Christian Molinier*, Berne : Peter Lang. 5-80.

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

BRAS Myriam (2018). « Tractament automatic de l'occitan : qualques piadas en abans », Obrador de Linguistica Occitana, Pau, Juillet 2018.

BRAS Myriam & Jean THOMAS (2011). « BaTelÒc : cap a una basa

tlse2.fr/accueil/equipes-de-recherche/sciences-du-langage/occitan-langues-romanes-langues-deurope-decrire-formaliser-outiller-comparer

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 Myriam & Marianne VERGEZ-COURET (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, 133-149.

BRAS Myriam & Marianne VERGEZ-COURET (à par. a). « Traitement automatique de l'occitan », in Louise ESHER & Jean SIBILLE (eds.) *Manuel de linguistique occitane.* Série Manuals of Romance Linguistics. De Gruyter.

BRAS Myriam & Marianne VERGEZ-COURET (à par. b). « Ressources pour l'occitan », in Louise ESHER & Jean SIBILLE (eds.) *Manuel de linguistique occitane.* Série Manuals of Romance Linguistics. De Gruyter.

BRAS Myriam & Marianne VERGEZ-COURET & Nabil HATHOUT & Jean SIBILLE & Aure SEGUIER & Benaset DAZEAS (2020). « Loflòc : 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. 141-15.

BRAS Myriam & Jean SIBILLE (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. 157-168.

BRAS Myriam & Jean SIBILLE (2021). "Preterit and perfect in Romance: new insights from Occitan", in Louis de Saussure and Laura Baranzini (eds.) Aspects of Tenses, Modality and Evidentiality, Leiden/Boston: Koninklijke Brill NV. 136-161.

BRAS Myriam, SIBILLE, J. (à par.). « Aquel temps, uèi, s'es escapat », in Joëlle Ginestet & Jean-François Courouau (eds.) Actes du colloque *Relire Joan Bodon,* Toulouse.

BRAS Myriam & Dejan STOSIC & Marianne VERGEZ-COURET & Delphine BERNHARD & Aleksandra MILETIC & Jean SIBILLE (2021). « Outiller les

langues régionales : expériences coopératives sur l'occitan et l'alsacien avec l'aide du français, de l'allemand, du serbe, du catalan... », Lettre de l'Institut des Sciences Humaines et Sociales du CNRS, n°69. 20-22. janvier 2021.

CARRUTHERS Janice & Marianne VERGEZ-COURET (2018.) « Méthodologie pour la constitution d'un corpus comparatif de narration orale en Occitan : objectifs, défis, solutions ». *Corpus*, 18.

CARRUTHERS Janice & Marianne VERGEZ-COURET (2021). « Temporal Structures in Occitan and French Oral Narrative: The Role of Frames and Connectives ». Linguisticae Investigationes, 44(1). 1-36.

MILETIC Aleksandra & Delphine BERNHARD & Myriam BRAS & Anne-Laure LIGOZAT & Marianne VERGEZ-COURET (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 Aleksandra & Myriam BRAS & Louise ESHER & & Jean SIBILLE & Marianne VERGEZ-COURET (2019b). « Building a treebank for Occitan: what use for Romance UD Corpora? », In Kim Gerdes & Sylvain Kahane, (Eds.) Proceedings of the International Conference on Dependency Linguistics, SyntaxFest – Depling 2019, Paris, France.

MILETIC Aleksandra & Myriam BRAS & Marianne VERGEZ-COURET & Louise ESHER & Clamença POUJADE & Jean SIBILLE (2020a). « Building a Universal Dependencies Treebank for Occitan ». Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), 2932–2939, Marseille, 11–16 May 2020.

MILETIC Aleksandra & Myriam BRAS & Marianne VERGEZ-COURET & Louise ESHER & Clamença POUJADE & Jean SIBILLE (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, 140–149, Barcelona, Spain, December 13, 2020.

NIVRE Joakim & Marie-Catherine DE MARNEFFE & Filip GINTER & Jan HAJIČ Christopher MANNING & Pyysalo SAMPO & Sebastian SCHUSTER & Francis TYERS & Daniel ZEMAN (2020). "Universal Dependencies v2: An Evergrowing Multilingual Treebank Collection", in: *Proceedings of the 12th International Conference on Language Resources and Evaluation (LREC 2020)*, Marseille, France, 11–16 May 2020, 4034-4043.

POUJADE Clamença (en prep.). « La linguistique outillée à l'épreuve de la variation : ressources et outils pour les parlers occitans de l'Ariège. »,

thèse de doctorat de l'Université Toulouse Jean Jaurès.

SORIA Claudia & Joseph MARIANI & Carlo ZOLI (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*, 73-79.

VERGEZ-COURET Marianne & Myriam BRAS (2014). « Annotation morphosyntaxique d'un corpus de textes occitans : l'expérience de BaTelÒc », XIème Congrès de l'Association Internationale d'Etudes Occitanes, Lhèida, 16-21 juin 2014.

URIELI Assaf (2013). *Robust French syntax analysis: reconciling statistical methods and linguistic knowledge in the Talismane toolkit*, Thèse de Doctorat, Université de Toulouse II-Le Mirail.

VERGEZ-COURET Marianne & Assaf URIELI (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.

Myriam Bras
CLLE UMR 5263
Université Toulouse Jean Jaurès et CNRS
Université de Toulouse
myriam.bras@univ-tlse2.fr

Bases de données orales, phonétique et diachronie (aires transitionnelles entre occitan et francoprovençal)

———————————————

Chapitre 6

On vowel nasalization in transitional Francoprovençal *and* Occitan areas[1]

Michela Russo[1] & Jonathan R. Kasstan[2]

[1]UJML 3 & SFL CNRS/Université de Paris 8 (FR), [2]University of Westminster (UK)

Abstract

An oft-cited characteristic feature distinguishing Oïl and Francoprovençal from Occitan varieties in the Romance literature is the presence of nasal vowels in the former, and their absence in the latter. Yet dialectological material from the turn of the century clearly shows variation in Occitan forms in the large transitional space where Francoprovençal and Occitan are in contact. The presence of nasalized and non–nasalized Occitan forms implies a more complex picture than is traditionally understood, and suggests that the phonological inventory of northern Occitan may have included extensive endogenous

[1] We thank Clara Duvert, Corina Curchi and Jimmy Josseron, students in the Linguistics and Dialectology Master's program at the University Jean Moulin Lyon 3 under Michela Russo's supervision, for their collaboration in the collection of materials in 2021; Clara Duvert for her fieldwork in Jaujac (Ardèche), and Jimmy Josseron and Corina Curchi for their help in the collection of materials from the departmental archives of Haute–Loire. In particular, Clara Duvert came into contact with Marcel Coudène, an aged Occitan speaker from Jaujac, author of the unpublished and handwritten grammar *Grammaire du Patois des Vallées des Cévennes Vivaroises*. This precious manuscript, a gift to his grandchildren, forms part of an ongoing project to preserve Occitan linguistic heritage. We thank Marcel Coudène for having written it, and his family for allowing us to make use of the data here.

nasalization in the past. How widespread can we claim this endogenous nasalization to be? And what phonological evidence is there in synchrony?

In addressing these questions, this paper has two aims. Using empirical data, we first seek to compare vowel nasalization in FP and northern Occitan in order to establish the acoustic and spectral properties of nasal targets in synchrony. Second, we test the hypothesis that nasality in northern Occitan has resulted from internal development, rather than from borrowing via its sister varieties. To do so, we marshal inter-disciplinary datasets gathered using methods from several fields in linguistics (acoustic phonetics, phonology, sociolinguistics, historical linguistics, dialectology). In synchrony, these datasets include: speech samples taken from sociolinguistic interviews among FP speakers in the Lyonnais mountains; recent radio and interview recordings from northern Occitan speakers in the wider Auvergne–Rhône–Alpes, Haute–Loire, Ardèche, and Drôme regions; dialectological surveys undertaken in Ardèche; and archival sound files taken from the open-access platform CoCoON (*Collections de corpus oraux numériques*).[2] We further triangulate these datasets using linguistic-atlas data dating back to the turn of the century, as well as historical northern Occitan texts (written in the Haute–Loire) that also offer dialectological evidence.

Taken together, these data suggest that northern Occitan demonstrably has nasal vowels and it has in some area also denasalized what were previously endogenous nasal vowels. Firstly, we show an acoustic/auditory analysis of nasalization in the varieties above based on spectrographic cues; secondly, we show that in the area where nasalization has disappeared, the denasalization has left residual phonological cues of nasalization. We demonstrate this in particular through a discussion of stressed /a/ +N sequences, where a velarization and rounding of /a/ can be observed. This, we argue, is a clear trace of phonological nasalization in the FP/Occitan transitional space. We then argue for other phonetics and phonological traces, including morpho-phonological alternations, whereby the nasal N feature has emerged as an affix in the expression of singular and plural paradigms.

[2] Available online: https://cocoon.huma-num.fr/

1. Introduction

This article provides the first diachronic and synchronic account of vowels nasalization in the transitional space where Francoprovençal (henceforth FP) and northern Occitan are spoken (see Figure 1). Vowel nasalization is often cited as a characteristic feature distinguishing Oïl and FP varieties on the one hand from Occitan varieties on the other (e.g. Ozawa 2007 and references therein).[3] However, in the northern Auvergne Occitan varieties (including sub-varieties such as Vivaro–Alpine Occitan, or *Vivarais* as spoken in Ardèche, Drôme and Isère), nasality is attested (see Nauton 1974: 63–71 who describes vowel nasalization as widespread).[4] Indeed, albeit in Central Occitan we do find oral vowels, but this is not the case in northern Occitan of transitional zones, such as in *Auvergne*, *Ardèche*, *Drôme* or *Isère* where Franco–Provençal and Occitan are in contact with one another. This vowel nasalization is first of all attested in the data available in 20th century linguistic atlases (see for instance ALMC [= *Atlas Linguistique et Ethnographique du Massif Central*], map 1665 'Les Quatre–temps/Ember days', map 1671 '(la) Toussaint/All Saints'; ALAL [= *Atlas linguistique et ethnographique de l'Auvergne et du Limousin*] map n° 29 'le vent/ the wind' pt 42 [lu vĩ], map 586 'front/forehead' pt 60 [frũ], ALF [= *Atlas linguistique de la France*] map 211 'cent/hundred', map n° 232 'chant/song', map n° 289 'cinq/five', map n° 509 'ils sont/they are', 648 '(des) glands/acorns', 1038 'plomb/lead', map n° 721 'à jeun/fasting' map 1060 'pons/bridge', map 1187 'sang/blood', 1292 'temps/time', map n° 1334 'tronc/trunk', 1353 'van').[5]

[3] See also Rostaing (1951), Straka (1955; 1979), Bourger (1964), Bouvier (1966), Ronjat (1930–1941), Gardette (1941), Lafont (1983, 1991), Bec (1995), Kristol (2016). See for instance Lafont (1991: 6): The Oc domain « does not have the problems of vowel nasalization that affect French and Portuguese ». In Occitan « the nasality of the vowel is only partial and always followed by a consonantal resonance » Bec (1995: 23) [our translation], see also Ozawa (2007: 393).

[4] See Nauton's claim for northernOccitan of the Haute–Loire (1974: 63): « [*scil*. Nasalization conditions] They apply to my whole field ».

[5] We note that this is not simply the case for northern Occitan, see ALG [=*Atlas linguistique et ethnographique de la Gascogne*] map 975 *pain* 'bread', 322 *grain* 'grain', 330 *foin* 'hay'. In addition, Dauzat (1897) has also remarked that, in closed internal syllables, « N falls, nasalizing the preceding vowel: CANTARE (in the Middle Ages, *chantar*) *tsãta*, SANCTUM *sẽ*, [...], punctum *pwẽ*, etc. » (p.50, our translation) « in some points in the

Figure 1 : FP and northern Occitan varieties in Auvergne–Rhône–Alpes (Wikimedia Commons)

The literature therefore attests to a more complex, transitional phenomenon than is traditionally appreciated, whereby oral vowels are most typically attested in central Occitan, though not necessarily in the northern Occitan varieties to be found in the transitional Oïl/FP/Oc space. However, there is as yet no acoustic or granular phonological description of the varieties spoken in this space, beyond the available impressionistic evidence. In the wider literature on sound change, particularly of under-studied language varieties, sub-

southern area the consonant [...] is maintained after nasalizing the vowel, but it remains rare in this northern Occitan speech [Bellac, Haute–Vienne]» (Lagueunière 1983: 102, our translation).

phonemic variation and acoustic analyses of gradient phonetic quality is acknowledged to remain in its infancy (e.g. Babel 2008).

In contributing to this wider literature on sound change of understudied varieties, we first seek to compare vowel nasalization in FP and northern Occitan in order to establish the spectral properties of nasal targets in synchrony. Second, in demonstrating abundant vowel nasalization, we will test the hypothesis that nasality in northern Occitan has resulted from internal development, rather than from borrowing via its sister varieties. In particular, we will show that indirect traces of nasalization can be observed in historical texts particular to Auvergne Occitan, which suggests (a) that nasality in Occitan must have been more extensive in the past, and (b) that nasal vowels may have undergone denasalization. However, phonological cues such as velarization and the rounding of stressed /a/ in V ([ɑ ɒ o ow /ɑ̃ ɒ̃ õ]) +N sequences, and parallel outcomes of /e i o u/ +N [e/i/ej]/[o/u/ow] + /N/ betray traces of nasality in modern northern Occitan varieties.

To address the paper's aims, and to document these phonological cues, we marshal disparate datasets from a number of paradigms in linguistics. In synchrony, these datasets include: spontaneous speech samples gathered from sociolinguistic interviews among FP speakers in the Lyonnais mountains; recent radio recordings from departmental archives of Haute–Loire and interview recordings from northern Occitan speakers in the wider Auvergne–Rhône–Alpes region, Haute–Loire, Ardèche, and Drôme regions; spontaneous speech samples gathered in Jaujac ; and archival sound files taken from the open-access platform CoCoON (*Collections de Corpus Oraux Numériques*). These materials include recordings of Occitan gathered from the Protestant area of Chambon–sur–Lignon (*commune* of Tence) by the dialectologist Théodore De Félice;[6] these recordings are also supplemented with local radio broadcasts from Haute–Loire and Ardèche.

We present a distributional and acoustic/spectral analysis of nasalization in recordings of semi–structured sociolinguistic interviews and wordlist elicitation tasks gathered in Lyonnais

[6] For a survey of this Protestant area see De Félice (1983, 1989); Martin (1997); Nauton (1974).

villages, as well as in recordings of northern Occitan (both wordlist elicitation tasks and spontaneous speech).

We then triangulate these datasets using linguistic-atlas data dating back to the turn of the century, as well as historical texts that further attest to the dialectological record (including one hand-written Occitan grammar).

In all, our inter-disciplinary datasets (which have been gathered by combining acoustic phonetic analysis with, phonological, sociolinguistic, historical linguistic, and dialectological methods) provide us with over 100 years of time depth.[7] Triangulating the data in this way presents some evidence for relative stability in the nasal vowel system.[8] More significantly, the data suggest the existence of nasal vowels in the northern Occitan area, both in diachrony and synchrony.[9] On a diachronic level, we will show that nasality is evidenced through phonological properties in the backing/rounding of the low vowel (velarized [ɑ/ɒ]); raising or lowering of mid–vowels ([ve/vi] 'wine' Lt VINU, [mezu] 'maison/house' LT MANSIONE); and the glidification of the nasal element /N/ ([bostɔw] 'batôn/stick' Lt –ONE). Furthermore, we show that nasality is deployed morpho–phonologically to express plurality, which distinguishes to a certain extent northern and central Occitan varieties, given the tendency in the former to not have sigmatic plurals. Our results are consistent with what is attested in the linguistic atlases, in that they show nasalization in the northern Occitan area of Haute–Loire, Puy–de–Dôme, Ardèche, Drôme and South of Isère. These same atlases show nasalization beyond the area examined in this article, including in the neighboring areas of the Creuse du Cantal, in the Haute–Vienne

[7] *Atlas linguistique de la France* [= ALF] Gilliéron & Edmont (1902–1910); ALLy Gardette (1950–1956); *Atlas Linguistique et Ethnographique du Massif Central* [= ALMC] Nauton (1957–1963); *Atlas linguistique et ethnographique du Jura et des Alpes du Nord* [=ALJA] Tuaillon & Martin (1971–1981).

[8] Bouvier (1976) studied the phonetic features of the *Drôme* languages; see recently, Russo, Curchi and Josseron (2021) for nasality in *Drôme*. The languages of the part of the Isère that borders the Pilat region to the east are described by Devaux (1892 and 1935); see recently Bert (2001) for the Pilat region.

[9] We refer specifically to transitional Occitan–speaking zones, between Auvergnat, Occitan *Vellave*, High–*Vivarais* (expanding westwards beyond the region called *Vivarais* into the Auvergne region), Low–*Vivarais*, and Vivaro–Alpin dialects.

(Limoge), in the South of the Corrèze.[10]

Owing to the numerous datasets that we draw upon in this paper, it is first necessary to offer some initial commentary on the research design of the present study before outlining vowel nasalization in FP and Occitan.

In the following sections, we provide an acoustic/auditory analysis of vowel nasalization in FP and in northern Occitan, focusing on 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.

Data collection involved both structured elicitation tasks (e.g. wordlists) and spontaneous speech (see details below). To facilitate acoustic and auditory analysis of V+N sequences, spectrogram analysis took place in PRAAT (version 6.1.40, Boersma & Weenink 2021)[11].

In order to triangulate our observations, we draw on a number of different historical materials, including observations from the following linguistic atlases: ALAL; ALF; ALG; ALJA; ALLy; ALMC; ATF; and lexicographic sources such as FEW. We also exploit historical Occitan Auvergnat written texts (available in recent new editions) and a hand-written Occitan grammar, the *Grammaire du Patois des Vallées des Cévennes Vivaroises*, written by a local Occitan speaker (Marcel Coudène) hitherto unknown and unpublished.

2. Vowel nasalization in FP and transitional FP/Oc areas

2.1 Vowel nasalization in FP

To describe vowel nasalization in FP, evidence is drawn from the

[10] For the ALF, see also Ozawa (2007). However, we do not share Ozawas's and Chambon (2004) hypothesis, that this is the result of a Frenchification of Occitan. The languages of the part of the Isère that borders the Pilat region to the east are described by the work of Devaux (1892 and 1935); see Bert (2001) for the Pilat region.

[11] Spectrogram settings were kept to a frequency view range of 5000 Hz, with a dynamic range of 40-50.0 dB, suitable for measuring formant values at F1, F2, and F3.

Variation and Change in Francoprovençal Corpus (VCF) (Kasstan 2015a; 2022), which is made up of 100+ hours of speech samples gathered from 74 speakers of FP in France and Switzerland. While data in this corpus were gathered from different speaker types (including 'older', 'younger', and 'new' speakers, for details see Kasstan 2019), only 'older' speakers are considered here (4f, 5m). These participants belong to an inter-war generation who acquired FP as an L1: they are retired rural dwellers. Data collection for this sample took place in the Lyonnais mountains, in fieldwork sites that broadly overlap with ALF and ALLy (=*Atlas linguistique et ethnographique du Lyonnais*) datapoints for this region (with a focus here on the *communes* of Saint–Martin–en–Haut and Saint–Symphorien–sur–Coise).

We focus on example tokens from the former here, and only instances of /i/ +N are considered, yielding n=100 tokens (see Appendix 1 for a list of lexical items extracted from a larger reading list).

Nasalization in FP applies to all contexts where Metropolitan French has nasal vowels. However, FP has a comparatively more complex vocalic inventory, including the high nasal vowels /ũ/ and /ĩ/ (cf. Bourger 1964, Bouvier 1966, Gardette 1941, Escoffier 1958a/b, Martin 1979, 1990; 1993; 1997, Tuaillon 1972; 2007, Bert 2001, Kasstan 2015b, Kristol 2016, Russo et al. 2021, Chong & Kasstan 2022). Historical atlas data attest to nasalized vowels in FP as being made up of an underlying high vowel + nasal coda sequences (see e.g. ALLy map 837 *chemin* 'path', below). It is also the case that the dialectological record attests to be a significant amount of variation in the quality of nasal szation, where [ĩ] and [ɛ̃] are frequently found to co-vary for datapoints in very close proximity (Figure 2):

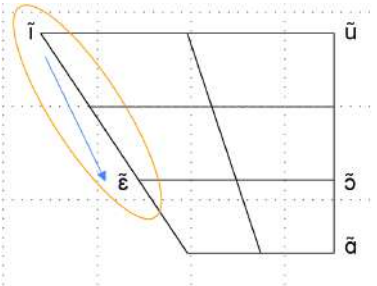


Figure 2 ALLy space of /I + N/ from Maps n° 438, 923, 835, 837

This is the case for example with ALLy maps 438 *un pin/des pins* 'pine/pines' and 837 *chemin* 'pathway', in which this oscillation between both [ĩ] and [ɛ̃] is visible for items containing a nasal /i/ +N cluster (cf. Figures 3 and 4, below)[12].

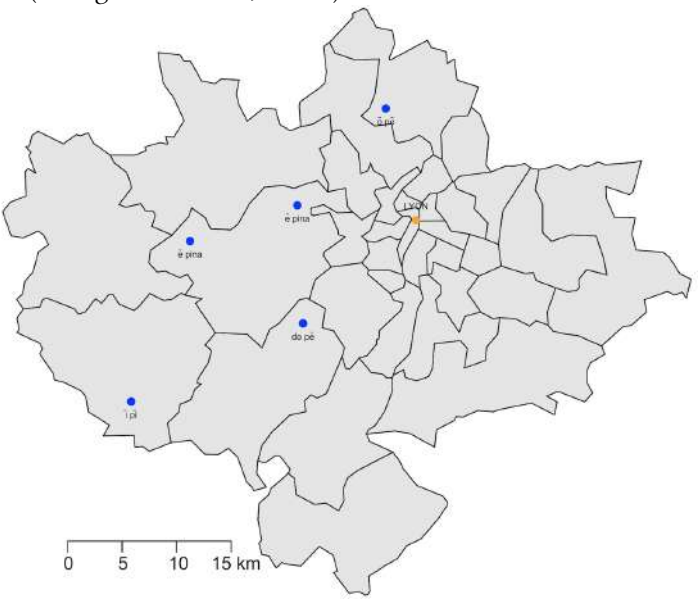


Figure 3 : ALLy map 438 *un pin/des pins* 'a pin/pins' – /ĩ/ ~ /ɛ̃/ (*Monts du Lyonnais*)

[12] ALLy map 438 'un pin/des pins' – /ĩ/ ~ /ɛ̃/ (after Gardette 1950–1956). Forms include [õ.pẽ] [ẽ.pẽ] [ĩ.pĩ] [ĩ.pĩ]. Data point pt. 40, 39, 49, 54. ALLy map 837 'chemin' /ĩ/ ~ /ɛ̃/. Forms include 0, [ʦumẽ] [ʃamĩ] [ʃamĩ]. Data point pt. 40, 39, 49, 54.

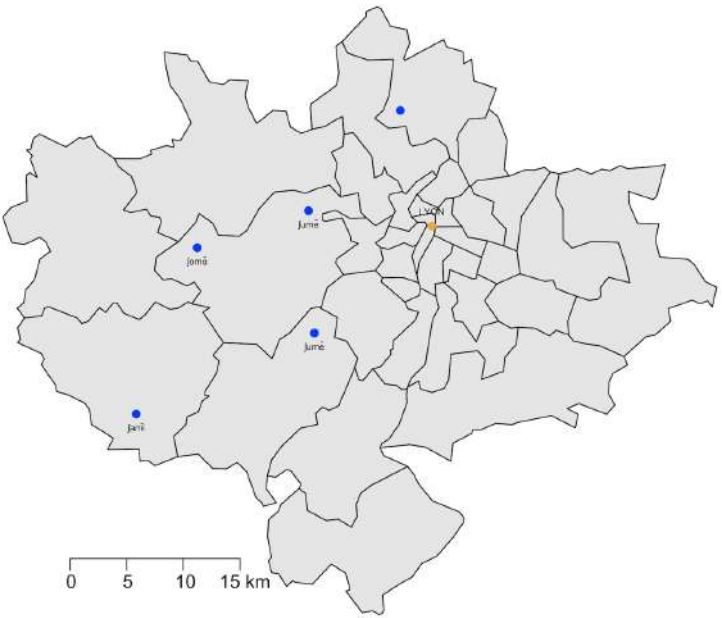


Figure 4 : ALLy Map 837 *chemin* 'pathway' – /ĩ/ ~ /ɛ̃/ (*Monts du Lyonnais*)

Much like Old French, for FP it is possible to posit two phases for the nasalization of /i/ +N: a first phase with a nasalized [ĩ] quality (which is still preserved along the FP/Oc border), followed by a second phase with a mid–open [ɛ̃] quality.[13] There is also evidence to suggest that speakers are aware of this variation: Kasstan & Russo (2020) present sociolinguistic-interview data to suggest that speakers can recall both variants, and are able to provide metalinguistic commentary on these forms when asked.

[13] Evidence for which can be gleaned from existing ALLy data (e.g., maps 438 'pin/pine', 923 'matin/morning', 835 'voie lactée' (chemin de Saint–Jacques), 837 'chemin/path'. In the Loire, the FP forms [ĩ] and [ɛ̃] are present until pt. 34 (Saint–Marcel–d'Urfé), pt. 48 (Essertines–en–Chatelneuf), pt. 55 (Sury/Loire), pt. 61 (La Valla), pt. 66 (Roizey), indicating an isogloss. At data pt. 62 (Sainte–Croix) and pt. 54 (Saint–Bonnet–les–Oules) we observe oral forms without nasalization such as [madzi] *maison* 'house' and [ʧami] *chemin* 'path'; pt. 55 (Sury) has [ʧami] as expected in Occitan.

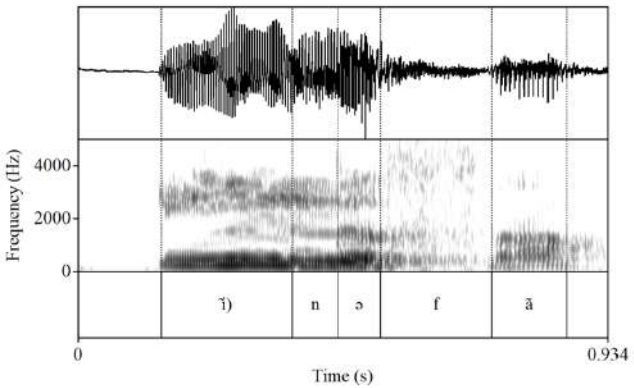
However, much like the VCF corpus as a whole, within the sub-sample of n=100 tokens analyzed here mid-open variants constitute but a small proportion of the total V+N token count (11%), suggesting that the nasalized [ĩ] is the dominant form.

Turning to the FP production data from the VCF corpus, among a sample of 9 older speakers (balanced for sex), we do observe variation in nasal and non-nasal realization, but vowel nasalization is the dominant manner of articulation (see Table 1). This observation is not statistically significant, but nonetheless presents a picture of nasalization in synchrony in the FP-speaking Lyonnais region. Turning next to the acoustic analysis, and comparing waveforms from the production data more carefully (cf. Spectrograms 1 and 2), it is clear that what we are positing here to be nasal vowels according to the dialectological record are indeed fully realized nasal vowels. While it can be difficult to give general acoustic characterizations of nasal vowels, given their complex vocal tract configuration, the segmentation at least neatly suggests a dynamic trend towards the nasal target with a broad F1 pattern.

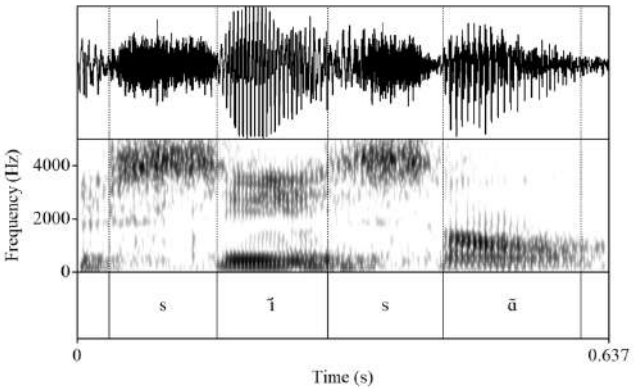
Table 1. Distribution of n=100 /i/+N tokens by speaker sex (VCF corpus)

	female	male	TOTAL
nasal	26	58	84
non-nasal	8	8	16
TOTAL	34	66	100

χ^2, two-tailed = 0.1405 (n.s.), df = 2.173



Spectrogram 1 : [ĩ)nəfã] *enfant* 'child' (*Monts du Lyonnais*)



Spectrogram 2 : [sĩsɑ̃] *cinq–cents* 'five–hundred' (*Monts du Lyonnais*)

Concerning nasalized back vowels in FP, /ũ/ can also derive from the nasalization of /o/ as can be gleaned from the ALLy maps in (1). It is also possible to establish an isogloss between FP and Occitan for /ũ/, as 4 out of 7 maps also suggest a parallel isogloss for /i/ +N (see Gardette 1973). In each case, it is possible to posit for /u, ũ/ a two-stage process of nasalization triggered by the raising of /o/ to [u].

(1) ALLy: nasal [ũ] from /o/ – maps #

121 : *timon* 'drawbar'
127 : *aiguillon* 'spur'
138 : *sep de l'araire* 'sep of the plow'
157 : *petite pioche* 'small pick'
314 : *bélier* 'ram'
678 : *maison* 'house'
1319 : *à bouchon* 'with a plug'

There is for these maps a fair amount of variation in the Occitan of Ardèche and Haute–Loire departments with nasalized forms such as [alamũ] *sep de l'araire* 'sep of the plow': while we find the Occitan non-nasalized form for *aiguillon* 'spur' at pt. 61 (La Valla/Loire), and for *maison* 'home' at pt. 62 (Sainte–Croix/Loire), there is also variation with the presence of a nasalized form [alamũ] at pt. 67 (Saint–Romain–les–Atheux/Loire), at pt. 69 (Saint–Sauveur–en–Rue/Loire), as well as in Ardèche and Haute–Loire, at pt. 71 (Vanosc/Ardèche), pt. 72 (Saint–Julien–Molhesabate/Haute–Loire), pt. 74 (La Louvesc/Ardèche[14]). Looking to the northern-most Occitan region, and particularly ALF data pts 827 (Vion, Ardèche), 838 (Saint–Nazaire–en–Royans, Drôme), 849 (Le Monestier de Clermont, Sud de l'Isère) we might suppose this presence of nasalization to be the result of long-term language contact with FP. After all, it is well-attested that the Lyonnais region (and Lyon in particular) has long been an important regional centre of innovation and diffusion (e.g. Gardette 1974; Chambon & Greub 2000). The limit of nasalization of /a/ +N can also be established using ALLy maps 419 *pain* 'bread' (Fr. [pɛ̃]); 1093 *main* 'hand' ([mɛ̃]); 1309 *demain* 'tomorrow' Fr. [dəmɛ̃]. A border separates two types of backness (velarization) with explicit nasalization in the forms [pɑ̃]/[mɑ̃]/[demɑ̃], and a backness without nasalization [po]/[mo]/[demo] (we return to these crucial variants in our analysis of Occitan nasalization below). This isogloss also tracks the ALLy maps for nasal [ĩ]. Further South, nasalized forms are still found into the very North of Ardèche (which is FP-speaking). However, in what follows we will see that nasal vowels are also widespread in Occitan-

[14] An area around *La Louvesc*, South of the Pilat region, is described by Dufaud (1986; 1998).

speaking Ardèche, particularly in the High–*Vivarais* region and in the Occitan locality of Jaujac (or Low–*Vivarais*, southern Ardèche).

To summarize, what are attested impressionistically to be nasal vowels in /i/+N clusters based on the dialectological record are indeed confirmed to be [ĩ] variants based on the available sociolinguistic and phonetic data presented above. Having sketched an account of nasalization of the high front vowels in FP using historical atlas data and more recent production data, we turn now to comparing observations of spectral properties of nasal targets to those of Occitan. We remind the reader here that word-final –n (formerly in Latin open syllables) ceased to be realized in large parts of the Occitan space (as a number of ALF maps attest). However, in what follows our analysis will demonstrate maintenance in some northern Occitan areas (e.g. in the Ardèche space where the Jaujac variety forms such as [pɑn] Lt. PANE or [mɑn] Lt. MANU can be found). This holds too for Auvergnat Occitan, which does not always nasalize synchronically: e.g. where a vowel followed by a nasal /n/ renders [i] in [vi] *vin* 'wine', for VINU in Occitan. However, below we argue that phonological traces remain of a prior stage of nasalization, where, for example, forms such as [ve] can be found to vary with other forms such as [vi].

2.2 ca + N in transitional Oc/FP areas

The FP vowel [ĩ] (also realized [ɛ̃], as above) is an outcome of diachronic palatalization and affrication of the Latin CA sequence /k+a/, which results in /i/ + N, giving rise to [i ĩ], as in (2):

(2) Diachronic palatalization/affrication of stressed Latin /KA–/ in FP {θĩ, tsĩ, ʃĩ}

ALLy map 372 *chien/chienne* 'dog' Lt. CANE(M)

[ʃɛ̃]	pt. 52	Marennes/Isère
[θɛ̃]	pt. 64	Pommier/Isère

While we do observe the Occitan non-nasalized high vowel forms in the transitional area in the North of Ardèche such as from two

ALLy points (where the FP feature /a/ > [i] after a palatal, here the coronal affricate [ts] combines with an Oc oral [i], as expected mostly from Haute–Loire results: [tsi] MSG pt. 73 (Ardoix/Ardèche) ; [tsino] FSG pt. 70 (Boulieu/Ardèche). The oral treatment of the vowel is characteristically Occitan and the nasal treatment is characteristically FP. In pt. 70 we see the Oc feature in the FSG syllable finally [–o] ([tsino]). The differences between transitional varieties where vowels are nasalized and varieties where the final vowels remain oral (as in pts 73 and 70) have not been sufficiently discussed in the literature so far, and it is a crucial issue in order to understand the transitional areas between FP and Occitan, such as in Ardèche [tsi] /[tsino], an area that belongs to northern Occitan and FP (Nauton's *amphizone*). In the North of the Drôme, Bouvier (1976: 367–369 and map 56) also indicates the nasalization of /i/ +N. Russo et al. (2021) too show nasalization in the Occitan-speaking Drôme, as in other northern Occitan varieties. Their findings in this area mainly show the nasalization of the vowels /a/ and /ɛ/ (from Latin A, Ĕ): [pɑ̃] *pain* 'bread'; [tɛ̃] *temps* 'time'. High vowels are mostly denasalized (as we see in ALLy map 372 above) and the final nasal consonant is quite regularly deleted [vi] *vin* 'wine'; [mezu] *maison* 'house'; [razi] *raisin* 'grape'. Next to the Drôme, in northern Isère, which encroaches on the traditional FP border, nasalization is more advanced. In final position, all nasal vowels are nasalized without a trace of the consonant, a situation which contrasts strongly with southern Isère.

2.3 Nasalization: morpho–phonological profile

The resulting picture is complex. Nasal and oral variants can coexist in the same transitional space and variety, which in turn impacts on phono–morphological properties and patterns. Indeed, it has been pointed out elsewhere that alternations between nasal and oral vowels can have a morpho–phonological value whereby nasality can behave as an affix (for Drôme and Isère see Bouvier 1966; Russo et al. 2021). For example, Russo et al. (2021) indicate that in Trièves (South of Isère) there is a morphological alternation between [jaˈnu] SG and [jaˈnũ] PL *genou(x)* 'knee(s)' Lt. DENUCULU, a type of alternation that Gardette (1941: 137–138) describes as not uncommon. Bert (2001: 362) describes a similar phenomenon based

on fieldwork in Riotord (Haute–Loire), where both oral and nasal endings are attested for word-final –ONE forms. For the Loire, Bert (2001: 238) also points out for *mouton* 'sheep' the alternation between SG [ũ] in [muˈtũ] and PL [u] in [muˈtu] for the variety spoken in Versanne (Pilat region). This variation in nasal production is thus indicative of a morpho-phonological functioning of vowels at work in nominal paradigms, which is also confirmed in diachrony. For example, Russo et al. (2021) cite von Wartburg (FEW 7, 95a)'s entry <nevon> as attested in Old Dauphinois (Grenoble 1338) which depicts a morphosyntactic variant corresponding to a subject case form (Latin NEPOS). Bouvier (1966: 129) further points out the Occitan form <nebon> in a 12th century document located in Valence (1160 Brunel no. 98; FEW ib.). In Occitan-speaking Drôme the form <nebon> would thus be a very old form, which would have in the past been more widespread (e.g. for Drôme see ALF forms [nəˈvõ] with a nasalized coda for pts 838, 907, and 920 *neveu* 'nephew'). Morpho–phonological alternations are also found by Bert (2001: 360) in transitional Oc/FP areas, where an [u/y] opposition can also signal an SG/PL alternation. This evidence suggests a phonological cue of nasality, since nasals trigger a raising of /o/ > [u] SG, while the plural (with a palatal element –j) triggers the fronting of [u] to [y] (e.g. in Saint–Marcel [moˈty] PL 'sheep', Davézieux [peˈʃy] 'fishes', although the converse is also possible, with a singular [y] form, and a plural [u] form.[15] This seems to us to be an innovative pattern which is furthermore consistent with the expression of non–sigmatic plurals in northern Occitan. Indeed, the Occitan-speaking Pilat region attests to the alternation between singular and plural forms in words ending in –ONE, based on the length of the final vowel [u] (/o/ +N being the underlying representation), as in Ardoix (Ardèche) [kaºju] 'pig' SG vs [kaºju:] PL 'pigs' (Lt –ONE); this strategy applies after a prior raising of /o/ before N to [u] (see Bert 2001: 360).

[15] Saint–Marcel, Thélis, le Bessat, Tarentaise, la Versanne and Davézieux (Bert 2001 : 361).

3. Nasal vowels in northern Occitan

We have already pointed out that in the northern Occitan speaking region of the Haute–Loire nasalization is present in the historical record (according to Nauton 1974: 63–71; Sumien 2009: 12).[16] Quintessentially, the indefinite article in Haute–Loire is also realized as a nasalized vowel [ũ/jũ] (Fr. [ɛ̃], Nauton 1974: 63), and according to Nauton, the production of nasals is well-advanced over a large region of Haute–Loire.[17] To familiarize the reader with the northern Occitan transitional zone (Haute–Loire, Ardèche, Drôme), we highlighting existing boundaries and contact areas with other Gallo–Romance regional languages and different Occitan varieties:

Figure 5 : northern Occitan

[16] Sumien (2009: 12) attests to nasalization in northern Occitan in which he highlights a neutralization between high nasal vowels and nasal mid–vowels: /ɛ̃/+/ẽ/+/ɪ/ > /ẽ/; /ɔ̃/+/ũ/ > /õ/: « La zòna probabla, mai que son isoglòssa l'an jamai dessenhada amb exactitud, de la nasalizacion completa (dança [ˈdãsɔ] en luòc de [ˈdansɔ, ˈdaⁿsɔ]). Conten de zònas pus reduchas ont certanei vocalas nasalas se son neutralizadas: /ɛ̃/+/ẽ/+/ɪ/ > /ẽ/; /ɔ̃/+/ũ/ > /õ/. Es lo cas a l'entorn de Rumans, dins una part dau nòrd–auvernhat e dins lo Creissent ».

[17] However, he excludes from nasalization an area along the Southern border of the department, which continues in Lozère and Ardèche (ALF pts 821, 824): from Chanaleilles, Pradelles, Goudet, Saint–Front, Mounedeyres, Boussoulet, Le Mazet–St–Voy.

Figure 6 : Departmental map of Auvergne–Rhône–Alpes: Haute–Loire, Loire, Ardèche, Drôme, Isère (© Brad Pict – Fotolia)

Before delving into the analysis of the material at our disposal, we define in the following section the Occitan varieties within the transitional space of Auvergne–Rhône–Alpes[18].

3.1 Northern Occitan language(s) in Auvergne–Rhône–Alpes

Northern Occitan varieties (which can broadly be grouped into three types *Auvergnat*, *Limousin* and *Vivaro–Alpin*) are typically distinguished from southern Occitan varieties based on the opposition between [ka]/[t͡ʃa/t͡sa] outcomes of Latin C+A

[18] See Hasselrot (1934); Gardette (1941); Tuaillon (1964); Martel (1983); Martin (1979, 1990); Bert (2001); Bouvier (2003); Sumien (2009); [FORA] Bert et al. (2009); Bert & Costa (2014), among others. Particularly, Bert (2001) conducted fieldwork on the Occitan/Franco–FP border in about twenty villages between the North of *Ardèche* and the South–East of the Loire.

palatalization (Sumien 2009, Brun–Trigaud et al. 2005).[19] This affrication of the Latin syllable C+A which results in [ʧa/tsa] is considered one of the major distinguishing features for geographical division between Languedoc Occitan and northern Occitan (see ALF 238 *chat* 'cat' and ALMC 567 *chat, deux chats, chatte* 'cat, two cats, cat (F)'. Auvergnat Occitan is a central-northern variety spoken in Auvergne and Velay which has its own specificities and it belongs instead to the northern Occitan grouping.[20] Historically, it has been subjected to the centripetal forces that have radiated from a more innovative LUGDUNUM (Lyon) on the one hand, and a more conservative southern–western Romance (Aquitaine) zone on the other.[21]

Ronjat (1930–1941) delimits the Auvergne area (referring to the Crescent route, De Tourtoulon & Bringuier 1875/2004) as comprising the area from the North to the West as far as the South of Vichy, where it butts up against the FP border. This boundary merges in the North with the department of Puy–de–Dôme and the southeast of Allier (Roux 2015; 2020). However, the borders are more complex: to the North we find the boundary established by Ronjat, and the Crescent, to the East the Occitan–FP boundary as studied by Gardette (1941). This FP/Oc boundary follows the Forez mountains,

[19] And also, for the [ga]/[ja] opposition. For other features of northern Occitan, see also Brun–Trigaud et al. (2005). Typical phonological features of northern Occitan also include: *monophtongization of [aw/ɔw] > [o], /Vs/ + C (= /p t k/) > [jC/V:C] (especially in Vivaro–Alpin). Recall also that these areas have a non–sigmatic distinction between nominal singular and plural (nouns and adjectives), the plural is expressed through vowel alternation*s: [ˈpalɔ/ˈpala:]/ (and if applicable stress shift) [ˈpalɔ/pɒˈla:] *pala* 'skin', *palas* 'skins'; *alternation between vowel and diphthong* [ˈɔme/ˈɔmej:*] òme* 'man', *òmes* 'men*', (in case of even final word accent)* [tsɒˈpe/tsɒˈpjaw, tsɒˈpe/tsɒˈpjo] *chapel* 'hat', *chapiaus* 'hats' *(*Brun–Trigaud et al. 2005; Sumien 2009*)*.

[20] Girard (1925), Dauzat (1938 ; 1944), Bec (1970; 1973), Bonnaud (1974; 2006), De Felice (1983; 1989), Chambon and Olivier (2000), Chambon (2012), Allières (2001), Roux (2015; 2020), Sumien (2006; 2009), Surrel (2022). From the 1940s-2015, the administrative region of Auvergne was composed of four departments: Allier (*Auvergnat* is spoken in the south of Bourbonnais), *Cantal*, *Haute–Loire*, *Puy–de–Dôme*. Since 2016, these former *departments* form part of the new region Auvergne–Rhône–Alpes.

[21] Bec (1970, 1973) sees Auvergnat as part of the *Aquitano–Pyrenean* macro–structure, which is also confirmed by Chambon (2000), while at an embedded step the macro–affiliation of Vivaro–Vellave (High–Vivarais) within Occitan is Averno–Limousin (see Martel 1983). Auvergnat includes the West of the Velay, the East towards Yssingeaux speaks Vivaro–Alpine. The Aurillac area is in Auvergne but is a Languedoc Occitan speaking zone (see Sumien 2009: 14).

the departments of the Loire to the East and the Puy–de–Dôme to the West (including the Noirétable area, in Loire, which is *Auvergnat* Occitan-speaking). To the South we find the Saint–Bonnet–le–Château area where the Vivaro–Alpine Occitan variety is also spoken. Besides Saint–Bonnet–le–Château, and the south-western part of the *Loire*, the following areas are equally Occitan-speaking: the Bourg–Argental area (on the Pilat Mountains) in the south-eastern part of the Loire (Vivaro–Alpine), then the area immediately North of the *Ardèche*, as well as Firminy (formerly Forez). In Firminy Occitan was spoken until the beginning of the XXth century (Noirie 2018; Martin 1979)[22]. Auvergnat Occitan is spoken in the Velay but also in the Ardèche region and in the northern part of the Lozère; the southern part of the department of Haute–Loire and the northern part of the Ardèche (*Vivarais*/Northern *Vivarois*) are considered by Martin (1979) to be varieties of Auvergne Occitan (Occitan– *Vellave*). The Occitan Vivaro–Alpine variety thus is spoken in the North of *Ardèche* (the North of the *Vivarais* is the present–day *Ardèche*)[23], in part of the Velay (Vivaro– *Vellave*), in the Loire (around Saint–Bonnet–le–Chateau), as well as in some localities of south-eastern Puy–de–Dôme.[24]

This western Vivaro–Alpine is also called Vivaro–*Dauphinois* (Bec 1973),[25] since is also spoken in southern *Dauphiné*, which includes part of the Drôme and Isère where this Occitan variety is also known as Vivaro–Valesian or *Dauphinois*. Ardèche seems to be a region where numerous Occitan varieties are present, particularly Vivaro–Alpine Occitan (the High– *Vivarois* in the North also called Occitan–Vellave), and a Languedocian variety in the southwestern

[22] Gardette (1968: 212) : an important linguistic border cuts across the Forez and Noirétable plateau, and, further South, the Saint–Bonnet–le–Château plateau and the Saint-Etienne mountains. These two regions (Noirétable, St–Bonnet and Bourg–Argental) have characteristic Occitan phonetics. See also Hasselrot (1934), Martin (1990), Escoffier (1958a/b), Russo (2021).

[23] The Vivaro–Alpine varieties of the *Ardèche* or North–*Vivarois* are found in the northern half of the *Ardèche*, they include to the South of Doux the varieties of Lamastre, Saint–Agrève, Saint–Péray, and those of Boutières, Vernoux, le Cheylard, Saint–Pierreville going down to Privas in the South (see also *Duvert*–Chenebert 2020–21; Garnier 2022).

[24] As well as in Livradois, around Ambert and Arlanc.

[25] It is for the extension to the East of the domain that Bec (1963) called it Provençal–Alpine, then Vivaro–Alpine or Dauphinois.

part of the region (Sumien 2009).[26] FP is spoken in Ardèche in northern Annonay.[27] The Vivaro–Alpine group was already called *Dauphinois* by Mistral; this designation corresponds to Ronjat's classification in his grammar. Ronjat uses the term *Group Alpin–Dauphinois*, including *Vivarais*, a group which, according to him, also covers the northeastern part of Velay, the South of Forez and the North of Ardèche. According to Martel (1983), it is subdivided into Rhodano–Dauphinois in the Drôme and Isère, and Vivaro–Vellave in northern *Vivarais*, eastern Velay and Forez.[28]

To summarize, the Vivaro–Alpine Occitan variety is spoken in some localities of Velay (Auvergne), as well as in some localities of the FP Forez (in Auvergne–Rhône–Alpes), notably in the Loire department, around Bourg–Argental (just North of Ardèche), and around Saint–Bonnet–Le Château, as well as in the southeast of the Puy–de–Dôme (around Arlanc). These geographical localities are also part of the FP/Oc border discussed in the previous section (see Gardette 1941; Tuaillon 1964; Escoffier 1958a/b; Russo 2021). In all of these areas, Vivaro–Alpine Occitan is considered Haut–*Vivarois* (which has spread beyond the *Vivarais* region of Ardèche); it is called Vivaro–Vellave especially in the Velay and in the North of Ardèche (Martin 1979; Philippe Martel 1983 *inter alia*). The Vivaro–Vellave (also known as western Vivaro–*Alpin* or *Rhône–Alpin*) is part of the Vivaro–*Alpin*.

Ronjat's definition of northern Occitan considers a dialectal 'Group E', including Alpino–*Dauphinois*, which consists of the northern *Vivarais*, the north-eastern Velay, and the south-western Forez (which is Occitan).[29] Ronjat also situates the subgroup of

[26] The Aurillacois Occitan (a Northern variety of Languedocian) is spoken instead in the South–West of the Cantal (Olivier & Chambon 2000).

[27] For the Ardèche department, we can also rely on the work of Bert (2001) who shows a large local field survey: in Brossainc, Vinzieux, Limony, Félines, Serrières, Saint–Marcel, Savas, Peaugres, Davézieux, Champagne, Andance.

[28] In other words, Martel (1983) classifies a western subset of Vivaro–Dauphinois varieties divided by the Rhône (on the right side of the bank) into the Vivaro–Vellave group and by the Rhône on the left bank into the Rhodano–Dauphinois group (which includes Drôme and Isère), then in the East the Alpin varieties. See also Sumien (2009: 15): « A l'oèst, lo vivaro-daufinenc se parla en Isèra Occitana, en Droma, dins lo nòrd de Vivarés, dins l'èst de Velai a l'entorn de Sinjau e dins lei franjas occitanas dau sud de Forés ». Bouvier (1976) considers that the South of the *Drome* could be classified either as Provençal = northern–Provençal or as Provençal = southern–Provençal.

[29] The East of the Haute–Loire includes: Tence, Chambon–sur–Lignon, Sainte–Sigolène,

southern *Vivarais* (which is also covered by our analysis) in the Alpino–Dauphinois group. On the other hand, he classifies *Auvergnat–Limousin* as 'Group D', generally characterized by the deletion of final /–n/ and by the closure of /e/, and /o/ before nasal consonants in [i u], whereas 'Group E' should transition with 'Group D' in Averno–Limousin. According to the supra–dialectal structuring provided by Sumien (2009: 20–21), the Arverno–Mediterranean group, which includes the Averno–Limousine varieties, eastern Occitan, *Auvergnat*, the Protestant dialects of the Velay, Sud–Vivarais, Vivaro–Alpin, Vivaro–*Dauphinois* and Alpin–*Niçard*, show a nasalization from partial [V^n/($\tilde{V}^n$] to complete [$\tilde{V}$]. D'Albera (1986) also considers northern Occitan as an area of innovation in relation to nasalization phenomena that are otherwise absent in Languedoc and Provençal varieties. Indeed, northern Occitan is linked by this feature to a more innovative north-western Romance according to Nauton; i.e. a north-western zone influenced by the centripetal force of Lyon (which includes Velay), Nauton (1974)'s 'amphizone'.

4. Nasal vowels in the atlases and sound archive material for Haute–Loire and Ardèche

In order to establish whether or not nasality was endogenous to northern Occitan (i.e., whether it is an autochthonous process or due to influence from FP, which has nasal vowels, including high nasal vowels), we present evidence from the transitional space between Occitan and FP-speaking areas (Haute–Loire, Puy–de–Dôme, Ardèche and Drôme) described in the previous section, with data taken from the ALF and regional atlases, as well as from a sound archive. Our data show both traces of synchronic nasalization as well as cues of a historical/phonological denasalization. The northern Occitan oral data presented in this section come from the Haute–Loire and the Ardèche region. After a general overview of the atlas data, we concentrate on sound recordings, most of which are from the departmental archive (recordings of speakers from Haute–Loire

Yssingeaux; the Bas–en–Basset, part of the ancient Velay is also Vivaro–Vellave (see Di Caro forthcoming; Grange 2021, 2008; Martin 1997; De Felice 1983, 1989).

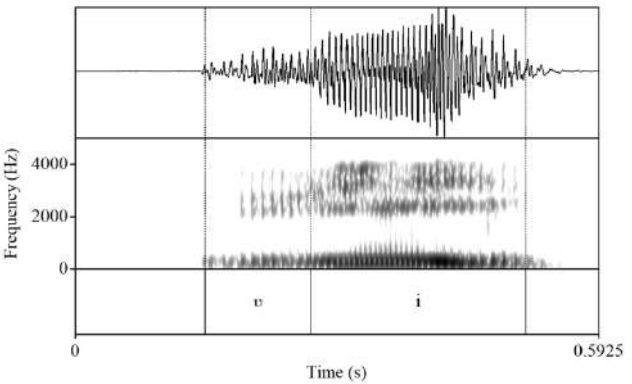
and Ardèche).

Some nasal contexts found in Haute–Loire are described by Nauton (1974): according to him all /m/ and /n/ consonants in coda position trigger nasalization of the preceding vowel in examples such as [tsɑ̃]/[tsõ] *champs* 'field' (Fr. [ʃɑ̃]) Latin CAMPUS (/M/ is in coda position in the Latin etymon). However, according to Nauton, Latin intervocalic /m/ and /n/, consonants that later became final (due to vowel elision) and lenited (including consonantal loss) do not always trigger the process: while the /m/ nasalized any preceding vowel, as in [fũ] *fumée* 'smoke' < Lt FUMU or [fõ] *faim* 'hunger' Lt FAME (/M/ is an onset in Latin), the final /n/ (onset Latin /N/) fell without nasalizing the preceding vowel, as in [ple] *plein* 'full' Lt PLENU, [vi] *vin* 'wine' Lt VINU, [bwo] *bon* 'good' Lt BONU. The [bwo] form, however, seems to be intermediate, or 'fudged' (Hornsby 2006: 87), as FP too has diphthongs from the mid–open vowels [ɔ], whereas northern *Auvergnat* Occitan does not. In fact, in our data below in the Puy–de–Dôme we have the Occitan form [bu] without a diphthong. The form [bwo] indicates the absence of Occitan nasality but does evidence the FP diphthong, while in the Pilat region we have [tsõ] or [tjõ] (see Bert 2001: 341), where glide insertion is not the outcome of Lt Ĕ[30]. For /i +N/ a trace of nasalization is visible in the denasalized Occitan form reported by Veÿ (1911: 33) <ve> *vin* 'wine' in old Saint–Etienne texts.[31] It should be noted that in the Pilat region, according to field surveys by Bert (2001: 364), the Occitan forms [ve/vi] for VINU, [ve'tse] *voisin* 'neighbour' for VECINU and [ma'te] *matin* 'morning' for MATUTINU, [pe] for PINU for example, are present in Félines (Haute–Loire), and in Vinzieux (Ardèche) [madzi] MATUTINU but [ve'tse] and [vi]. Here, [e] in place of the etymological /i/ indicates nasalization in Ardèche and Haute–Loire, even though synchronically the vowels are oral ([e]). But most varieties in Ardèche and Haute–Loire show the primitive nasal /i +N/ in the synchronic oral Occitan state [i], see ALLy 438 and ALF 1667 *pin* 'pine'. We therefore observe variation in the system. Furthermore, Nauton (1974) reports that nasalization is triggered

[30] See for TĔMPUS diphthongized nasal vowels in the FP-speaking Forez to the West, Gardette (1941: 39–40), as well as the Dauphiné (ATF 366; Devaux 1892: 157).

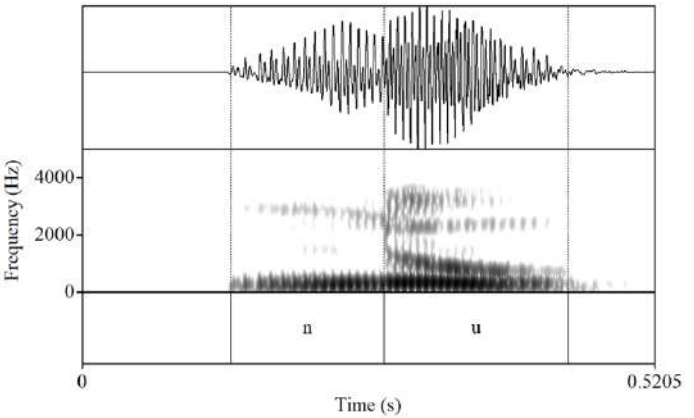
[31] See also in northern *Dauphiné* the ATF maps n° 197 'jardin/garden' and n° 391 'voisin/neighbour'.

when a lexical item is inserted in a constituent before a consonantal onset, e.g. [ɛ̃ plɛ̃ sa] 'a full bag' (where plɛ̃ replaced [ple] Lt PLENU). It is important to note (as we did for nominal paradigms above) that the verbal oppositions are also morphological in this Occitan area between the 3rd pers. SG of the present and the 3rd PL in the present tense of the indicative: ALF 574 *ceux qui finissent* 'those who end up' pt. 833 [fəˈnisũ] (southern Ardèche) vs. ALF 576 *que ça finisse* 'that it ends' (see Figure 7). The situation described by Ronjat in relation to nasalization in group D (i.e., with deletion of nasal /N/ and high vowels [i u] before nasal, also with raising of /o/ to [u]) is the one we find in the maps (for example *chien* 'dog') of ALLy and ALF and in the sound examples of *Auvergnat* Occitan as seen in Neschers in the Puy–de–Dôme (see Spectrograms 3, 4, 5, taken from archival recordings of the Haute–Loire):[32]

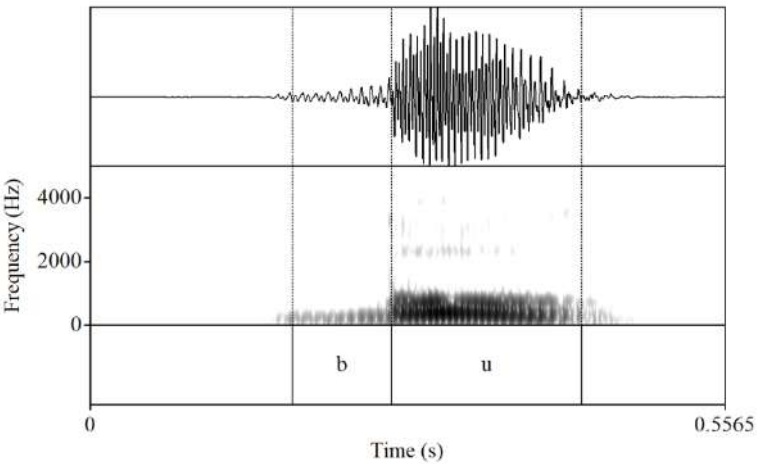


Spectrogram 3 : [ʊi] *vin* 'wine' (Occitan Auvergnat Neschers in the Puy–de–Dôme)

[32] The general weakening of the final /n /of oxytons is also characteristic of eastern and northern Languedoc dialects.



Spectrogram 4 : [nu] *nom* 'name' (Auvergnat Occitan Neschers in the Puy–de–Dôme)[33]



Spectrogram 5 : [bu] *bon* 'good' (Auvergnat Occitan Neschers in the Puy–de–Dôme)

[33] For [u] F1 = 300 Hz, F2 = 860 Hz (F3 = 2400), see *infra*.

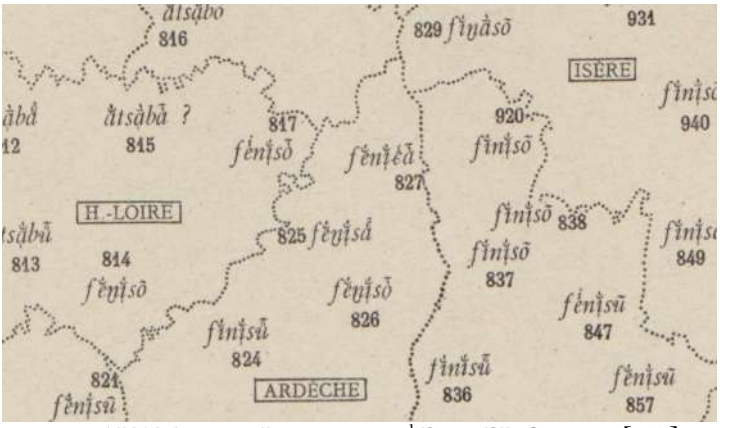


Figure 7 : ALF Map 574 Latin UNT 3rd Pers. PL Occitan [ũ/õ] *ceux qui finissent* 'those who end up' õ < –ANT?

This follows what Nauton observes for Haute–Loire, namely that /n/ seems to fall without nasalizing (triggering a raising), while /m/ systematically nasalizes the preceding vowel (see Spectrogram 3). The Alpine Vivaro–Vellave results follow group D as indicated by Ronjat which features deletion (rarely weakening) and raising of /e o/ + N to /i u/. This analysis is also supported by Bert's (2001: 360) investigation in the Pilat transitional zone, which provides cases such as Bourg–Argental [mo'tu] 'sheep' Lt MULTONE, [ka'ju] 'pig' Lt –ONE, [dar'bu] 'mole' Lt, [me'zu] 'house' Lt MANSIONE, [sa'vu] 'soap' Lt SAPONE (see also ALLy 321 'pig' pt 68 Sainte–Sigolène [ka'ju]).[34]

In the following sections we will only deal with the treatment of high and original mid vowel qualities before nasal consonant N ([–high > [+high]]) and the treatment of open [+low] vowel /a/ + Nasal sequences ([a] > [ɑ/ɒ]). In particular, the first treatment allows us to compare Occitan outcomes with FP, which nasalizes high vowels.

Returning to map 372 of the ALLy, we can see that nasalization in the FP space intersects with denasalization in northern Occitan. We have argued that the FP vowels [ĩ]/[ɛ̃] correspond to /i, e + N/

[34] Same forms for Andance, Peaugres, Savas, Saint-Marcel, La Versanne, Thélis, Le Bessat, Tarentaise, Planfoy, La Valla (ib.).

sequences. In addition, we have stated above that this [i] may also be the result of diachronic palatalization of /k+a/ sequences: the combined palatalization/affrication of /k+a/ and the FP realization of /a/ as [i] spread in northern Occitan too, e.g. in forms such as [tsi]:

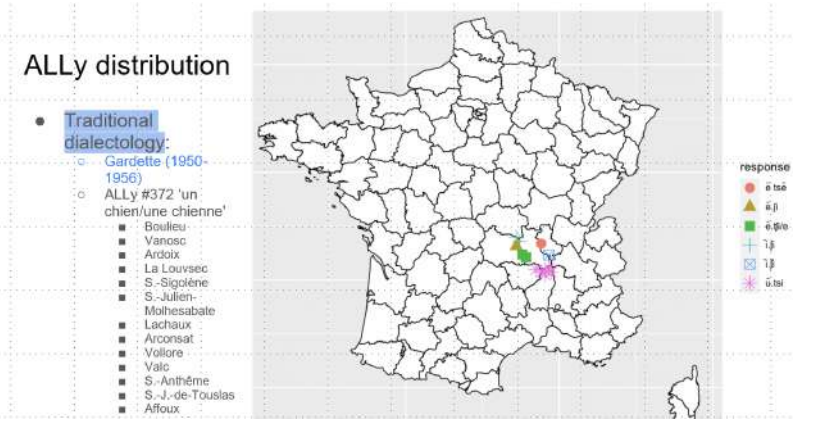


Figure 8 : Distribution of FP/Occitan outcomes based on ALLy Map 372

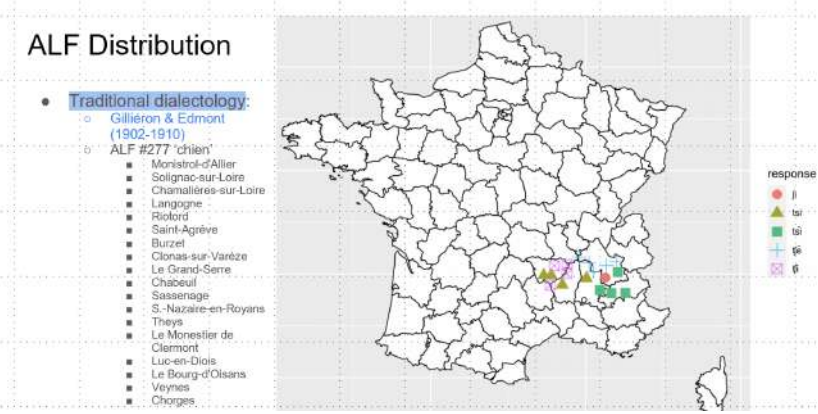


Figure 9 : Distribution of FP/Occitan outcomes based on ALF Map 277

However, we note for high vowels a decline in nasalization in the ALLy localities of Ardèche and Haute–Loire at the FP/Occ border:

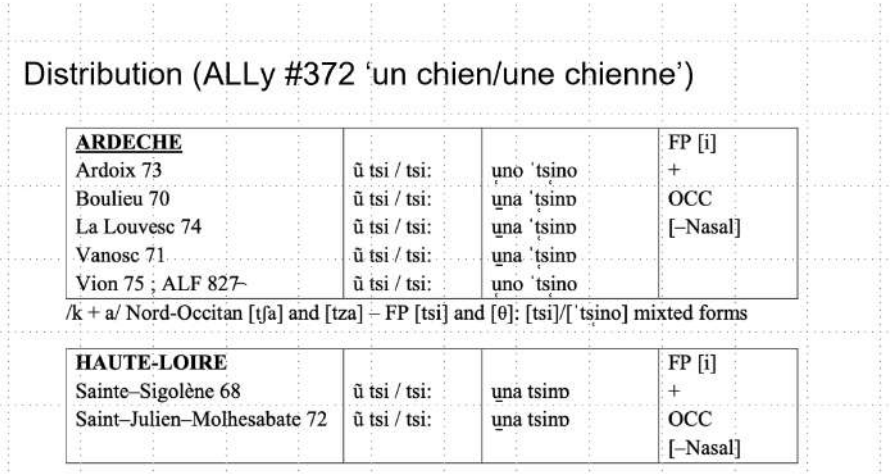


Figure 10 : Occitan distribution of oral based on ALLy 372

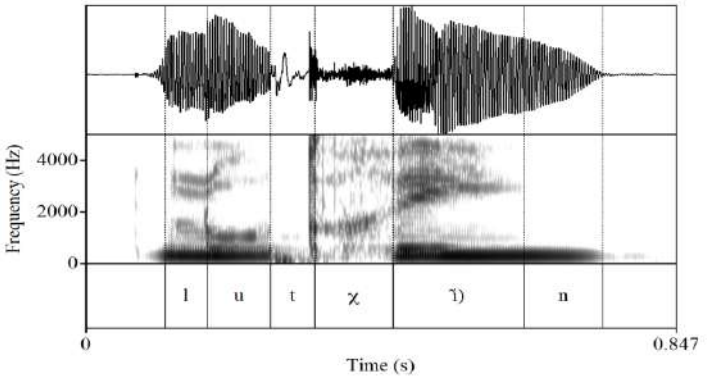
We have hence the FP/Occitan definition for these forms, in the sense that the FP high vowel [i], due to the FP palatalization/affrication, is not nasalized, which is consistent with central Occitan, and partly consistent with northern Occitan). For example, we do not have any high nasal vowels around the Noirétable plateau, which is Occitan-speaking, or in the South at Saint–Bonnet–le–Château, which is related to Vivaro–Alpine Occitan (point 816 of the ALF in the *Loire*), and we do not have high nasal vowels in the Haute–Loire either. These two maps show that in the FP area there is a decline of nasalization with regard to high vowels in contact with mainly Auvergne Occitan, but also Vivaro–Alpine Occitan in the Loire. The same situation is represented by the data we have analyzed from Auvergnat Occitan from Neschers in the Puy–de–Dôme department.

4.1 Nasalization in the Southern of Ardèche (Low–Vivarais): field surveys

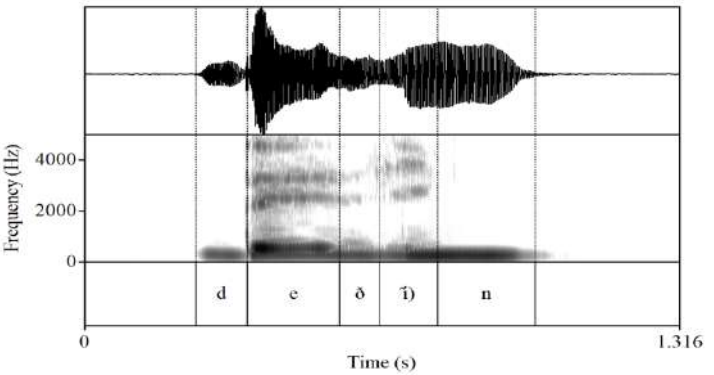
The situation depicted above can also be reversed in the southern part of Ardèche (Jaujac)[35], if we consider Spectrograms 6 and 7 ([lu tχĩːⁿ] *le train* 'the train' and [de dĩːⁿ] *dedans* 'inside' (vs. French [tχɛ̃]

[35] [d͡ʒowd͡ʒa] in this Occitan variety.

and [dədɑ̃]):



Spectrogram 6 : [lu tχĩ)^{n}] *le train* 'the train' (Occitan Vivaro–Alpin Southern Ardèche – Jaujac)



Spectrogram 7 : [de dĩ)^{n}] *dedans* 'inside' (Occitan Vivaro–Alpin Ardèche – Jaujac)

In both spectrograms, nasalization is visible in the lowering of F1 due to the low nasal pole (around 250 Hz). The acoustic association of the nasal cavity shows additional formants, but particularly a clear anti–formant between F1 and F2, i.e., a zone where a decrease in harmonic intensity is observed. This anti–resonance or anti–formant is visible in the second part of the nasal vowel. A reduction in energy (amplitude) is also visible starting from the nasalized part of [ĩ)],

triggering a progressive weakening of highest formants up to formation of a final nasal coda.

Based on the two spectrograms one might deduce in Jaujac that nasalization on high vowels is an adaptation of the French loan ([tχɛ̃] and [dədɑ̃]). However, we show with other data also from the sound archives of Haute–Loire department (e.g. Saint–Agrève in Ardèche) that high vowels in Ardèche can also be nasalized, and that vowel raising (even when it comes without synchronic nasalization) is a trace of a (previous) nasalization, as it is the case for Spectrograms 4, 5 in Puy–de–Dôme [bu] < Lt BONU and [nũʲ] Lt NOMEN. Jaujac indicates more an intermediate stage, where denasalization is not reached, but /e o/ are nonetheless raised; the same pattern is found for back high vowels [mũ] 'mon/my' (French) and [pɔpiˈjũ] 'papillon/ butterfly', next to [de dĩː].

For the southern Ardechean Vivaro–Alpine dialectal area, our corpus consists of surveys conducted in the village of Jaujac (surveys 2019–2021),[36] as well as of sound files achieved at departmental Sound Archive of the Haute–Loire Saint–Agrève (southern Ardèche), and from the Vivaro–Alpine Occitan of Albon (open access on the platform *CoCoON= COllections de Corpus Oraux Numériques*).

As for Jaujac, it is a village located in the southern half of the department in the *Cevennes* region (see Moulin 2006):

[36] Several surveys were carried out by Clara Duvert (2019–21), a Masters student in Linguistics and Dialectology at the University Jean Moulin Lyon 3, with a Jaujac resident who speaks Occitan fluently, Raymond Constant, aged 94, born in La Souche. Raymond Constant is the author of a book entitled of *L'année du Lazare*, which recounts the daily life of his grandparents at a time when Occitan was the dominant language of the region (see Duvert 2021).

Figure 11 : Map of Jaujac (Cevennes, after Moulin, 2006)

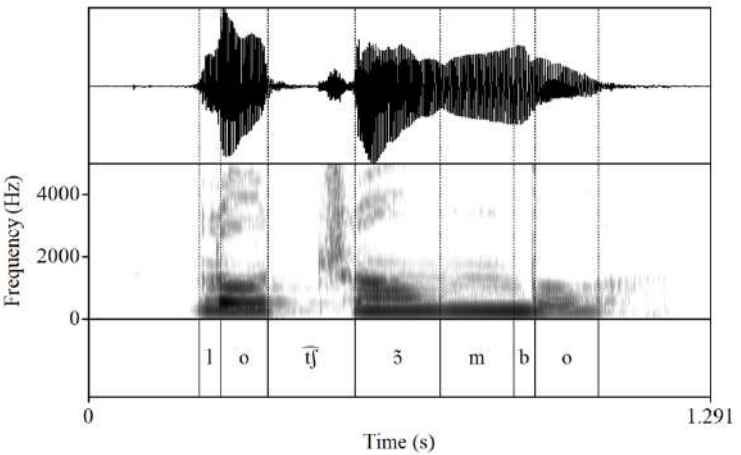
The contact between different Occitan dialects and the influence of FP in the North makes the Ardèche an interesting territory for the analysis of V+N sequences. Southern of Ardèche represents a transitional zone between an area to the South (that is part of the Occitan domain, close to Languedoc language) and an area to the North (High *Vivarais* or Vivaro–*Vellave*); FP is also spoken in the extreme North of the Ardèche. Recall that western Ardèche is a sub–dialect of Auvergne; indeed, as in section 2.1, the Ardèche is part of the Vivaro–Alpine, the eastern of close to Provençal domain, while the western periphery also belongs to the Auvergnat Occitan. The Occitan variety of Jaujac is usually classified as sub–dialect Vivarais, part of the group Dauphinois which is known as Vivaro–Alpine (see

Section 2.2). The Lower–Vivarais is distinct from the Vivaro–Alpine of Haut–Vivarais, spoken in the northern half of the Ardèche. Ronjat and Bec classify the Jaujac dialect as a southern dauphinois variety (see Duvert 2021), even if the Jaujac variety seems to them rather oriented to Languedoc.[37] However, in Jaujac there is systematic palatalization and affrication of /k/ in onset position; the northern Occitan outcome is [lu.tʃa] 'the cat' ALMC 567 *un chat/deux chats/chatte* 'a cat, two cats, cat (F.)', see also ALF 250, 1498), and in Jaujac the variant [tsa] can be found (Duvert 2019–21).[38] In Low–Vivarois we find /a/ realized as [ɔ] (i.e. velarization) in many positions, for example, *chastanha* 'chesnut' is pronounced [ʧɔsˈtɔɲɔ] or [ʧasˈtɔɲɔ]; this feature is shared with eastern or northern Languedoc too; this also concerns nasalized /a/ throughout Low–Vivarois (pretonic and posttonic).[39] In Jaujac this velarization and rounding process is also visible in stressed vowels followed by a nasal consonant /a + N/ (see also Ozawa 2007). In addition, unlike in Languedocian, the nasal consonant is kept in internal coda (see also Garnier 2020). This characteristic is specific to the Vivaro–Alpine varieties. In Spectrograms 8 and 9 the stressed low vowel is velarized and rounded, the nasal consonant holds as an internal syllable in coda [lo.ˈʧɔ̃mbo] 'la jambe/the leg' (French ([ʒɑ̃b]), [mɔ̃nʤa] 'Fr. (il) mange/(he) eats':

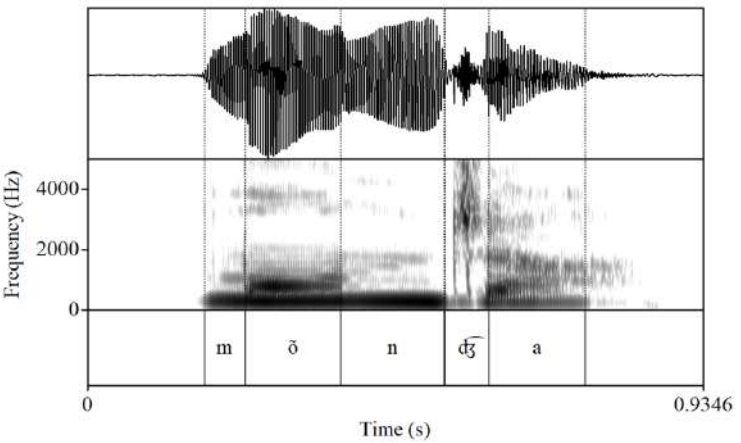
[37] For the conservation of Latin intervocalic dentals, for instance, which is precisely one of the main features characterizing Vivaro–Alpine. In [dʒ͡owdʒ͡a] intervocalic –T– is maintained as voiced stop, see [kuˈpadɔ] 'cut', while in Vivaro–Alpine we expect in past participle from Lt –ATU, lenition in V_V gives –T– > ∅. The other differences are retention of intervocalic –s– (*susar* vs. *suar* [high–*Vivarois*] < Lt SUDARE 'to sweat'), lack of palatalization of /s/ and of /z/ preceding a vowel ([saˈzu] vs. [ʃaˈʒu] 'season' Lt SATIONE), as well as lack of obstruents palatalization in general, the retention of /s / before /p t k/: *escoba* [ɛsˈkuβɛ] [Jaujac] vs. *eicoba* [ɛjˈkuβǫ] [high–*Vivarois*], use of the first person verbal inflection –e (usually –o in Vivaro–Alpine), see Jaujac [βajle] 'je vais/I go. The idea is widely thought that the Low–*Vivarois* is a northern Occitan Languedoc variety, as the Gévaudanais variety (Ronjat § 850). This belief is based on Ronjat (Vol. 4, page 46): « The general habitus [of Low–*Vivarois* speakers] gives the listener the impression of a Languedocian with *cha* and most of the analysis confirms this impression ».

[38] Note that in Jaujac there is no betacism as there is in Languedocian, this is consistent with northernOccitan Vivaro–Alpine and with Haut–Vivarois, see ALMC n° 397 'vache/vaches, bœuf/bœufs/cow(s)/oxen'. Duvert (2021) points out that pt 31 (*Chirols*, Ardèche) of the ALMC, is the closest to Jaujac [dʒ͡owdʒ͡a].

[39] It is important to note that another feature that distinguishes Jaujac from Languedocian is the absence of betacism, see close pt 31 ALMC 397 'vache, vaches; boeuf, boeufs' [vatso].



Spectrogram 8 : [lo ˈt͡ʃɔ̃mbo] ‘la jambe/the leg’ (Vivaro–Alpin Ardèche – Jaujac)



Spectrogram 9 : [ˈmõnd͡ʒa] ‘mange/eat’ (Occitan Vivaro-Alpin Ardèche – Jaujac)

The height of F1 and F2 indicates the velarization and rounding of the vowel. In addition, we note an additional low nasal formant.

In both spectrograms, we notice too the appearance of anti–formants, where we observe a strong drop in the intensity of harmonics, as well as the appearance of extra–formants (in these surveys there is also [blɔ̃] 'blanc/white' French [blɑ̃]). We also note the modification of the formant values of the corresponding oral vowel: a higher F1, and a weakening in amplitude of the higher formants. The formants of the nasal vowel [õ] are in both cases around F1 = 600 Hz, F2 =1400 Hz, F3 = 2000 Hz. The internal nasal coda in such examples is also confirmed by atlas data, see ALMC 1319 '(la) jambe/(the) leg' pt 31 [ˈʧsɔ̃ᵐbo] (see also pt 27 or pt 35 [ˈʧʃɔ̃ᵐbo] among others), ALF 709 ' jambe/leg' and ALLy 1104 'la jambe/the leg' where nasalization in Haute–Loire is complete (pt 60 Sainte–Sigolène and pt 72 Saint–Julien–Molhesabate [ˈʧsãbo]. Complete nasalization is also present in the Ardèche in the pts of the ALLy (73 Ardoix, 74 La Louvesc, 75 Vion) [ˈʧsãbo].

Figure 12 : ALF map 709 *jambe* 'leg' V̆ + N internal coda (Haute Loire, Ardèche, Drôme), e.g. pt 833 (Vogüé –Ardèche) [ˈʧʃɔ̃ᵐbo]

It is important to highlight that the Jaujac variety is also situated in a transitional zone at the crossroads of Auvergnat and Provençal (i.e. spoken South of Drôme). Indeed, as it can also be seen from map ALF 709, Ardèche shares with Provençal spoken in Drôme the sequences V̆ + N in internal syllables (e.g. pt 855 [ˈʧʃɑ̃mbo] Nyons). To triangulate our analysis, we draw on examples from an available unpublished and entirely handwritten grammar of Jaujac Occitan,

(Marcel Coudène, unpublished).[40]

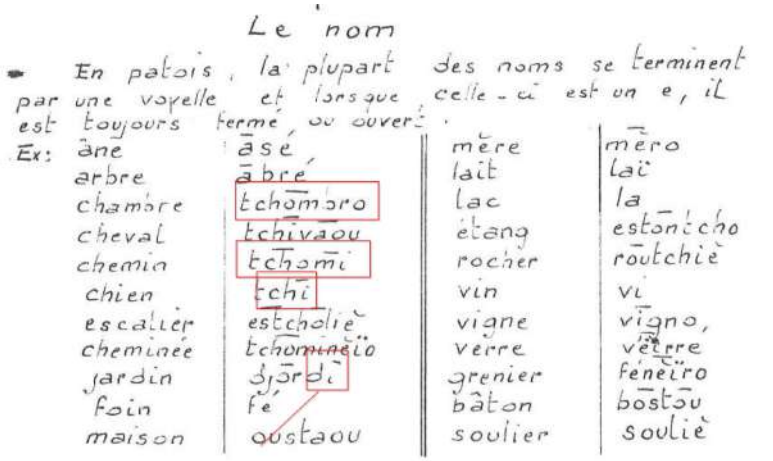


Figure 13 : Jaujac Occitan grammar variants [ʧĩ] CANE 'chien/dog', [ʤɔʁˈdĩ] 'jardin/garden', [ʧɔ̃mi] 'chemin/path' *CAMMINU, [ʧɔ̃mbʁo] CAMERA 'chambre/room', [bostɔw] 'batôn/stick', [vin] 'vin/wine' VINU

Figure 14 : Jaujac Occitan grammar depiction of syllables *an* and *am* which become *on* or *om* in this variety, e.g. [ʧɔ̃mbʁo] CAMERA 'chambre/room', etc.

Among the similarities linking the Jaujac variety to *Drôme* Occitan is the production of nasal /n/ word-finally (see e.g. high vowels in Spectrograms 6/7). This is also visible from atlas data with

[40] *Grammaire du Patois des Vallées des Cévennes Vivaroises.*

low vowels derived from Latin open syllables. If we look at ALMC 1324 '(la) main droite, gauche/(the) right hand, left' Lt MA.NU; as we can see in pt 31 close to Jaujac the effect of nasalization is a further backing of the /a/ vowel which is nasalized and followed by a coronal nasal: [mõn] 'main/hand'. The final consonant is also indicated in the grammar by the unpublished grammar:

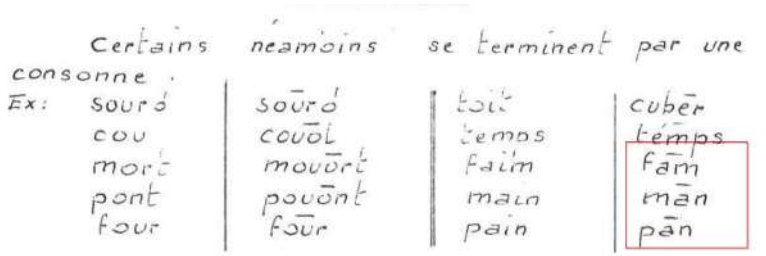


Figure 15 : Jaujac Occitan grammar [fɑm] 'faim/hunger', [pɑn] 'pain/bread', [mɑn] 'main/hand'

The same phenomenon is visible in ALF map 796 'main, les mains/hand, (the) hands', e.g. pt 833 (Vogüe, Ardèche), pt 857 (Luc–en–Diois, Drôme) [mɑ̃n], whereas the form given by the ALLy 1093 'main/hand' for Sainte–Sigolène (pt 68) in the Haute–Loire is back and rounded without nasalization, i.e. [mɔ]; the same non nasalized and velarized form is present for Ardèche, pt 73. The nasalization boundary of /a/+N can also be established with the help of ALLy maps 419 'pain/bread' (French [pɛ̃]), 1093 'main/hand' (French [mɛ̃]), 1309 'demain/tomorrow' (French [dəmɛ̃]). The ALLy shows that the boundary which separates the FP nasalized forms [pɑ̃]/[mɑ̃]/[demɑ̃] from the velarized/rounded [–nasal] forms [po]/[mo]/[demo] follows the same boundary as for the nasalization of [ĩ]). To the South, the ALLy shows that nasalized forms are still found in the North of the Ardèche, and our own survey work shows that nasalized forms are present too in the southeastern village of Jaujac.

Thus, the tendency for /a/+N to become [õ] is also inherent to FP as can be seen in the FP–speaking part of the Pilat region (see Bert 2001: 358ff's description). We note too that the Bourg–Argental area of the

Pilat Mountains is considered to be Vivaro–Alpine Occitan.[41] This nasalization is found in the Stéphanois region, [põ] < PANE and [grõ] < GRANU (Veÿ 1911: 14), which is also confirmed by Bert (2001)'s fieldwork, where he notes [grõ], [demɑ̃] < DEMANE, [mɑ̃] < MANU 'hand', then in the varieties of the *Terres froides* (in the Auvergne–Rhône–Alpes region, precisely the Isère department, see ATF), or in Pélussin (Champailler 240).

4.2 Haute–Loire speech archives: Chambon–sur–Lignon (recordings of Theodore de Felice), and Saint–Agrève (local radio, southern Ardèche)

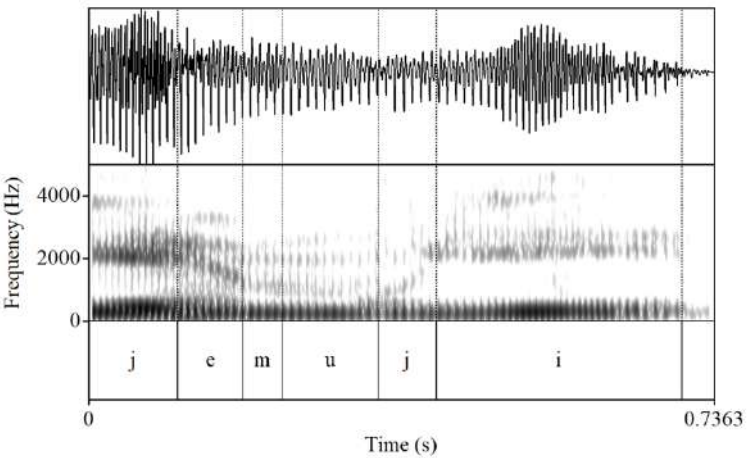
We also analyzed nasal vowels production in the departmental archive sound files from Saint–Agrève in Ardèche. These recordings of a Vivaro–Alpine variety located in southern of Ardèche, are to be included in the High–Vivarois (in the Vivaro–Vellave region, see section 2.1), particularly as a conservative language of the High–Lignon. We recall that the situation of Vellave Occitan is related to the very particular situation in the Velay where the Reformed asserted themselves against Catholicism (as well as in the neighboring region of Vivarais, to which Saint–Agrève also belongs). The state of the country of reformed catholic religion contributed to the specificity of this language which remains isolated and conservative to the point that Nauton designates it with the name of 'butte–témoin' (*witness*) (for details see De Félice 1983; 1989). The audio files we have analyzed come from the Departmental Archives of the Haute–Loire, especially from Chambon–sur–Lignon (Haute–Loire), and Saint–Agrève (Ardèche, France). For northern Occitan, our materials include De Félice's own recordings related to the Protestant area of Chambon–sur–Lignon (canton of Tence), as well as the recordings of a local radio from the same area. We further triangulate our data with analyses of recordings taken from *Le patois vous parle*, a regularly scheduled segment of the *Cimes du Lizieux* (1984–1990) radio program.

Below we illustrate segments extracted from one recording in

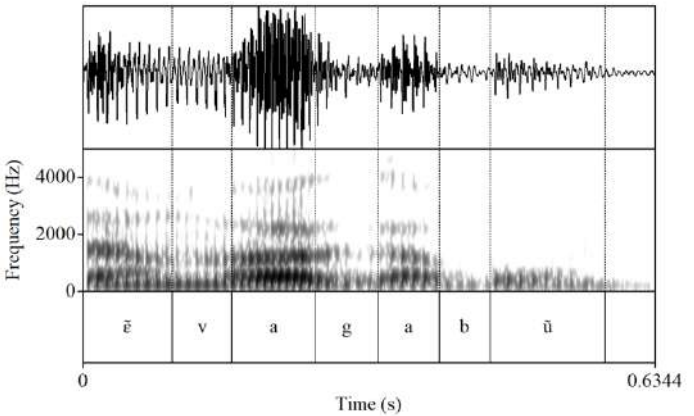
[41] We recall also that this region crossed by the boundary between Franco–Provençal and Occitan Vivaro–Alpine (see Gardette 1983: 176–178, also Ronjat (1930–1941[vol. I]; Tuaillon 1964 for the Eastern part ; Nauton 1966; Martin 1979, who has carried out numerous field surveys).

particular (*La poule noire*, 'the black hen'). The document identification is as follows:

Archival recordings
Tale : La poule noire 'The black hen' / Paul Paya
Translator (Occitan): Joseph Deléage
Speaker (Occitan): Joseph Deléage
Speaker (French): Théodore de Félice
Collector : Jean–Michel Guinet
Places : Le Chambon–sur–Lignon (Haute–Loire), Saint–Agrève (Ardèche)



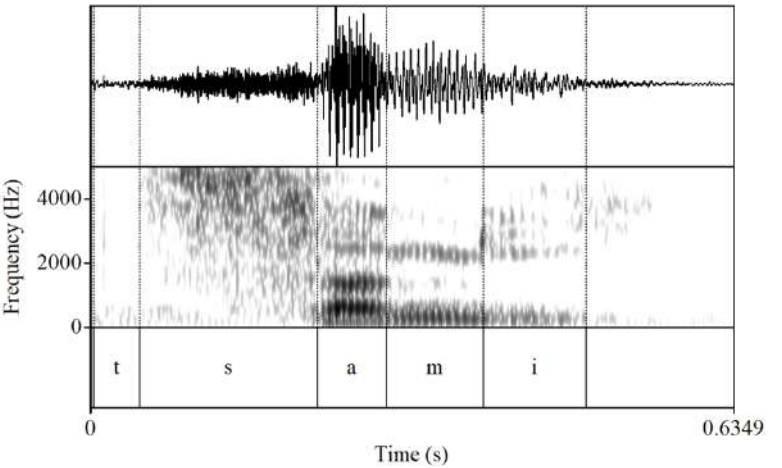
Spectrogram 10 : [muji] 'moulin/mill' (Occitan, Ardèche – Saint–Agrève)



Spectrogram 11 : [ɛ̃.vɑgabũ] 'un vagabond/a vagabond' (Occitan, Ardèche – Saint-Agrève)

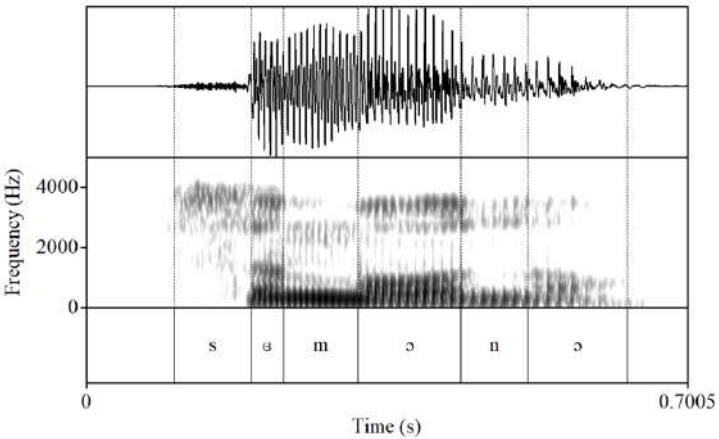
Spectrograms 10 and 11 suggest that the variety coalesces structurally with what we would expect for Vellave Occitan, with clear realization of nasal vowels or denasalized [i]. In Spectrograms 12 and 13 we further illustrate segments extracted from a local radio broadcast, we give below the details:
(source http://www.marraire.eu/Escotar/Escotar11.php)

An Occitan cultural space /Un espace culturel occitan/ Paraulas de tèrras occitanas fichier V1. Las muaas sègan totjorn le mèsme chamin
Type: Audio document, tale in Occitan.
Tale : Las muaas 'milling'
Speaker (occitan) Sinjau / Abbat Baure
Place: Haute–Loire



Spectrogram 12 : [tsami] 'chemin/path' (Haute–Loire local radio)

As Spectrogram 12 shows, with [tsaˈmi] 'chemin/path' (vs. French ([ʃəmɛ̃] < Lt /i + N/) we observe the synchronic result of non–nasality, i.e. high vowels and deletion of the final nasal consonant, which has already been demonstrated for Occitan as spoken in Auvergnat (cf. Spectrograms 3, 4, 5). Conversely, when we consider the low vowel in sequences /a/ + N (Spectrogram 13), the formants clearly illustrate the velarization of /a/ with F1 at 486 Hz and F2 at 870 Hz.



Spectrogram 13 : /a/ + N [sɐˈmɔnɔ] *semaine* 'week' with velarization of /a/ = [ɔ] < Late Lt SEPTIMANA (Occitan, Vellave – Saint–Agrève Ardèche)

4.3 The *CoCoON* corpus: nasalization in Vivaro–Alpine Occitan of Albon

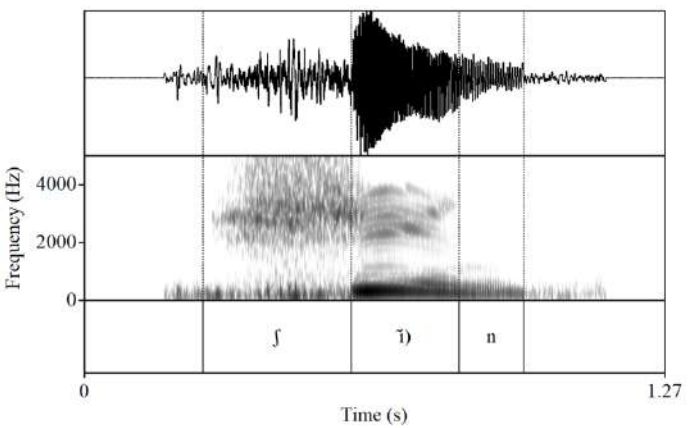
Our analysis is further supported by the oral surveys and texts available for the Occitan variety of Albon, Vivaro–Alpine (see Quint 1994; 1999), deposited on the platform *CoCoON COllections de Corpus Oraux Numériques* which is hosted by the TGIR Huma–Num:[42] https://cocoon.huma-num.fr/.

The comparison of these varieties is useful for our purposes since the Albon variety is geographically close to Jaujac (southern Ardèche). However the Albon variety belongs to the Vivaro–Vellave Occitan variety.

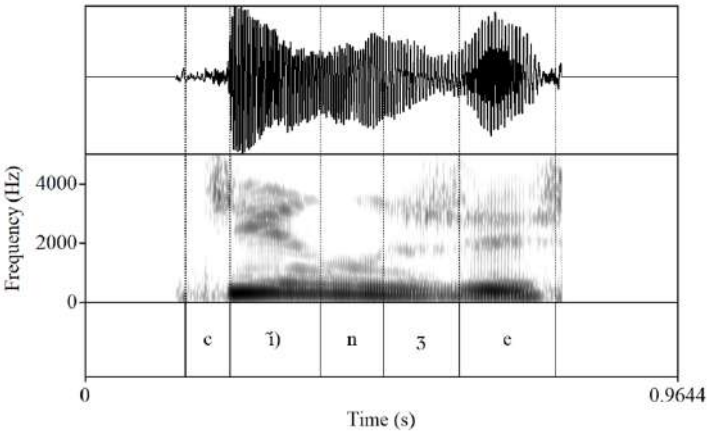
Both the Jaujac and Albon varieties are *Vivarois* and are part of the *Dauphinois* group. We have mentioned that the Jaujac variety is a South–Dauphinois or Low–Alpine variety (close to pt 31 of the ALMC). The Ardèche Vivaro–Alpine variety of Albon (1994) is also defined by Quint (1999) as Vivaro–Vellavian Alpine speech. We

[42] TGIR = *Très Grande Infrastructure de Recherche*, supported by CNRS. *Huma-Num* is a research infrastructure dedicated to digital humanity (*Humanités Numériques*).

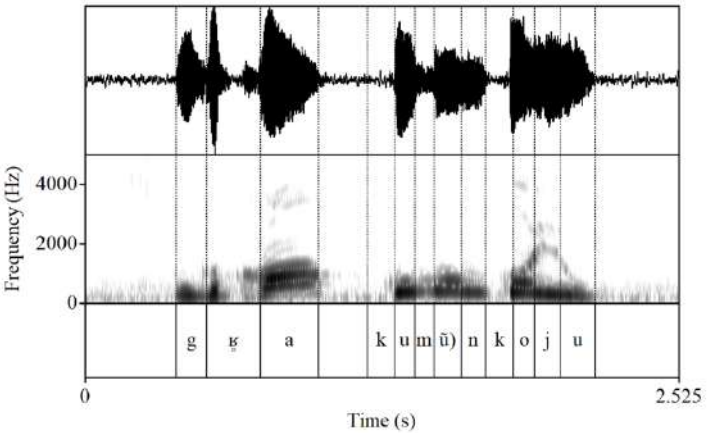
have seen clear traces of nasality in the southern Dauphinois or Low–Alpine variety of Jaujac, and the situation is not very different for Albon: in those varieties South of Ardèche high nasal vowels can be found in internal coda position (also followed by a weak consonantal nasal), e.g. [ʃĩⁿ] 'cinq/five' (cf. Spectrograms 14 and 15). However, we find next to it [ʃu] (< SUNT, with palatalization of /s/ preceding a vowel typical of the High–Vivarois, cf. Spectrogram 18), [ũ koju] where we have the Vivaro–Vellave outcomes as in Haute–Loire: the nasalized indefinite article and the raising of /o/ to [u] in N < Lt –ONE ([koju] is a FP lexical item (cf. Spectrogram 16 and [tsi] from Lt CANE, as we shown in ALF 277 and ALLy 372; Spectrogram 17, which matches [tsami], [vi], and at a nasalized stage [mujɪ̃ːˀ]).



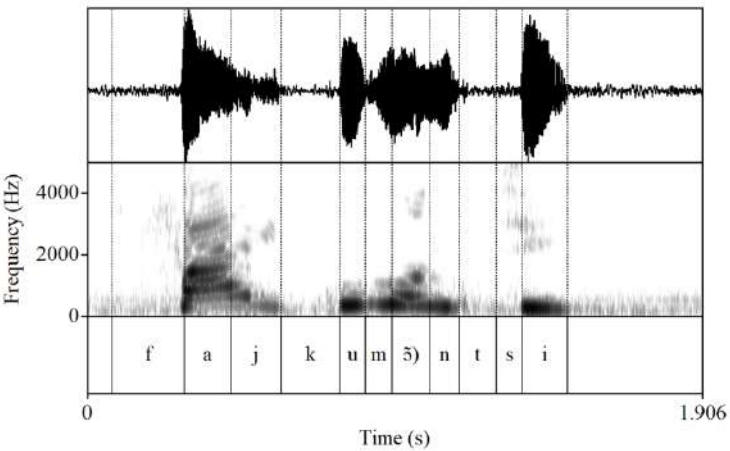
Spectrogram 14 : [ʃĩ)ⁿ] 'cinq/five' with intermediate nasalization of high vowels in internal coda (Albon, southern Ardèche)



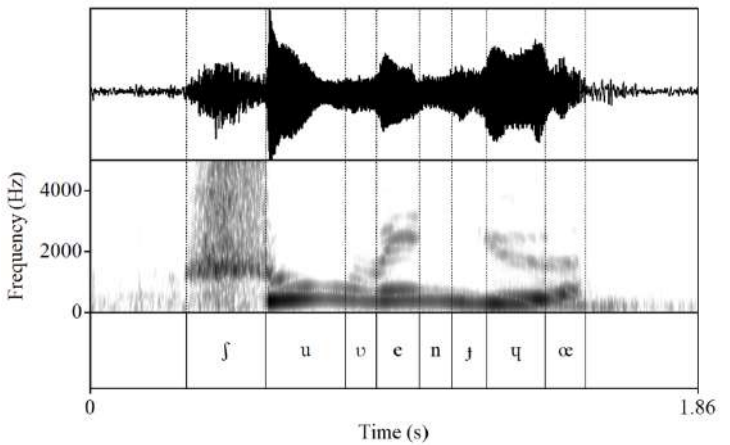
Spectrogram 15 : [cĩ)nʒe] 'quinze/fifteen' (Albon , southern Ardèche)



Spectrogram 16 : [gɣ̞a kumũ)nkoju] 'porc/pig' – Lt –ONE [koju] (Albon, Southern Ardèche)



Spectrogram 17 : [faj kumɔ̃)ntsi] *il fait un froid de canard* 'it's as cold as hell' (Albon, southern Ardèche)



Spectrogram 18 : [ʃu.ʊenɟɥœ] *ils sont venus* 'they came' (Albon, southern Ardèche)

For Occitan then, the evidence thus far suggests that the N in V+N sequences is not always deleted but instead promotes nasalization of

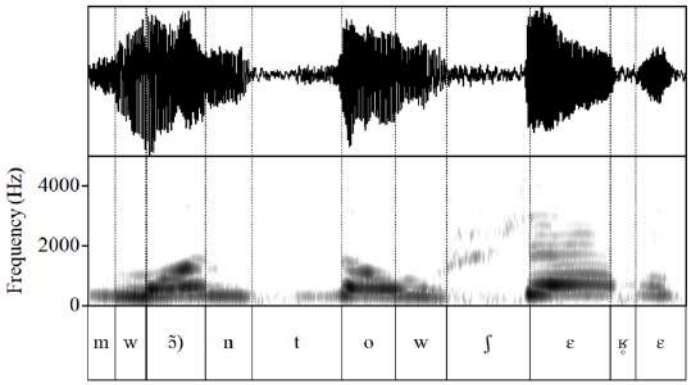
V. Similar observations have been made by Russo et al. (2021) for the *Drôme*. Across *Drôme* and *Montélimar–Dieu–le–fit* axis (Bouvier 1966), we also find some specific similarity (to the Spectrograms 11, 14, 15): the /n/ after the high vowels /i u/ is not always deleted (in the Romans–Valence plain and in the *Drôme* Valley), and it shows nasalization of the vowel.

This type of nasalization is very similar to that reported for FP speakers of Forézien (Straka 1955: 270). This phenomenon can be observed all along the border of the *Drôme* (e.g. *Chambrillan*) where atlas data captures it as diphthongal realizations, e.g. [raˈzjẽⁿ] *raisin* 'grape', even if in a number of cases the /i/ has been recorded intact with a weakly articulated nasal consonant, e.g. [veˈriⁿ] *venin* 'venom', or zero [vi] *vin* 'wine', etc. In this area we are therefore faced with tripartite variants of the type /i/ + N [djẽⁿ/dĩ], but also [din] and [di] (Bouvier 1966; Russo et al. 2021). This evidence further corroborates our discussion above concerning e.g. Spectrogram 7 [de.dĩ)ⁿ] 'dedans/inside' from the Ardèche (Jaujac).

5. Nasalization of /a/ + N in northern Occitan : Phonological cues

This situation between nasalized and non–nasalized Occitan forms raises questions and suggests that the northern Occitan inventory may have included extensive endogenous nasalization in the past. Moreover, the cases of non–nasalization observed in the Albon variety of Ardèche or Chambon–sur–Lignon in Haute–Loire may well be indicative of a recent phase of denasalization. We have observed intermediate stages of nasalization, especially in southern Ardèche, but the departmental archive materials coupled with historical atlas data clearly show that we also observe an intermediate phase of nasalization in northern Ardèche, the Haute–Loire, Drôme and Lozère. In addition, we have observed two phonological tendencies triggered by underlying nasals, following V, in open and closed syllables: a tendency to raise mid rounded and unrounded vowels /e ɛ o ɔ/ to [i u] without nasalization of the surface forms (especially in Puy–De–Dôme and Haute–Loire, e.g. in forms such as [bu], or [vi] in variation with [ve], where the lowering is a trace of previous nasality) as well as evidence for an intermediate

phase of variation demonstrating closed vowels +N with nasalization (e.g. [vɑgabũ], with generalized nasalization of [i u] as in FP in all the sequences with mid and high vowels + N). The second phonological tendency that we have described is clearly the velarization and rounding of the low vowel /a/ in the sequences of underlying stressed /a/+N from open and closed Latin syllables (e.g. PANE, GRANDE, CAMBA, MANE, DEMANE etc.), which results in the variable realization of [ɑ ɒ o] as well as [ɑ̃ ɒ̃ õ] depending on the region. We have also noted in the Haute–Loire sound archives velarized forms such as [aw] resulting from /a/+N. These forms are also attested in descriptive work such as Bert (2001: 357) for the Pilat region, who reports for DEMANE and MANU [deˈmaw] and [maw] in La Valla (Loire); [deˈmao] and [mwao] in Tarentaise (Loire); [deˈmwo] and [mwo] in Saint–Genest (Loire); and [deˈmoœo] and [mao] in Le Bessat (Loire). This particular velar treatment of nasals is observed elsewhere in Romance, e.g. in the phonological nasals of Portuguese: e.g. /beN/ [bẽj̃] *bien* 'well', /boN/ [bõw̃] *bon* 'good', in which the nasal consonant is interpreted as a nasal diphthong with a glide element [j/w] which is the outcome of /N/ (e.g. Meireles 2014). It is a further phonological trace of the fact that the variation we describe is indicative of a more recent process of nasalization and denasalization in the transitional Oc/FP regions, and may attest to the phonological status of nasals. Moreover, this phenomenon does not seem to be exclusive to the transitional Occitan varieties of the Loire, as we find the phenomenon in the e.g. Occitan–Vellave of the Ardèche in Albon too (cf. Spectrogram 19):



Spectrogram 19 : [mwɔ̃)ⁿˈtow ˈʃɛʁ̥e] / mwɔnˈtun ow ˈʃɛʁe/ (S) montent sur la colline, '(S) climb the hill' (Albon, Ardèche)

It is interesting to observe that in a Richigny song, written at the end of the 19th century in the dialect of a village northern of Riotord (Haute–Loire), the final vowel from –ONE gives <*lou garsou<u>s</u>*> 'les garçons the boys' (see Bert 2001: 362).

In spite of the variable realizations, we observe the phonological trace of velarization and rounding of Latin sequences in transitional Oc/FP areas of the type /A+N/ = [aw ao o ɑ ɒ o]/ [ãw̃ ãõ õ ɑ̃ ɒ̃ õ]. We note too that in the FP=speaking part of the country the development of /A +N/ is also either [õ] or [ɑ̃] (see ALLy map 419 *pain* 'bread', also Bouvier 1976 map 54; ALF map 527 *faim* 'hunger', ALF 762 *levain* 'yeast'). Thus, the sequences of /a/ +N provide us some interesting cues in the sense of a diachronic phonological effect triggered by nasalization. We have shown that in ALLy maps n° 419 *pain* 'bread', n° 1093 *main* 'hand', n° 1309 *demain* 'tomorrow' we can observe an underlying /a/ vowel from the velar realization [ɑ ɒ ɔ], as above.

As a crucial step in triangulating our data, it is significant for us that these forms are also attested in the written language as evident from 16th century Occitan texts of Vellave, which we have defined as a sub–Avergnate variety. In the texts of Noëls de Cordat (1631–

1648), edited by Surrel (2022), we observe these orthographic representations which are indicative of nasalization. These texts mark the passage to modern Auvergnat varieties between the 15th and 16th centuries. In these Vellavian Occitan texts, the labialization and velarization of tonic /a/ followed by a nasal consonant in coda or syllable onset position is represented orthographically with <o>, and this happens only in Vellavian texts from 1600 onwards. This 17th century orthography is reminiscent of the current situation of the northern Occitan dialects examined here, and also of the forms indicated in ALLy maps 419, 1093, 1309, and diffused in transitional FP zones as the Pilat area (see above). If in the texts of the 17th century these orthographic representations could indicate nasalization, in the modern varieties which show only the back velar vowels for /A+N/, this indicates a recent denasalization and suggests that these forms were previously nasalized. This denasalization, which shows traces in the velarized /a/ as we have argued, does not occur only in final open syllables and may be independent of the loss of the final nasal, as we have seen for CAMBA in the Jaujac Ardechian variety, e.g. in [ˈʧʃɔ̃mbo]. We have seen in Ardèche from recent surveys that nasal consonants can be maintained or lost in internal or final coda (depending on the variety, Vivaro–Vellavian, Vivaro–Valesian, etc.), and that the preceding vowel may or may not be nasalized (denasalization and subsequent backing and rounding of A seems more apparent from our data in Puy–De–Dôme, Haute–Loire than in Ardèche), see cases such as indicated by the spectrogram [lo ˈʧʃɔ̃mbo] ‘la jambe/the leg’ (Occitan Vivaro–Alpin Ardèche, Jaujac). As we can see in the following examples, we have early nasalization noted by the velar back vowel in 17th c. texts below <chombes> (plural) (Surrel 2022, see also Roux 2020). This is reflected in the ALMC examples that still indicate nasalized forms in modern outcomes as seen in <chombro> (NoëlsCordat) *chamber* ‘room’ < Lt camera, pt 23 [ˈtsɔ̃bro] in the ALMC 719, see ((3) – (5)):

(3) /a+N/ Word–internally = [o/ɔ̃] Old and modern nasalization

Antepenultimate Stress

a. CAMERA : N *chombro* FSG ‘chambre/room’ NoëlsCordat *chombre* CletSM–CD

ALMC 719 pt 21 22 [tsɔ̃bra], pt 23 [ˈtsɔ̃bro]

b. CAMBA : N *chombes* ‘jambes/legs’ FPL (CletSM-D)

ALMC 1319 pt 21 [ˈtsɔ̃ba]

c. CAMPANA : N *campono* FSG ‘cloche/bell’ NoëlsCordat

d. LITANĬA : N *letonios* FPL ‘litanies’ NoëlsCordat

(4) /a+N/ Final stressed vowel = [o/ɔ̃] Old and modern nasalization

a. DE MANE : Adv. *demo* ‘demain/tomorrow’ NoëlsCordat CletSM, *démo* CletSM–C

ALMC 1443 pts 21 22 [demo], pt 23 [dɪmo]

ALMC 1444 pt 21 [lĩdɪmo]

b. *LĔVĀMEN : N *levon* ‘levain/yeast’ M NoëlsCordat

ALMC 1109 pts 21 22 23 [lɪvo]

(5) Monosyllables

(a) ANNU : N ‘année/year’ MSG *on* NoëlsCordat, PL *ons* CletSM–D SocConstPuy

ALMC 1415, 1417 pts 21 22 [o] [õ], 23 [?]

(b) MANU : N *mo* FSG ‘main/hand’ NoëlsCordat CletSM–ABCE, *mos* FPL NoëlsCordat SocConstPuy

ALMC 1324 pt 9 21 22 23 [mo]

(c) PANE : N *po* ‘pain/bread’ M NoëlsCordat

ALMC 1124 pt 9 21 22 23 [po]

Surrel (2022) too notes examples from the author Antoine Clet, which come from a modern play (ca. 1750), and in particular from witness D, which dates to 1836. We have seen in the examples that, with a century's difference, the language used by Antoine Clet does not differ from that of Cordat and we find therein all the characteristic features of the Occitan of the present–day Central Velay or Occitan–Vellavian, including this typical velarization/rounding which for us indicates nasalization in the northern Occitan written texts as well as an indigenous Occitan nasalization. The same velarization/rounding of /a/ + nasal in tonic syllables has been also noted by Roux (2020: 104) in Cordat; Roux observes that Cordat's language, which represents the southern part of northern Occitan, is conservative. We find here too the same cue of nasalization represented by velarization in his examples:

Cordat (17th c.)	Roux (2020)
<chon>	'champ/field'
<on souput>	'an sauput/(they) have known'
<pô queit>	'pan cuèit/ baked bread'
<sont el>	'sans–t–el/without him'
<de gron jaue>	'de grand jaug/of great joy'
<capitoni>	'capitani/captains'
<effon>	'enfant/child'
<bron>	'bram/cry'

These outcomes are attested in our data, as illustrated in the spectrograms and atlases examples provided above, and clearly show velarization/rounding and nasalization as a feature of Vivaro–Alpine Occitan in synchrony, as in Jaujac (Ardèche) [lo ʧʃɔ̃mbo] 'jambe/the leg' and [mɔ̃nʤa] (Spectrograms 8 and 9). The Ardèche data correspond to [sɞˈmɔnɔ] shown in Spectrogram 13 which represents the /a + N/ velarization of Auvergne Occitan. In modern varieties of Occitan Auvergnat, we have indeed found (in the Occitan Vellave,

Saint–Agrève – Ardèche) [sɞˈmɔnɔ] 'semaine/week' (an example we have taken from De Felice's sound archives, see Spectrogram 13), with rounding and velarization of A, among many other examples with the same categorization /A +N/ = [ɑ ɒ o /ɑ̃ ɒ̃ õ].

6. Conclusion

This paper has argued that northern Occitan (and particularly the Auvergnat varieties of Haute–Loire and Puy–De–Dôme) has demonstrably denasalized what were previously autochthonous nasal vowels . We have argued that this process of denasalization has left residual phonological cues of nasalization, as we have seen in /A + N/ sequences, i.e. a phonological velarization and rounding of A (/A+N/ = [o ɑ ɒ]/ [õ ɑ̃ ɒ̃] up to [aw ao/ ãw̃ ãõ]. This is a clear trace of phonological nasalization in the Occitan/FP transitional space. Our picture is supported by Auvergnat Occitan (Vellavian) texts from the 17th century: [on] ANNU and modern pt [o] or pt 22 [õ] ALMC 1415, where we attest to these two variants in nearby villages. Furthermore, we have shown that the raising of /o e/ +N to [u i/ũ ĩ] is also a phonological process triggered by nasality, which allows for the phonological reconstruction of nasalization in northern Occitan for mid vowels (after loss of the final nasals), such as in [bu] 'good' BONU, [saˈvu] 'soap' SAPONE, [muˈtu] 'sheep' MULTONE, [kaˈju] 'pig' –ONE, etc. These forms as we have seen can alternate in the transitional Oc/FP space synchronically with the nasalized variants [ũ ĩ] ([muˈtu]/ [muˈtũ]) and also with diphthongal forms as in Jaujac [bostõw̃] *batôn* 'cane', where [w̃] or [w] is what remains as a phonological trace from the lenited /N/. The treatment of /o e/ +N is not dissimilar from what is observed for FP in Auvergne–Rhône–Alpes, producing [i u/ũ ĩ], which are systematically nasalized in FP and denasalized in most cases in northern Occitan. However, intermediate stages are also found with a nasal coda [ũⁿ ĩⁿ]. Mid–open vowels are instead distinguished between a more FP–like form with diphthong as opposed to a more Occitan–form without a diphthong. This gives rise to mixted variants in FP/Oc areas [dj̃ẽⁿ] vs. [dĩⁿ]/ [dĩ] or [bwo]/[bu]. Intermediate forms such as these have also been found in Jaujac with oral denasalized /i N/ [vin] *vin* 'wine' VINU which alternate with a high nasal [t͡ʃĩ] CANE *chien* 'dog',

[ʤɔʁˈdĩ] *jardin* 'garden'. The pattern is frequent in Ardèche, where we observe for the same villages forms such as [vi/ve] for *vin* 'wine', where [i/e] forms indicate clearly a previous stage of nasalization of the high oral vowels (here VINU).

Unlike central Occitan, northern Occitan thus does have phonological nasalization and denasalization which is readily demonstrable in diachrony and synchrony. We therefore contend that it would not be accurate to continue to characterise nasal vowels as a distinguishing feature between FP and Occitan, even if we can recognize an Oc transitional zone that shows clear denasalization {o ɑ ɒ i u e}. However, we have shown many cues which indicate a recent denasalization after a previous nasalization. This is common to all transitional areas, such as Haute–Loire, which has maintained nasalization under the phonological raising of the mid–vowels /e o/ +N and backing of the low vowel in the stressed /a/ +N sequences. Furthermore, as we have seen, a cue of phonological nasalization is also betrayed by morpho–phonological alternations whereby the N nasal feature became an affix to express morphosyntactic properties, particularly in the SG vs. PL paradigms, to the extent that the nasal affix can replace in northern Occitan the sigmatic plural of central Occitan.

These findings contribute to the wider literature in at least two respects: we not only eschew vowel nasalization as a characteristic feature distinguishing FP and Oc, but we have also shown the advantage of triangulating very disparate datasets, from a number of paradigms (acoustic phonetics, phonology, dialectology, sociolinguistics, historical linguistics), which we have brought to bear on our primary aims. We would encourage further collaboration of this type in subsequent work on endangered and under-described language varieties.

BIBLIOGRAPHY

ALAL = POTTE, Jean–Claude (1975–1992), Atlas linguistique et ethnographique de l'Auvergne et du Limousin, 3 vol. Paris: CNRS.

ALF = GILLIERON, Jules and Edmont, Edmond (1902–1910), *Atlas linguistique de la France*. Paris: Honoré Champion. Online http://lig-tdcge.imag.fr/cartodialect5/#/ .

ALG = SEGUY, Jean (1954-1973), Atlas linguistique et ethnographique de la Gascogne, 1965-1985 (1954-1973). Paris: CNRS.

ALJA = TUAILLON, Gaston and Martin, Jean–Baptiste (1971–1981), *Atlas linguistique et ethnographique du Jura et des Alpes du Nord*. 4 vol., Paris : CNRS.

ALLy = GARDETTE, Pierre (1950–1976). *Atlas linguistique et ethnographique du Lyonnais*. 5 Volumes. Paris : CNRS. Online www.ortolang.fr .

ALMC = NAUTON, Pierre (1963, 1972, 1976, 1977), *Atlas Linguistique et Ethnographique du Massif Central*. 4 Volumes. Paris : CNRS.

ATF = DEVAUX, André (1935), *Atlas linguistique des Terres Froides*, Antonin DURAFFOUR & Pierre GARDETTE (posthumous work published by). Lyon : Bibliothèque de la faculté catholique de Lyon.

ALLIERES, Jacques (2001), *Manuel de Linguistique Romane*. Paris : Honoré Champion Éditeur.

ASCOLI, Graziadio Isaia [1873] (1878), Schizzi franco–provenzali. *Archivio Glottologico Italiano* 3 : 61–120.

BABEL, Molly. (2008). The phonetic and phonological effects of moribundity. *Penn Working Papers* 14(2). 25-34.

BEC, Pierre (1973 [1983]), *Manuel pratique d'occitan moderne*. Paris : A. & J. Picard.

BEC, Pierre (1970–71), *Manuel pratique de philologie romane*. Vol. II. *Français, roumain, sarde, rhéto–frioulan, francoprovençal, dalmate*. Paris : A. & J. Picard.

BEC, Pierre (1970), *L'occitan. Manuel pratique de philologie romane*. Volume I. Paris : Éditions A. & J. Picard. 395–462.

BEC, Pierre (1963 [1986]), La langue occitane. Que–sais–je ? Paris: Presses Universitaires de France (PUF).

BERT, Michel & James COSTA (2014), What Counts as a Linguistic Border, for Whom and with What Implications? Exploring Occitan and Francoprovençal in Rhône–Alpes, France. 186–205.

BERT, Michel & Jean–Baptiste MARTIN (2013), Le Francoprovençal. In Georg KREMNITZ (ed.). Histoire sociale des langues de France. 489–501. Rennes : Presses Universitaires de Rennes (PUR).

BERT, Michel, COSTA, James & Jean–Baptiste MARTIN (2009), Étude FORA: francoprovençal et occitan en Rhône–Alpes. Lyon: Institut Pierre Gardette, INRP, ICAR, DDL.

BERT, Michel (2001), *Rencontre de langues et francisation: L'exemple du Pilat*. PhD dissertation. Lyon: Université Lumière Lyon 2.

BONNAUD, Pierre (2006), *Grammaire Générale de l'Auvergnat à l'usage des arvernisants*. Chamalières: Cercle Terre d'Auvergne.

BONNAUD, Pierre (1974), *Nouvelle grammaire auvergnate*. Clermont–Ferrand Cercle Occitan d'Auvergne–Auvernha Tarra d'Oc.

BOUVIER, Jean–Claude (2003), L'occitan en Provence: le dialecte provençal, ses limites et sesvariétés. In Jean–Claude BOUVIER (ed.). *Espaces du langage: géolinguistique, toponymie, culture de l'oral et de l'écrit*. Claude Mauron & Jean–Noël Pelen. 11–25. Aix en Provence: Université de Provence.

BOUVIER, Jean–Claude (1976), *Les parlers provençaux de la Drôme. Étude de géographie phonétique*. Paris: Klincksieck.

BOUVIER, Jean–Claude (1973), Les paysans drômois devant les parlers locaux. *Ethnologie française* 3/3–4. 229–234.

BOUVIER, Jean–Claude (1970), Le vocabulaire franco–provençal dans la Drôme provençale. In *Quatrième Congrès de langue et littérature d'oc et d'études franco–provençales, Avignon.* 7–13 September 1964. 455–469. Rodez: Revue de Langue et littérature d'oc.

BOUVIER, Jean–Claude (1966), Quelques aspects de la diversité phonétique dans la Drôme provençale : remarques sur la nasalisation. *Revue de linguistique romane* 30. 122–133.

BURGER, Michel (1964), La nasalisation spontanée dans les dialectes de la plaine vaudoise et fribourgeoise, *Revue de linguistique romane*, 28.111–112. 290–305.

CALVET, Maurice (1960), *Le parler occitan de Saint–Victor en Ardèche : lexique et locutions selon les principaux thèmes de la vie rurale d'autrefois*. Université de Grenoble (Masters thesis).

CALVET, M. (1969), Le système phonétique et phonologique du parler provençal de Saint-Victor en Vivarais. Grenoble : Faculté des lettres et sciences humaines de Grenoble.

CHAMBON, Jean–Pierre (2012), Histoire de l'occitan (et du français) dans le domaine auvergnat : progrès récents en linguistique et en philologie (bilan et bibliographie). *Bulletin historique et scientifique de l'Auvergne*, CXIII/2. 106–123.

CHAMBON, Jean–Pierre (2004), Les centres urbains directeurs du Midi dans la francisation de l'espace occitan et leurs zones d'influence : Esquisse d'une synthèse cartographique. *Revue de Linguistique Romane* (RLiR) 68: 5–18.

CHAMBON, Jean-Pierre & Philippe OLIVIER (2000), L'histoire linguistique de l'Auvergne et du Velay : notes pour une synthèse provisoire. *Travaux de linguistique et de philologie* 38: 83–153.

CHAMPAILLER = CHARPIGNY, Florence, GRENOUILLER, Anne–Marie, MARTIN, Jean–Baptiste (1986), *Marius Champailler, paysan de Pélussin*. Aix–en–Provence: Edisud/éditions du CNRS Parlers et Cultures de France.

CHONG, Adam & Jonathan R. KASSTAN (2022). Acoustic characteristics of fricatives in Francoprovençal. *Journal of the International Phonetic Association*. FirstView. 1-34.

CLET, Antoine (1757), « Monsieur Lambert », 1842, *Histoire poétique et littéraire de l'ancien Velay*, Francisque Mandet. Paris: Rozier. 299–346.

CORDAT, Natalis (1876), « Noëls vellaves », 1631–1648. Publiés avec introduction et notes par Jean–Baptiste Payrard. Le puy en Velay: Freydier.

DAUZAT, Albert (1944), *La géographie linguistique*. Paris: Flammarion.

DAUZAT, Albert (1938), Géographie phonétique de la Basse Auvergne. *Revue de linguistique romane* 14: 1–210.

DAUZAT, Albert (1897), *Études linguistiques sur la Basse–Auvergne: Phonétique historique du patois de Vinzelles (Puy–de–Dôme)*. Paris : Alean.

DE FELICE, Théodore (1983), Le patois de la zone d'implantation protestante du nord–est de la Haute–Loire. Paris/Genève: Champion–Slatkine.

DE FELICE, Théodore (1989), Nouvelles recherches sur le patois de la zone d'implantation protestante du nord–est de la Haute–Loire. Paris/Genève: Champion–Slatkine.

DEVAUX, André (1892), *Essai sur la langue vulgaire du Dauphiné septentrional au Moyen Âge*, Paris, H. Welter.

DI CARO, Alexandre (forthcoming), Description du parler nord–occitan de Sainte–Sigolène. Une incursion linguistique dans l'occitan vivaro–vellave à travers des enquêtes de terrain. Mémoire de Master / Master's thesis. Supervisor : Michela Russo, Linguistics Dep., UJM Lyon 3, France.

DORIAN, Nancy C. (ed). (1989). *Investigating Obsolescence*. Cambridge: Cambridge University Press.

DORIAN, Nancy C. (1982). Defining the Speech Community to Include its Working Margins. In Suzanne Romaine (ed.). *Sociolinguistic Variation in Speech Communities*. 25–33. London: Edward Arnold.

DUFAUD, Joannès (1981–1988), *Chansons anciennes du Haut–Vivarais*. 4 vol. Davézieux: L. Volozan.

DUFAUD, Joannès (1986), *L'occitan Nord–Vivarais. Région de La Louvesc*. Davézieux: L. Volozan.

DUFAUD, Joannès (1998), *Dictionnaire Français – Nord–Occitan. Nord du Vivarais et du Velay*. Saint–Julien–Molin–Molette: Jean–Pierre Huguet.

DURAFFOUR, Antonin (1932), *Phénomènes généraux d'évolution phonétique dans les parlers francoprovençaux d'après le parler de Vaux–en–Bugey (Ain)*. Grenoble : Institut de phonétique.

DUFAUD, J. (2004), *Des mots à la phrase occitane, complément à l'occitan nord–vivarais* (*Parlarem en Vivarés*). Saint–Julien–Molin–Molette: Jean–Pierre Huguet.

DUVERT–CHENEBERT, *Clara* (2020–21), L'occitan en Ardèche: l'exemple du parler de Jaujac. Mémoire de Master 1/ Master's thesis (1st year). Supervisor : Michela Russo, Linguistics Dep., UJM Lyon 3, France.

ESCOFFIER, Simone (1958a), *La rencontre de la langue d'oïl, de la langue d'oc et du francoprovençal entre Loire et Allier. Limites phonétiques et morphologiques. Publications of the Institute of Romance Linguistics of Lyon, vol. 11;* Paris: Belles–Lettres.

ESCOFFIER, Simone (1958b), *Remarques sur le lexique d'une zone marginale aux confins de la langue d'oïl, de la langue d'oc et du francoprovençal. Publications of the Institute of Romance Linguistics of Lyon*, vol. 12 ; Paris: Belles–Lettres.

FEW = Wartburg, Walther von, *et al*., 1922–2002. *Französisches Etymologisches Wörterbuch. Eine darstellung des galloromanischen sprachschatzes*, 25 vol., Bonn/Heidelberg/Leipzig-Berlin/Bâle, Klopp/Winter/Teubner/Zbinden.Garnier, Quentin (2020). Le vivaro–alpin: progrès d'une définition. *Géolinguistique* 20.OpenAccess https://journals.openedition.org/geolinguistique/1992 .

GARDETTE, Pierre (1983), *Études de géographie linguistique*. Paris : Klincksieck.

GARDETTE, Pierre (1983), Deux itinéraires des invasions linguistiques dans le domaine provençal. In Id., *Études de géographie linguistique*. Paris: Klincksieck. 615–630.

GARDETTE, Pierre (1973), Frontières linguistiques et limites intérieures en lyonnais d'après l'Ally. In Georges Straka and Gardette, Pierre (eds.). *Les dialectes romans de France*. 141–171. Paris: CNRS.

GARDETTE, Pierre (1970), Nécrologie de Pierre Nauton. *Revue de linguistique romane* 34: 446–447.

GARDETTE, Pierre (1964), En marge des atlas linguistiques du Lyonnais, du Massif Central, du francoprovençal du Centre. Les influences des parlers provençaux sur les parlers francoprovençaux. *Revue de linguistique romane* 28: 69–81.

GARDETTE, Pierre (1957), Le Lyonnais et le Massif Central d'après les atlas linguistiques régionaux (à propos de la publication du premier volume de l'*Atlas linguistique et ethnographique du Massif Central* de P. Nauton). *Revue de linguistique romane* 21, 209–230.

GARDETTE, Pierre. (1955), Deux itinéraires des invasions linguistiques dans le domaine provençal. *Revue de linguistique romane* 19: 183–198.

GARDETTE, Pierre (1941), *Géographie phonétique du Forez*. Macon: Imprimerie Protat Frères.

GARDETTE, Pierre (1939), Limites du provençal au pays du Forez. In *Mélanges en hommage à* Antonin Duraffour. *Romanica Helvetica*, 14. 22–36.

GIRARD, Augustin–Marie (1925), *Grammaire vellave*. Le Puy–en–Velay: Imp. l'Avenir.

GLESSGEN, Martin–Dietrich et Pfister, Max (1995), Okzitanische Koine/La koinè occitane. In Günter Holtus, Michael Metzeltin, Christian Schimtt, *Lexikon der Romanistischen Linguistik* (LRL) II/2. 406–412.

GLESSGEN, Martin–Dietrich & Max PFISTER (1995), Okzitanische Skriptaformen. I. Limousin / Périgord, In Günter Holtus, Michael Metzeltin, Christian Schimtt, *Lexikon der Romanistischen Linguistik* (LRL) II/2: 412–419. Tübingen: Niemeyer.

GRANGE, Didier (2021), Un parler roman : le patois de Sainte-Sigolène (ebook marraire.com).

GRANGE, Didier (2008), Le lexique descriptif occitan-français du vivaro–alpin au nord du Velay et du Vivarais (ebook marraire.com)

GOURGAUD, Yves (1976), *Natalis Cordat. Nadaus 1632–1648*. Le Puy–en–Velay, Section 43 de l'Institut d'études occitanes.

GPFP = DURAFFOUR, Antonin (1969), *Glossaire des patois francoprovençaux*, published by Laure Malapert, Marguerite Gonon, Pierre Gardette dir. Paris : CNRS.

GRANDGENT, Charles Hall (1905), *An outline of the phonology and morphology of old Provencal*, Boston/New York/Chicago, D. C. Heath & Co.

GREUB, Yan & Jean–Pierre CHAMBON (2009), Histoire des dialectes dans la Romania : Galloromania. *In* Ernst, Gerhard, Glessgen, Martin–Dietrich, Schmitt, Christian and Schweickard, Wolfgang (eds.). *Romanische Sprachgeschichte/Histoire linguistique de la Romania. Ein internationales Handbuch zur Geschichte der romanischen Sprachen/Manuel international d'histoire linguistique de la Romania*. Berlin/New York: Mouton De Gruyter. Vol. .: 2499–2520.

GUITER, Henri (1980), Limites linguistiques du Velay méridional. *Bulletin historique, scientifique, littéraire, artistique et agricole de la Société académique du Puy et de la Haute–Loire* 56. 109–116.

HASSELROT, Bengt (1974), Adieu au francoprovençal. *Revue de Linguistique Romane* 38. 265–275.

HASSELROT, Bengt (1934), Le francoprovençal se compose–t–il de deux groupes principaux, un septentrional et un méridional?, *Studia Neophilologica* 7. 1–17.

HORNSBY, David. (2006), *Redefining Regional French*. Oxford: Legenda.

KASSTAN, Jonathan R. (2015a), Variation and change in Francoprovençal: A study of an emerging linguistic norm. Unpublished PhD thesis, University of Kent.

KASSTAN, Jonathan R. (2015b), Illustrations of the IPA: Lyonnais (Francoprovençal). *Journal of the International Phonetic Association*. 45(3). 349-355.

KASSTAN, Jonathan R. (2019), Emergent sociolinguistic variation in severe language endangerment. *Language in Society* 48(5). 685-720.

KASSTAN, Jonathan R. & Michela RUSSO. (2021), Maintenance in shift: On nasalization in trasitional Francoprovençal & Occitan areas. Conference presented at the *Journée annuelle de la Société Linguistique de Paris*, 12th June 2021.

KASSTAN, Jonathan R. (2022), A variationist analysis of the subjunctive mood in a sociolinguistic corpus of spoken Francoprovençal. LINX *Revue des linguistes de l'Université Paris Ouest Nanterre La Defésense*. Thematic Issue : Russo, Michela (ed.), *Nouvelles perspectives sur les langues romanes à l'interface de la grammaire. Regards sur la variation syntaxique, morphosyntaxique et phonologique* 84(1). 1-23.

Online https://journals.openedition.org/linx/8593

KRISTOL, Andrés (2016), Francoprovençal. In Adam LEDGEWAY & Martin MAIDEN (eds.). *The Oxford Guide to the Romance Languages*. Oxford: Oxford University Press. 350–362.

LAFONT, Robert (1991), Okzitanisch: Interne Sprachgeschichte I. Grammatik. In: Holtus, Günter / Metzeltin, Michael / Schmitt, Christian *Lexikon der Romanistischen Linguistik* (LRL) V/2, 1–18. Tübingen: Niemeyer.

LAFONT, Robert (1983), *Éléments de Phonétique de l'Occitan*. Valderiès: Vent Terral.

LAGUEUNIERE, France (1983), *Études de géographie linguistique dans l'arrondissement de Bellac (Haute–Loire): phonétique historique et phonologie*. PhD thesis. Paris: Université de Paris–Sorbonne.

LODGE, R. Anthony (1995), Okzitanische Skriptaformen II. Auvergne. In: Holtus, Günter / Metzeltin, Michael / Schmitt, Christian *Lexikon der Romanistischen Linguistik* (LRL) II/2. 420–424.

MARTEL, Philippe (1983), L'espandi dialectau alpenc: assag de descripcion. *Novel Temps* 21. 4–36.

MARTEL, Philippe & Franc BRONZAT (1977), L'aira dialectala 'vivaròupenca'. *Quaserns de Linguistica Occitana* (Clarmont, IEO Puei de Doma), 6. 38–48.

MARTIN, Jean–Baptiste (1973), État actuel du bilinguisme à Yssingeaux (Haute–Loire). *Ethnologie française* 3/3–4: 309–316.

MARTIN, Jean–Baptiste (1974), Les textes anciens de la Haute–Loire. In Moignet, Gérard and Lassalle, Roger (eds.). *Actes du 5e Congrès international de langue et littérature d'oc et d'études* francoprovençales, Nice, 6–12 September 1967. 353–357. Nice: Publications de la Faculté des lettres et des sciences humaines.

MARTIN, Jean–Baptiste (1979), La limite entre l'occitan et le francoprovençal dans le Pilat. *Études foréziennes* 10 : 75–88.

MARTIN, Jean–Baptiste (1990), Le francoprovençal. In Metzeltin, Michael & Christian Schmitt (eds.). *Lexikon der Romanistischen Linguistik* (LRL), vol. 5.1. 671–685. Tübingen: Max Niemeyer.

MARTIN, Jean–Baptiste (1991), Nommer la langue pour les linguistes et pour les locuteurs : l'exemple du francoprovençal. In Bouvier, Jean–Claude (ed.). *Actes du colloque : Les français et leurs langues,* Montpellier 5–7 September 1988. 495–501. Université de Provence, Aix–en–Provence.

MARTIN, Jean–Baptiste (1993), Découpage linguistique. Domaine francoprovençal et région Rhône-Alpes. In Paul BACOT & Philippe DUJARDIN (eds.). *Sociologie du découpage et de ses usages politiques*. Lyon: Centre de Politologie de Lyon. 93–112.

MARTIN, Jean–Baptiste (1997), *Le parler occitan d'Yssingeaux (Haute–Loire)*. Yssingeaux : Histoire et Patrimoine.

MARTIN, Jean–Baptiste (2021), *La langue francoprovençale*. Gleizé : Édition du Poutan.

MEIRELES, Vanessa (2014), *Analyse phonologique et métrique des glides etdiphtongues en portugais brésilien*. PhD Thesis : University of Paris.

MOULIN, Bernard (2006), *Grammaire occitane: le parler bas–vivarois de la région d'Aubenas*, Saint–Étienne de Fontbellon: Section Vivaroise de l'Inst. d'Études Occitanes.

NAUTON, Pierre (1974), *Géographie Phonétique de la Haute–Loire*. Publications de l'Institut de Linguistique Romane de Lyon (published by Jean–Baptiste Martin). Vol. 29. Paris: Société d'Edition: Les Belles Lettres.

NAUTON, Pierre (1966), Occlusives intervocaliques dans la région amphizone de l'Atlas linguistique du Massif Central. *Travaux de linguistique et de littérature* 4 : 357–369.

NAUTON, Pierre (1956), Atlas linguistique et ethnographique du Massif Central (domaine, réseau, questionnaire, but). *Revue de linguistique romane* 20: 41–65.

NAUTON, Pierre (1948), *Le Patois de Saugues (Haute–Loire). Aperçu linguistique, terminologie rurale, littérature orale*. Clermont–Ferrand: Faculté des Lettres de l'Université de Clermont.

NAUTON, Pierre (1952), Une butte–témoin linguistique: le patois des protestants du Velay. *Mélanges de linguistique et de littérature romanes offerts à Mario Roques par ses amis, ses collègues et ses anciens élèves de France et de l'étranger*, Paris, Didier, 185–193.

OLIVIERI, Michèle & Patrick SAUZET (2016). Occitan. In Adam LEDGEWAY & Martin MAIDEN (eds.). *The Oxford Guide to the Romance Languages*, Oxford: Oxford University Press. 319–349.

PADEN, William D. (1998), *An Introduction to Old Occitan*. New York: Modern Language Association of America.

PELLEGRINI, Giovanni Batista. (1965), *Appunti di grammatica storica del provenzale*. Pisa: Libreria Goliardica.

PERRE, Didier. (2004), Les mélodies des noëls de Natalis Cordat (env. 1610–1663). Premiers résultats. *Cahiers de la Haute–Loire*, 163–193.

PFISTER, Max. (2002), L'area galloromanza. In Boitani, Piero, Mancini, Mario and Vàrvaro, Alberto (eds.). *Lo spazio letterario del Medioevo. II. Il Medioevo volgare.* Vol. II. *La circolazione del testo*. Roma: Salerno Editrice. 13–96.

PFISTER, Max. (1958), Beiträge zur altprovenzalischen Grammatik. *Vox Romanica* 17: 281–363.

QUINT, Nicolas. (1999), Le parler occitan ardéchois d'Albon, canton de Saint–Pierreville, Ardèche. Description d'un parler alpin vivaro–vellave du boutiérot moyen. Paris: L'Harmattan.

RAVIER, Xavier. (1991), Okzitanisch: Areallinguistik/Les aires linguistiques. In Holtus, Günter, Metzeltin, Michael and Schmitt, Christian (eds.). *Lexikon der Romanistischen Linguistik* (LRL) V/2: 80–105.

RIDEAU, Jean–Yve. (2018), *Tresor des parlers occitans du Velay oriental et du sud Forez*.

Online : http://bartavel.com/tresor/1.Introduction.pdf

ROHLFS, Gerhard. (1970), *Le gascon. Études de philologie pyrénéenne*. Tübingen: Niemeyer.

RONJAT, Jules. (1930–1941), *Grammaire (h)istorique des parlers provençaux modernes*. 4 vol. Montpellier: Société des Langues Romanes.

ROSTAING, Charles. (1951), Phénomènes de nasalisation en provençal. *Mélanges de linguistique offerts a Albert Dauzat par ses élevés et ses amis*. Paris: Editions d'Artrey. 275–278.

ROUX, Jean. (2020), *De la renaissance d'une langue occitane littéraire en Auvergne au début du XXe siècle, au travers des œuvres de Bénezet Vidal et Henri Gilbert*. Thèse de doctorat. Université Paul Valéry Montpellier 3.

ROUX, Jean. (2015), Huit siècles de littérature occitane en Auvergne et Velay ; Morceaux choisis, Régionales, Lyon : EMCC (= *Édition* média conseil communication).

RUSSO, Michela, CURCHI, Corina & Jimmy JOSSERON (2021). La nasalisation dans le nord–occitan de la Drôme. Conference presented at the 5th edition of the Congress *International Langue et Territoire*. U. Montpellier Paul–Valéry 14–20 June 2021. https://let2021.sciencesconf.org/

RUSSO, Michela (2021). Les limites du Croissant dans l'Est francoprovençal (Forez) et dans le Nord occitan (Auvergne). In Louise Esher, Maximilien Guerin, Nicolas Quint, Michela Russo (eds.). *Le Croissant linguistique entre oc, oïl et franco–provençal, Des mots à la grammaire, des parlers aux aires*. Paris : L'Harmattan, 65–106.

SAMPSON, Rodney. (1999), *Nasal Vowel Evolution in Romance*. Oxford: Oxford University Press.

SIMIAND, Pierre. (1991), L'Ardèche dialectale. *L'Ardèche*. Aubenas: Curendera, 459–466.

OZAWA, Hiroshi. (2007), La nasalisation en occitan d'après l'Atlas Linguistique de la France. *Revue de linguistique romane* 71(283–84). 393–410.

STRAKA, Georges. (1955), Remarques sur les voyelles nasales, leur origine et leur évolution en français, *Revue de Linguistique Romane* 19.75–76, 245–274.

STRAKA, Georges. (1979), Remarques sur les voyelles nasales, leur origine et leur évolution en francais. *Les sons et les mots,* Paris: Klincksieck, 501–531.

SUMIEN, Domergue. (2009), Classificacion dei dialèctes occitans. *Linguistica occitana* 7 : 1–56. Online http://linguistica-oc.com/?page_id=244

SUMIEN, Domergue. (2006), La standardisation pluricentrique de l'Occitan: Nouvel enjeu sociolinguistique, développement du lexique et de la morphologie. Turnhout, Belgium: Brepols.

TAVERDET, Gérard. (1985), *Les noms de lieux de la Haute-Loire*. Fontaine lès Dijon: A.B.D.O.

TOURTOULON, Charles & Octavien BRINGUIER. (1875), Etude sur la limite géographique de la langue d'oc et de la langue d'oïl (1875) (avec une carte), Paris, Imprimerie Nationale, [réé. IEO dau Lemosin – Chamin de Sent Jaume, Masseret – Meuzac, 2004].

TUAILLON, Gaston (1964), Limite nord du provençal à l'est du Rhône. *Revue de Linguistique Romane*, 28. 129–143.

TUAILLON, Gaston (1972), Le franco–provençal: progrès d'une définition'. *Travaux de linguistique et de littérature* 10: 1. 293–339.

TUAILLON, Gaston. (1990), Méditations sur les langues régionales en Dauphiné. *Inventer le Monde. Les Rhônalpins et leurs langages*. Grenoble: Musée Dauphinois. 9–21.

TUAILLON Gaston. (2007), *Le francoprovençal. Définition et délimitation. Phénomènes remarquables*. Tome 1. Quart (Vallée d'Aoste): *Musumeci*.

VEŸ, Eugène. (1911), *Le dialecte de Saint-Etienne au XVII^e siècle*. Paris: Honoré Champion.

WHEELER, Max W. (1988), Occitan. In Martin HARRIS & Nigel VINCENT (eds.). *The Romance Languages*. London: Croom Helm. 246–278.

WOLF, Lothar & Sarcher WALBURGA. (2003), Autour de l'indexation électronique de l'ALMC. Identification lexicale et étymologique : État des travaux ». In Jacques GOURC, François PIL & Jean-Claude BOUVIER (éd.), *Sempre los camps auràn segadas resurgantas. Mélanges offerts au professeur Xavier Ravier*, Toulouse, Presses universitaires du Midi. 299–302.

WOLF, Lothar. (1988), Les fichiers de Pierre Nauton. In *Espaces romans. Études de dialectologie et de géolinguistique offertes à Gaston Tuaillon*. Grenoble: Ellug. 279–281.

WOLF, Lothar (1978), Aspects de l'influence du français sur les parlers de la Haute–Loire et du Massif Central. In Robert LAFONT (ed.), *Mélanges de philologie romane offerts à Charles Camproux*. vol. 2. Montpellier : C.E.O. (*Centre d'études occitanes*). 979–986.

WOLF, Lothar. (1970), Remarques sur la pénétration du français dans le Massif Central. *Revue de linguistique romane* 34. 306–314.

Wolf, Lothar. (1968), *Sprachgeographische Untersuchungen zu den Bezeichnungen für Haustiere im Massif Central. Versuch einer Interpretation von Sprachkarten*. Tübingen: Niemeyer.

Russo, Michela
UJML 3 & SFL CNRS
Université de Paris 8 (FR)
michela.russo@cnrs.fr

Kasstan, Jonathan
School of Humanities
University of Westminster
j.kasstan@westminster.ac.uk

Appendix I

List of target words (VCF corpus)

	Target	Item	Part of speech	Gloss
1.	/i/+N	cendres	N	'embers'
2.	/i/+N	chien	N	'dog'
3.	/i/+N	cinq	N	'five'
4.	/i/+N	manger	V	'to eat'
5.	/i/+N	mangez	V	'eat' (2.PL)
6.	/i/+N	mangez	V	'eat' (1.PL)
7.	/i/+N	un	DET	'a'
8.	/i/+N	un enfant	DET+N	'a child'

Outils numériques, bases de données orales et implication des locuteurs ou du public (langues du Croissant, occitan, rromani)

Chapitre 6

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

Nicolas Quint[1]
LLACAN – UMR 8135 (CNRS / INALCO / EPHE)

Abstract

The area of the Linguistic Crescent extends across the northern fringes of the French Massif Central, and it derives its name from the half-moon shape it takes on maps. The Gallo-Romance varieties traditionally spoken throughout this area simultaneously display features considered typical of the three following languages: Occitan, French (and other

[1] Je tiens à remercier ici les nombreux informateurs locuteurs des diverses variétés du Croissant mentionnées dans cet article. Merci aussi à Maximilien Guérin pour son précieux concours bibliographique et à Guylaine Brun-Trigaud pour ses cartes. Je suis le seul responsable des erreurs qui pourraient subsister.

Ce travail s'insère dans les trois projets suivants :

(i) "Oc/Oïl : textes, identité et contact de langues aux confins gallo-romans", financé par le dispositif Émergence(s) de la Ville de Paris ;

(ii) ANR-17-CE27-0001-01 (Projet « Les parlers du Croissant : une approche multidisciplinaire du contact oc-oïl ») et (iii) ANR-10-LABX-0083 (programme « Investissements d'Avenir », Labex EFL, Axe 3, Opération VC2 - « Au cœur de la Gallo-Romania : caractérisation linguistique et environnementale d'une aire de transition »), tous deux gérés par l'Agence Nationale de la Recherche.

Il contribue à l'IdEx Université de Paris - ANR-18-IDEX-0001.

Oilic varieties), and – to a lesser extent – Francoprovençal. To date, mainly due to their mixed characteristics, the Crescent varieties remain understudied.

In this contribution, I provide a general overview of the Linguistic Crescent. After introducing the area and its scientific interest (section 1), I discuss in section 2 its main characteristics (geographical extension, diatopic variation, and sociolinguistic situation). Then, in section 3, I give an account of linguistic research on the Gallo-Romance varieties spoken in the Crescent, with special emphasis on the collective projects that have dealt with this topic in the course of the last decade. Section 4 takes stock of the various deliverables and other results that the researchers involved in these projects have been able to achieve. Section 5 is devoted to the future of the Crescent varieties. It stresses that, although these varieties are highly vulnerable and actually on the verge of extinction, their high scientific and cultural value justifies that they should be carefully and thoroughly studied.

1. Introduction[2]

Située aux franges nord du Massif Central, 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 nommée à cause de sa forme géographique évoquant une demi-lune) est constituée d'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 classer, les parlers du Croissant ont fait l'objet de moins d'études que la plupart des langues régionales traditionnellement pratiquées en France et ils comptent aujourd'hui encore parmi les plus méconnues des variétés gallo-romanes.

Or ces parlers, justement du fait de leur caractère mixte et intermédiaire, constituent précisément un patrimoine aussi précieux

[2] Liste des abréviations : ANR = Agence Nationale de la Recherche ; DGLFLF = Délégation Générale à la Langue Française et aux Langues de France ; 3DMA = 3 degrés médians d'aperture ; F = féminin ; lit. = traduction littérale ; M = masculin ; OLL = Occitan Languedocien Littéraire ; NTR = neutre ; PL = pluriel ; S = sujet ; SG = singulier, TAL = Traitement Automatique des Langues.

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 75 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 contribution, 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 consacrées aux 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 ferai le point sur les résultats issus de ces projets et j'insisterai en particulier sur la façon dont les linguistes et les locuteurs participant à de telles 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.

2. Les parlers du Croissant : une brève présentation

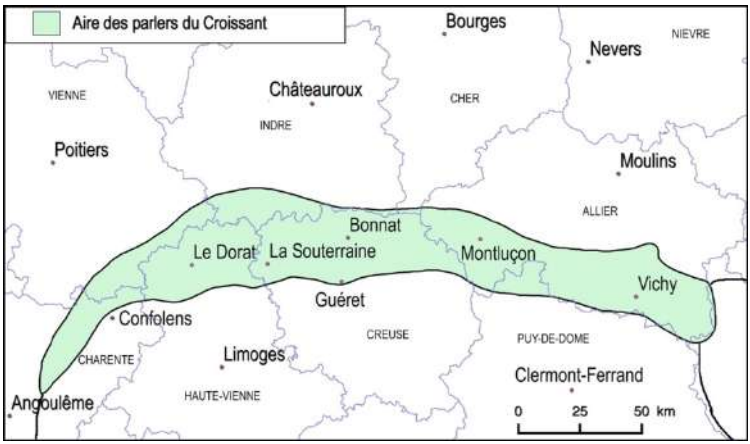
2.1. Situation géographique et délimitation

D'un point de vue géographique, le Croissant linguistique dessine une demi-lune aux franges nord du Massif Central (Carte 1). Il

s'étend sur environ 300 km d'est en ouest et atteint 40 km d'épaisseur (du nord au sud) en son centre (du nord de la Creuse au sud de l'Indre). Il est cependant beaucoup plus mince en Charente (pas plus de 5 à 10 km de large).

Carte 1. Le Croissant linguistique en France [auteur : Guylaine Brun-Trigaud]

Le Croissant s'étend sur trois régions (Centre, Nouvelle-Aquitaine et Auvergne-Rhône-Alpes) et au moins sept départements français (Carte 2), à savoir d'est en ouest : (i) le tiers sud de l'Allier, (ii) quelques communes de l'extrême-nord du Puy-de-Dôme, (iii) le tiers nord de la Creuse, (iv) treize communes de l'extrême-sud-ouest de l'Indre (Quint, Guérin & Brun-Trigaud : à paraître), (v) le quart nord de la Haute-Vienne, (vi) quelques communes du sud-est de la Vienne et (vii) une bande de territoire courant au centre nord de la Charente (Quint 2021, 2022).



Carte 2. Le Croissant linguistique (détail) [auteur : Guylaine Brun-Trigaud]

L'existence de cet ensemble linguistique est essentiellement justifiée par l'existence de deux isoglosses (Quint 2021, Brun-Trigaud 1992) :

- (i) au sud, les parlers du Croissant se distinguent des parlers typiquement occitans (auvergnat à l'est et limousin à l'ouest) par l'absence de distinction de timbre pour les voyelles post-toniques. Ainsi, la plupart des parlers occitans pratiqués au sud du Croissant distinguent au moins deux voyelles après l'accent tonique :

(Ex1) occitan languedocien littéraire (OLL)

*femn**a*** /'fenn**o**/ 'femme' vs. *òm**e*** /'ɔm**e**/ 'homme'.

En revanche, dans le Croissant, les voyelles atones sont toutes réalisées [ə] (sud du Croissant (Ex2)) ou totalement amuïes (nord du Croissant (Ex3)).

(Ex2) Gartempe (Creuse)

/'fan**ə**/ 'femme', /'ɔm**ə**/ 'homme'.

(Ex3) La Châtre-Langlin (Indre)

/fãn/ 'femme', /um/ 'homme'.

- (ii) au nord, les parlers du Croissant se distinguent des parlers d'oïl (d'est en ouest : bourbonnais d'oïl, berrichon et poitevin-saintongeais) par le fait que, dans la quasi-totalité des parlers croissantins, l'infinitif et le participe passé des verbes du 1er groupe ont une désinence comportant les voyelles /a/ ou /ɑ:/. Cette voyelle

ouverte, héritée des terminaisons latines –ĀRE (infinitif) et –ĀTUM (M.) / –ĀTAM (F.) (participe passé) se retrouve dans presque tous les parlers occitans (Ex4) :

(Ex4) occitan languedocien littéraire (OLL)
*cant**a**r* /kanˈt**a**/ 'chanter' (infinitif).
*cant**a**t* (M.) /kanˈt**a**t/ 'chanté', *cant**a**da* (F.) /kanˈt**a**do/ chantée'.

En revanche, les parlers d'oïl ont généralement des voyelles antérieures plus fermées (cf. la réalisation /ʃ̃ãˈt**e**/ commune à l'infinitif *chanter* et au participe passé *chanté/e* en français standard) pour ces mêmes désinences.

Ces deux isoglosses soulignent bien le caractère intermédiaire des parlers du Croissant, qui possèdent à la fois des caractéristiques typiques de l'occitan, en particulier sur le plan morphologique (cf. les désinences d'infinitif et de participe passé[3]) et des langues d'oïl, en particulier sur le plan phonologique (cf. la réduction des voyelles atones post-toniques)[4].

Historiquement, il est vraisemblable que le Croissant remontait autrefois davantage au nord et que, du fait de la domination croissante du français, il est progressivement descendu vers le sud. Il est donc très probable que les parlers croissantins actuels ont présenté davantage de caractéristiques occitanes dans le passé et qu'on peut les décrire, en prenant en compte la diachronie, comme des variétés occitanes francisées. Quoiqu'il en soit, en synchronie, c'est bien la coexistence de traits occitans, oïliques et (dans une moindre mesure) francoprovençaux qui fait la spécificité des variétés du Croissant (Guérin 2020a).

[3] C'est généralement la morphologie, de type plus nettement occitan chez les parlers du Croissant, qui a été prise en compte pour tracer des limites entre variétés croissantines et oïliques (cf. Terracher 1914).

[4] Pour des raisons de clarté de l'énoncé et d'espace disponible, je ne parlerai pas ici de la troisième langue romane impliquée dans l'existence du Croissant, à savoir le francoprovençal. En effet, l'influence francoprovençale ne s'exerce que sur une partie du Croissant (en l'occurrence l'extrême-est de l'Allier) tandis que la coexistence de traits occitans et oïliques est une caractéristique de l'ensemble de l'aire linguistique croissantine. Pour plus de détails sur les traits francoprovençaux dans le Croissant oriental, cf. Maurer-Cecchini (2023 : 237-246), Escoffier (1958).

2.2. Structuration interne

2.2.1. Multiplicité des parlers

Si le Croissant contraste assez nettement avec les aires linguistiques qui l'entourent, il n'en reste pas moins que l'aire croissantine présente une très importante diversité interne. En effet, il n'a jamais existé d'unité politique correspondant peu ou prou au Croissant ni de tradition écrite autochtone, le Croissant ayant été administré successivement en latin puis en français (à partir du Moyen-Âge (Bec 1986 : 77)), sans que les variétés parlées localement aient jamais accédé au statut de langues officielles. De ce fait, les parlers croissantins n'étant pas soumis à la pression d'une norme autochtone, ils ont donc évolué librement pendant des siècles, chaque micro-terroir développant ses propres caractéristiques grammaticales et lexicales.

Deux cas d'étude permettront de prendre la mesure de l'extrême variation que l'on peut rencontrer pour une forme donnée à travers le Croissant.

(i) Trait morphologique : le Tableau 1 ci-dessous propose un choix de variantes observées pour un même tiroir verbal.

Commune	Phonétique	Commune	Phonétique
Cellefrouin (Charente)	[i ʃãˈtɑv]	Crozant (Creuse)	[i ʃãˈtọv]
Pleuville (Charente)	[i ʃãˈtiː]	Gartempe (Creuse)	[i tsãˈtavə]
Azérables (Creuse)	[i ʃãˈtø]	Genouillac (Creuse)	[i ʃãˈtɛv]
Vareilles (Creuse)	[i ʃãˈtœv]	Saint-Pierre-le-Bost (Creuse)	[e ʃãtaˈvo]
Éguzon (Indre)	[i ʃãˈtez] ~ [i ˈʃãtj]	Archignat (Allier)	[e ʃãteˈvo]
		Naves (Allier)	[i ʃɛ̃ˈtʃɲɔ]
Saint-Sébastien (Creuse)	[i ˈʃãtj]	Châtel-Montagne (Allier)	[i ʃãˈtɛ̃]

Tableau 1. Formes attestées de la 1re personne du singulier de l'imparfait de l'indicatif du verbe 'chanter' (français standard : 'je chantais') dans divers parlers du Croissant

(ii) Trait lexical : les équivalents croissantins de l'adverbe français 'aussi' se répartissent en quatre types nettement différents :

- à l'extrême-ouest et au nord (Charente, Indre), on a des formes du type [aˈtu] ~ [øˈtu] ~ [iˈtu] ~ [eˈtu], apparentées au français régional *itou* et aux formes équivalentes du poitevin-saintongeais ;

- à l'ouest (Haute-Vienne et ouest de la Creuse), on a des formes [oˈsi] ~ [ɔˈsi], proches du français standard *aussi* ;

- dans l'est de la Creuse et l'ouest de l'Allier, on a des formes du type [eˈɾje] ~ [aˈrje] ~ [aˈrjɛ], qui semblent également employées dans certains parlers berrichons ;

- à l'extrême-est, on a des formes [mweˈtu] ~ [aˈmwɛ], contenant l'élément [mwe] ~ [mwɛ], apparenté aux termes occitans (OLL) *mai* [ˈmaj] 'plus', et *e mai* [eˈmaj] 'aussi, en outre'.

2.2.2. Marchois et bourbonnais

En dépit de ce foisonnement dialectal, on peut classer les parlers locaux en deux sous-groupes principaux :

- le marchois, qui regroupe les parlers de l'ouest ;

- le bourbonnais d'oc, qui regroupe les parlers de l'est.

La limite entre marchois et bourbonnais passe un peu à l'est de Bonnat (cf. Carte 2) et peut être caractérisée par un certain nombre de marqueurs linguistiques (Tableau 2)[5].

[5] Certains parlers, en particulier à la limite des deux zones, présentent simultanément des traits marchois et bourbonnais d'oc mais, dans l'ensemble, les traits présentés dans le Tableau 2 ont une répartition spatiale cohérente et permettent de départager assez précisément les deux grands groupes de variétés qui se partagent l'aire du Croissant.

Critère	Marchois	Bourbonnais d'oc	Ex
Pronom S1SG	toujours /i/	/e/ à l'ouest (en limite du marchois), /i/ à l'est	(5)
Pronom S3SG devant voyelle	/oz, uz, øz, v, w/	/ol/	(6)
Désinence verbale S1SG[6]	pas de marque spécifique	marque spécifique fréquente /u, o, ɛ̃/…	(7a/b)
Restricteur *ne…que*	pas de marque spécifique	/mɑ/ ~ /mɑk(ə)/	(8)
Radical parfait verbes irréguliers	marque /g/ ~ /dʒ/	marque /j/	(9)
Pronoms 3e personne masculins[7]	antécédents uniquement masculins	antécédents féminins inanimés	(18)
Pronoms toniques 3e personne PL	pas de marque spécifique	marque -/zot/ ~ -/zɔt/ fréquente	(10)
Verbe cognat de 'entendre'	maintenu	remplacé par cognat de 'écouter'	(11)

Tableau 2. Traits différenciant les parlers croissantins occidentaux (marchois) et orientaux (bourbonnais d'oc)

Exemples illustratifs :

(Ex5) 'je veux'

= marchois [**i** ˈvo] (Genouillac, Creuse)

vs. bourbonnais d'oc [**e** ˈvul] (Saint-Pierre-le-Bost, Creuse).

(Ex6) 'il a'

= marchois [**oz** ˈa] (Azérables, Creuse), [**ǫz** ˈa] (Éguzon-Chantôme, Indre), [**uz** ˈa] (Crozant, Creuse), [**øz** ˈɔ] (Oradour-Saint-Genest, Haute-Vienne), [**v** ˈǫ] (Alloue / Pleuville, Charente), [**w** ˈa]

[6] Hors futur et verbes très irréguliers.

[7] Ce critère, qui fait intervenir des phénomènes d'accord sémantique fondés sur la notion d'ANIMÉITÉ, est expliqué en détail plus loin dans ce chapitre.

(La Celle-Dunoise, Creuse), [**w** ˈọ] (Sainte-Colombe, Charente)

vs. bourbonnais d'oc [**ol** ˈa] (Archignat / Châtel-Montagne / Naves, Allier ; Nouzerines, Creuse).

(Ex7a) 'je chante' – 'il chante'

= marchois [i ˈʃãt] – [o ˈʃãt] (Azérables, Creuse)

vs. bourbonnais d'oc [i ʃɛ̃ˈt**u**] – [o ˈʃɛ̃t] (Naves, Allier).

(Ex7b) 'je trouvais' – 'il trouvait'

= marchois [i tɾuˈvɑv] – [o tɾuˈvɑv] (Saint-Front, Charente), [i tɾuˈvez] – [ọ tɾuˈvez] (Éguzon-Chantôme, Indre), [i tɾuˈviː] – [øː tɾuˈviː] (Pleuville, Charente)

vs. bourbounnais d'oc [e tʁuvaˈv**o**] – [o tʁuˈvɛv] (Saint-Pierre-le-Bost), [i tɾuˈv**ɛ̃**] ~ [o tɾuˈvo] (Châtel-Montagne).

(Ex8) '[le petit prince] (…) **ne** rencontra **qu**'une fleur' (Saint-Exupéry : 2007 [1946] : 78)

= marchois (…) ***n'trouvé qu'****une fioure* (Alloue, Charente (Barbier 2021 : 62)), (…) *o rencontri* ***qu'****une fleur* (La Châtre-Langlin, Indre (Larose 2021 : 62)), *trouvè* ***qu'****une flëur* (Noth, Creuse (Pradeau 2021 : 62))

vs. bourbounnais d'oc (…) ***n'****trouvi* ***mâ*** *na fleur* (Châtel-Montagne, Allier (Moutet 2021 : 63)), (…) *rincontrait* ***mâ*** *in-na flaou* (Naves, Allier (Grobost 2020 : 62)), (…) *ô l'a trouva* ***mâqu'****qu'une fleùr* (Nouzerines, Creuse (Contarin-Penneroux 2022 : 62)), (…) *ô trëuvé* ***mâqu'****ane fleur* (Toulx-Sainte-Croix, Creuse (Guy & Razet-Salessette 2022 : 62)).

(Ex9) 'il eut'

= marchois [w oˈ**g**e] (Dompierre-les-Églises, Haute-Vienne), [w ọˈ**dʒ**ẹ] (Noth, Creuse), [w aˈ**g**i] (Sainte-Colombe, Charente), [uz ọˈ**dʒ**i] (Saint-Plantaire, Indre)

vs. bourbounnais d'oc [ol aˈ**j**e ~ ol aˈ**ɟ**e] (Archignat, Allier), [ol aˈ**j**i] (Châtel-Montagne, Allier), [ol aˈ**j**e] (Saint-Pierre-le-Bost, Creuse).

(Ex10) '[pour] eux – elles'

= marchois [ji – jɛˈle] (Azérables, Creuse), [ji – jeˈle] (Lourdoueix-Saint-Michel, Indre), [i(ː) – ɛˈle(ː)] (Luchapt, Vienne), [i – ɛˈlɑ] (Oradour-Saint-Genest, Haute-Vienne), [iː – zɛl] (Sainte-Colombe, Charente)

vs. bourbonnais d'oc [jiˈ**zot** – jẹlˈ**zot**] (Nouzerines, Creuse), [jyˈ**zot** – jɛlˈ**zot**], (Saint-Pierre-le-Bost, Creuse), [jøˈ**zot**] (M./F.)

(Beaune d'Allier, Allier), [øˈ**zɔt** – ɛlˈ**zɔt**] (Châtel-Montagne, Allier).

(Ex11) 'écouter' vs. 'entendre'

= marchois (distingués) [ãtãd] vs. [ekuˈta] (Saint-Plantaire, Indre), [ãtãd] vs. [ekuˈtɑ] (La Châtre-Langlin, Indre), [ãtãd] vs. [akuˈtɑ] (Genouillac, Creuse), [ɛ̃ˈtɛ̃d] vs. [ekuˈta] (Sainte-Colombe, Charente), [ãtãdɾ] vs. [ekuˈta] (Dompierre-les-Églises, Haute-Vienne), [ãtãdʁ] vs. [ekuˈta] (Saint-Front, Charente)

vs. bourbonnais d'oc (confondus) [akuˈtɑ] (Nouzerines, Creuse), [ekuˈta] (Saint-Pierre-le-Bost, Creuse), [ekuˈtɑ] (Châtel-Montagne, Allier ; Toulx-Sainte-Croix, Creuse).

Étant donné que les parlers croissantins partagent de nombreux points communs avec les variétés occitanes voisines, on peut également considérer que le marchois est globalement davantage influencé par le limousin tandis que c'est l'influence auvergnate qui prédomine dans le cas du bourbonnais d'oc.

2.3. Des caractéristiques linguistiques originales

Au sein des langues romanes, les parlers du Croissant présentent souvent des caractéristiques originales et encore peu connues de la communauté scientifique, du fait du nombre réduit d'études qui leur ont été consacrées. Je me contenterai ici de citer quelques-unes de ces caractéristiques, mises au jour suite à des travaux effectués avec des locuteurs natifs, dans la plupart des cas depuis 2018.

2.3.1. Caractéristiques phonétiques et phonologiques

(i) D'assez nombreuses variétés du Croissant central distinguent trois degrés d'aperture (contre deux en français) pour les voyelles moyennes antérieures (Ex12, Ex13) et postérieures (Ex13).

(Ex12) Crozant (Creuse) :

vè /ˈvɛ/ 'ver de terre' vs. *vé* /ˈvẹ/ 'fois' vs. *vé* /ˈve/ 'regarde !' (lit. 'vois !') (Deparis : à paraître).

(Ex13) La Celle-Dunoise (Creuse) :

/ˈbw**ɛ**/ 'bois' vs. /ˈv**ẹ**/ 'fois' vs. /saˈv**e**/ 'savoir'.

/ˈk**ɛ**s/ 'caisse' vs. /trisˈt**ẹ**s/ 'tristesse' vs. /(kọ) ˈpr**e**s/ '(ça) presse'.

/ˈf**ɔ**ɾt/ 'forte' vs. /ˈp**ọ**ʃ/ 'poche' vs. /jãˈp**o**rt/ 'j'emporte'.

Ces trois degrés médians d'aperture (3DMA) ont, selon les parlers, une valeur phonologique (comme ci-dessus) ou simplement

phonétique. Les 3DMA ont pu être relevés dans les communes suivantes : La Châtre-Langlin, Éguzon-Chantôme, Lourdoueix-Saint-Michel et Saint-Plantaire (Indre), Crozant, La Celle-Dunoise, Mortroux, Noth, Nouzerines, Saint-Pierre-le-Bost et Saint-Sébastien (Creuse), Saint-Léger-Magnazeix (Haute-Vienne). Le phénomène des 3DMA s'inscrit donc visiblement dans une zone continue et peut être considéré comme aréal.

(ii) Dans l'ouest du Croissant (Charente et zones limitrophes), plusieurs parlers possèdent des oppositions de longueur (également attestées dans des parlers limousins voisins) avec un rendement élevé, tant au niveau de la flexion nominale (expression du nombre et du genre : cf. Tableaux 3 et 4) que verbale (distinctions temporelles (Ex14)).

Singulier	Français	Pluriel	Français
/dy/	'du'	/d**yː**/	'des'
/ˈʃmi/	'chemin'	/ˈʃm**iː**/	'chemins'
/ˈtu/	'tout'	/ˈt**uː**/	'tous'
/ˈpwe/	'puits'	/ˈpw**eː**/	'puits'
/ˈptjo/	'petit'	/ˈptj**oː**/	'petits'
/ˈøj/	'œil'	/ˈ**øː**j/	'yeux'
/ˈfjuʃ/	'fleur'	/ˈfj**uː**ʃ/	'fleurs'

Tableau 3. Oppositions de nombre fondées sur la quantité vocalique en parler de Sainte-Colombe (Charente)

Masculin	Français	Féminin	Français
/ʒɔˈli/	‘joli’	/ʒɔˈl**iː**/	‘jolie’
/(tu) ˈsu/	‘(tout) seul’	/(tut) ˈs**uː**/	‘(toute) seule’
/du/	‘deux (M.)’	/d**uː**/	‘deux (F.)’

Tableau 4. Oppositions de genre fondées sur la quantité vocalique en parler de Sainte-Colombe (Charente)

(Ex14) Pleuville (Charente) :
/øː tɾuˈvi/ ‘il trouva’ (passé simple)
vs. /øː tɾuˈv**iː**/ ‘il trouvait’ (imparfait).

(iii) Au niveau consonantique, les variétés croissantines de l’extrême-sud (p.ex. Fursac et Gartempe, à la limite des parlers occitans limousins) ont des systèmes caractérisés par la coexistence simultanée d’occlusives palatales et d’affriquées dentales et palatales, dans des contextes phoniques comparables (Ex15).

(Ex15) Gartempe (Creuse) :
/ˈ**c**a(ɾ)/ ‘clair’ vs. /ˈ**ts**a/ ‘chat’ vs. /ˈ**tʃ**a/ ‘chié (participe passé)’.
/ˈ**c**ɛ(də)/ ‘barrière’ vs. /(i) ˈ**ts**ɛ(rtsə)/ ‘je cherche’ vs. /ˈ**tʃ**ɛ(brə)/ ‘chèvre’.

2.3.2. Caractéristiques morphosyntaxiques

(i) Dans les parlers du Croissant charentais (p.ex. Sainte-Colombe et Saint-Front (Charente)), la marque de pluriel d’un syntagme génitival peut être portée par le nom déterminant (ou complément de nom) situé en fin de syntagme, tandis que le nom déterminé reste invariable (Ex16).

(Ex16) Sainte-Colombe (Charente) :
/lə ʃã d bja/ ‘le champ de blé’.
> /l**uː** ʃã d bj**ɑ(ː)**/ ‘les champs de blé’, lit. ‘les champ de blé**s**’, où le nom déterminant /bja/ ‘blé’ est fléchi au pluriel (/bjɑ(ː)/ ‘blés’) tandis que le nom déterminé /ʃã/ ‘champ’ reste invariable.

(ii) Dans plusieurs parlers croissantins du centre-nord (Allier, Creuse, Indre), l'accent remonte sur le radical verbal à la 3e personne du futur du singulier, ce qui fait qu'on observe deux radicaux futurs différents (un pour la 3e personne du futur singulier et un pour le reste du paradigme).

(Ex17) : 'je serai' – 'il sera'

= Archignat (Allier) [e ˈ**sʁ**e] – [u ˈ**sœʁ**], Azérables (Creuse) [i ˈ**sɾ**e] – [o ˈ**siɾ**], La Châtre-Langlin (Indre) [i ˈ**sɾ**ẹ] – [o ˈ**seɾ**], Crozant (Creuse) / Lourdoueix-Saint-Michel (Indre) [i ˈ**sɾ**ẹ] – [u ˈ**søɾ**], Éguzon-Chantôme (Indre) / Saint-Sébastien (Creuse) [i ˈ**sɾ**ɛ] – [ọ ˈ**sœɾ**], Genouillac / Mortroux (Creuse) [i ˈ**sɾ**e] – [u ˈ**siɾ**], Nouzerines (Creuse) [e ˈ**sɾ**e] – [o ˈ**səɾ**], Saint-Pierre-le-Bost (Creuse) [e ˈ**sʁ**ɛ] – [o ˈ**sœʁ**], Saint-Plantaire (Indre) [i ˈ**sɾ**ẹ] ~ [u ˈ**səɾ**].

(iii) Dans les parlers de l'est du Croissant (bourbonnais d'oc, p.ex. Archignat (Allier) et Saint-Pierre-le-Bost (Creuse)), on observe, pour divers paradigmes pronominaux de 3e personne, des phénomènes d'accord traduisant une distinction entre animé et inanimé (rarement décrite pour des variétés gallo-romanes) (Ex18).

(Ex18) Saint-Pierre-le-Bost (Creuse) :

*Mé pâ bin d'**yèlzautes*** [jɛlˈzot] (F.) *s'en rap'lant* (Dubac 2021 : 7) 'mais peu d'entre **elles** s'en souviennent (lit. 'rappellent')'. Le pronom [jɛlˈzot] 'elles' fait référence aux *grandes peursounes* 'grandes personnes' |ANIMÉ FÉMININ| et s'accorde en genre avec son antécédent.

vs. (…) *jusquant'que qua yi peurne fantésie à yune de **yuzautes*** [jyˈzot] (M.) *de s'rév'ya* (Dubac 2021 : 23) 'jusqu'à ce qu'il prenne fantaisie à l'une d'**elles** (lit. 'd'eux') de se réveiller'. Le pronom [jyˈzot] 'eux' fait ici référence aux *gueurnes* 'graines' |INANIMÉ FÉMININ| et ne s'accorde pas en genre avec son antécédent : c'est la forme masculine, que l'on peut ici considérer comme le genre non-marqué, qui est sélectionnée.

(iv) Dans certains parlers croissantins de l'extrême-nord, l'usage du pronom personnel sujet (fonctionnant de fait comme un indice) est obligatoire dans toutes les configurations syntaxiques, y compris dans les relatives (Ex19).

(Ex19) La Châtre-Langlin (Indre) :

(...) quèle grande peursoune ***ale*** *rèste en France* (Larose 2021 : 7) ‘cette grande personne habite en France’, lit. ‘elle reste en France’ : utilisation de l’indice sujet de 3e personne (ici *ale* [al] ‘elle’ F.SG) même quand le sujet (ici le syntagme nominal *grande peursoune* ‘grande personne’) est ouvertement exprimé.

*Ét y veugui un p’tit bounhoume (...) qu’****o*** *m’guéteu* (Larose 2021 : 12) ‘et je vis un petit bonhomme (...) qui me considérait’, lit. ‘qu’**il** me regardait’ : l’indice sujet de 3e personne (ici *o* [o] ‘il’ M.SG) est utilisé en sus du pronom relatif *qu’* [k].

*(...) ales vous d’mandant ryin su s’qu’****ou*** *compte* (Larose 2021 : 19) ‘elles ne vous questionnent jamais sur l’essentiel’, lit. ‘elles vous demandent rien sur ce que **ça** compte’ : l’indice sujet de 3e personne (ici *ou* [u] ‘ça’ NTR.SG) est utilisé en sus du pronom relatif *qu’* [k].

2.4. Un patrimoine menacé

De nos jours, les parlers du Croissant constituent clairement un patrimoine menacé (Guérin 2022). Dans la plupart des localités où un parler du Croissant est traditionnellement parlé, la transmission a généralement cessé avant 1945, parfois plus tôt, en particulier dans les villes (Commentry, Guéret, Montluçon, La Souterraine, Vichy). Souvent, dans une même commune, le centre-village (appelé ‘bourg’ en français régional) est passé au français avant les hameaux (ou ‘villages’) isolés, où certains enfants ont encore appris la langue locale parfois jusqu’aux années 1950 ou 1960, voire plus tard : ainsi, dans quelques communes rurales creusoises (p.ex. Crozant), on trouve des locuteurs natifs nés dans les années 1980. En tout état de cause, les parlers du Croissant sont désormais partout minoritaires sur leur territoire d’origine et leurs usagers ne représentent aujourd’hui jamais plus de 10% de la population locale. Si l’on extrapole ce chiffre (optimiste) à l’ensemble du Croissant (peuplé d’environ 350 000 personnes), on aurait donc un nombre maximal de 35 000 locuteurs. La réalité est probablement en-dessous de cette estimation et on peut considérer que le nombre total de ‘croissantophones’ ne dépasse pas 10 000 personnes en 2022. De plus, la grande majorité de ces locuteurs a plus de 75 ans et, les diverses variétés croissantines n’étant plus transmises aux enfants, il est vraisemblable que la majorité d’entre elles seront éteintes d’ici au

plus deux à trois décennies. Il y a donc une urgence toute particulière à étudier ce patrimoine linguistique original, encore peu décrit et sur le point de s'effacer.

3. Les recherches sur les parlers du Croissant : hier et aujourd'hui

Depuis maintenant plus de deux siècles, un certain nombre de chercheurs, originaires ou non de la zone du Croissant, se sont penchés sur les variétés romanes traditionnellement pratiquées sur ce territoire. Nous distinguerons deux phases successives dans cette aventure scientifique :

- une première phase allant des origines à 2013 ;
- une seconde phase (de 2013 à nos jours), qui correspond au renouveau des recherches consacrées aux parlers du Croissant.

3.1. Bref panorama historique des recherches sur les parlers du Croissant (des origines à 2013)

Des origines à 2013, les recherches sur les parlers du Croissant peuvent être à leur tour divisées en deux grandes périodes[8] :

- la genèse conceptuelle du Croissant (de la Révolution Française à 1913).
- l'accumulation des connaissances : de 1913 à 2013.

3.1.1. La genèse conceptuelle du Croissant

La Révolution Française amène les nouvelles élites du pays à considérer d'une façon également nouvelle la question de la langue. Or, à la fin du 18e siècle, les faits sont là : le français standard est la langue maternelle d'une minorité de Français et une très importante partie de la population ne le parle ni ne le comprend. En-dehors du bassin parisien, ce sont d'autres langues (basque, breton, occitan…) qui dominent le quotidien des citoyens. Les décideurs se préoccupent donc de recenser et de décrire les divers idiomes dans lesquels s'exprime une proportion considérable du peuple français et dont les révolutionnaires ambitionnent de restreindre l'usage au profit de la

[8] Pour une présentation plus détaillée des recherches sur le Croissant depuis les origines jusqu'à la fin du 20e siècle, cf. Brun-Trigaud (1990).

langue nationale, considérée par beaucoup comme l'idiome du progrès.

C'est dans cette atmosphère que, au début du 19e siècle, des relevés systématiques des variétés locales pratiquées en France sont entreprises par Coquebert de Montbret, qui se fonde pour ses recherches sur un corpus de traductions en vernaculaire de la parabole biblique de l'enfant prodigue. Environ 30 de ces traductions sont ainsi produites pour la zone du Croissant et, dès 1808, Coquebert de Montbret, en analysant les matériaux recueillis, commence à poser la question de la de la limite entre les langues d'oc (ou occitan) et d'oïl.

Dans la seconde moitié du 19e siècle, l'abbé Jean-Pierre Rousselot (1846-1924), l'un des plus grands linguistes français, lui-même locuteur natif du parler de Cellefrouin (Charente), s'intéresse en profondeur à son parler familial – auquel il consacre sa thèse en 1891 – ainsi qu'aux nombreuses variations qu'on rencontre dans le Croissant charentais et à la question de la limite entre langues d'oc et d'oïl.

En 1876, les deux linguistes méridionaux Charles de Tourtoulon et Octavien Bringuier publient les résultats d'enquêtes de terrain qu'ils ont menées eux-mêmes dans le Croissant et mettent en évidence l'existence d'une aire linguistique contenant des parlers intermédiaires entre langues d'oc et d'oïl. Octavien Bringuier étant mort avant la fin des enquêtes, seule la partie occidentale du Croissant (qui ne porte pas encore son nom) sera cartographiée, de la Charente jusqu'au centre de la Creuse (au niveau de Guéret). La zonation proposée par les deux linguistes pour les différentes variétés étudiées est extrêmement rigoureuse et la plupart de leurs résultats restent toujours valables en ce début de 21e siècle.

En 1913, le linguiste Jules Ronjat forge finalement le terme *Croissant* :

> « notre limite [oc/oïl] (…) a la forme d'un arc de cercle dont la corde serait sensiblement dirigée de l'O[uest]. à l'E[st]., et elle sépare notre domaine d'un territoire présentant la forme générale d'un **croissant** (largeur maximum 40 à 50 km. ; longueur de la corde d'arc, environ 240), dans lequel on rencontre des parlers intermédiaires » (Tome 1, p. 14)[9].

[9] Le mot graissé et les passages entre crochets sont mes propres ajouts.

Comme le montre l'extrait ci-dessus, c'est à la forme de demi-lune qu'il dessine sur les cartes géographiques que le Croissant doit son nom. C'est désormais de la sorte que les linguistes désigneront cette aire particulière.

3.1.2. L'accumulation des connaissances (1913-2013)

Le Croissant, ainsi caractérisé, sera étudié au cours de cette période par différents chercheurs et à différents niveaux. On peut distinguer trois grands types d'abordages de cette aire linguistique :

(i) les travaux de synthèse, eux-mêmes subdivisables en deux catégories :

- les thèses doctorales, comme celle de Simonne Escoffier (1958), qui s'intéresse aux limites entre occitan, langues d'oïl et francoprovençal, à l'extrémité orientale du domaine (est de l'Allier) ou celle de Marie-France Lagueunière (1983), consacrée aux parlers de l'arrondissement de Bellac (Haute-Vienne) ou encore celle de Guylaine Brun-Trigaud (1990), qui traite du Croissant dans une perspective historique et épistémiologique, ou bien celle de Stephan Mietzke (2000), qui propose une approche microdialectologique de la variation dans le Croissant.

- d'autres ouvrages de portée générale, p.ex. Baldit (1980) sur les parlers de la Creuse, Vignaud & Manville (2007) sur la région de Guéret, Baldit & al. (2010) sur le marchois ou Reichel (2012) sur les parlers arverno-bourbonnais.

(ii) les atlas linguistiques, en particulier ceux produits par le CNRS entre la seconde guerre mondiale et les années 1980 et qui incluent la zone du Croissant, écartelée entre quatre projets géographiques distincts : l'ouest de la France (Massignon & Horiot 1971-1983), l'Auvergne et le Limousin (Potte 1975-1987), le centre de la France (Dubuisson 1971-1982) et le Lyonnais (Gardette 1950-1976).

(iii) les monographies descriptives, produites par des linguistes et des érudits locaux et témoignant des caractéristiques de différents parlers pratiqués sur des espaces géographiques restreints, p.ex. Bouchard (2009) sur Saint-Priest-en-Murat (Allier)[10], Brun-Trigaud

[10] Le parler de Saint-Priest-en-Murat n'est pas à proprement parler une variété du Croissant. Sur le plan morphologique, il s'agit en effet d'un parler du bourbonnais

(1993) sur Lourdoueix-Saint-Michel (Indre), Dupeux (2014 [2013]) sur la Basse-Marche (Haute-Vienne), Pasty (1999) sur Fleurat et ses environs (Creuse), Quint (1991, 1996), sur Saint-Priest-la-Feuille et Gartempe (Creuse), Yvernault (2013) sur Archignat (Allier).

Ainsi, plusieurs dizaines de publications au moins ont porté sur les parlers croissantins dans le siècle qui a suivi la définition du Croissant par Ronjat. Malgré cette production relativement conséquente, il n'existait en 2013 aucune donnée disponible pour de nombreuses variétés locales.

3.2. Le renouveau des recherches sur le Croissant linguistique

3.2.1. Les projets collectifs

En 2013, à la demande d'une association culturelle locale, un premier colloque sur les parlers du Croissant, rassemblant des scientifiques (linguistes mais aussi historiens et anthropologues) ayant travaillé sur cette région se tient à Crozant, en présence d'un public constitué d'une majorité de locuteurs qui veulent en savoir davantage sur le patrimoine dont ils sont détenteurs. Cette rencontre entre chercheurs et locuteurs va insuffler une nouvelle dynamique aux recherches sur les parlers du Croissant et va déboucher sur la mise en place de plusieurs projets collectifs financés par des fonds publics.

Tout d'abord, en 2015, la DGLFLF (Délégation Générale à la Langue Française et aux Langues de France du Ministère de la Culture) soutient, dans le cadre de l'Observatoire des Pratiques Linguistiques, une première série d'enquêtes sur les parlers du Croissant.

Dès l'année suivante, le Labex EFL (*Empirical Foundations of Language* / Fondements Empiriques du Langage) permet de prolonger l'effort entrepris en soutenant successivement deux opérations pluriannuelles :

- LC4 « Les parlers du Croissant » (2016-2019), avec LC = *Language Contact* / contact de langues.

---

d'oïl. Il est cependant pratiqué à moins de cinq kilomètres de la frontière linguistique et, en particulier d'un point de vue lexical, il partage énormément d'éléments avec les parlers croissantins *stricto sensu*, d'où mon choix de l'inclure dans cette liste.

- VC2 « Au coeur de la Gallo-Romania : caractérisation linguistique et environnementale d'une aire de transition » (2020-2024), avec VC = *Variation and Contact* / contact et variation.

Ces appuis institutionnels permettent de mettre en place :

- une véritable équipe de plus d'une dizaine de scientifiques intéressés par les parlers du Croissant et disposés à y investir une partie de leur temps de travail ;

- un réseau de locuteurs natifs désireux de préserver la mémoire de leurs parlers respectifs et de transmettre leurs connaissances aux linguistes et aux générations futures.

Ce groupe de travail, mis sur pied grâce à la DGLFLF et au Labex EFL, parvient alors à obtenir successivement deux projets plus ambitieux :

- le projet « Le Croissant linguistique : une approche multidisciplinaire du contact oc-oïl[11] » (2018-2022), soutenu par l'ANR (Agence Nationale de la Recherche), permet aux équipes constituées précédemment d'agir avec davantage de moyens, et en particulier de recruter du personnel dédié : un chercheur postdoctoral, Maximilien Guérin, et une doctorante, Amélie Deparis, qui prépare une thèse intitulée « Le Croissant linguistique, contact entre langue d'oc et langue d'oïl dans l'aire gallo-romane : étude des traits linguistiques significatifs et de leurs représentations en cartographie. » C'est la première fois dans l'histoire des études sur le Croissant que des chercheurs sont salariés à temps plein pour travailler dans ce domaine.

- le projet « Oc/Oïl : textes, identité et contact de langues aux confins gallo-romans » (2021-2024), soutenu par le dispositif « Émergence(s) » de la Ville de Paris, prolonge l'effort initié dans les projets précédents et permet lui aussi de rétribuer un chercheur postdoctoral pendant quatre ans.

3.2.2. Les colloques

En parallèle avec l'éclosion de ces projets collectifs, des colloques dédiés (les « rencontres sur les parlers du Croissant ») ont été tenus à trois reprises dans le Croissant : au Dorat (Haute-Vienne) en 2017,

[11] https://anr.fr/Projet-ANR-17-CE27-0001

à Montluçon[12] (Allier) en 2019 et, après une interruption due aux perturbations engendrées par le covid, à Boussac[13] (Creuse) en 2022. Des quatrièmes rencontres sont en préparation pour 2024 à Saint-Amant-de-Boixe. Ces colloques, tous ouverts au public, ont rassemblé à chaque fois plusieurs dizaines de personnes (locuteurs et non-locuteurs) et ont contribué à développer les interactions entre les scientifiques (linguistes) et la population vivant dans l'aire croissantine.

3.2.3. Récapitulation sur le renouveau des études croissantines

Avant 2013, comme je l'ai dit plus haut (cf. **3.1.**), il y avait déjà eu d'assez nombreux travaux consacrés aux parlers du Croissant. Le grand changement survenu au cours de la dernière décennie porte sur plusieurs points :

- *la recherche sur les parlers du Croissant* est devenue *collective*. C'est désormais une équipe de chercheurs qui travaille sur l'objet et met en commun les fruits de ces recherches, alors qu'auparavant chaque initiative partait d'un seul individu ;

- *le territoire du Croissant est systématiquement quadrillé* par l'équipe constituée afin de rendre compte de la façon la plus complète possible de l'énorme diversité interne du Croissant (cf. **2.2.1.**) ;

- *les locuteurs sont plus étroitement impliqués dans les recherches*. Ils fournissent des données aux linguistes, assistent aux colloques, s'efforcent de trouver d'autres locuteurs. Cette implication est liée à une prise de conscience : de nombreux croissantophones sont désormais conscients que les diverses variétés locales qu'ils ont pratiquées dans leur enfance risquent de disparaître avec eux et qu'il est à la fois urgent et important d'en préserver la mémoire ;

- la recherche contemporaine sur les parlers du Croissant se veut également *multidisciplinaire*. Un certain nombre de chercheurs s'efforcent de collecter des données sur le terrain auprès des dépositaires des parlers locaux tandis que d'autres proposent des approches transversales dans différents domaines : cartographie, morphologie, phonologie, psycholinguistique (études sur le bilinguisme des locuteurs), syntaxe, sémantique, sociolinguistique,

[12] https://croissant2019.sciencesconf.org/
[13] https://croissant2022.sciencesconf.org/

TAL (traitement automatique des langues). Cette multiplicité des approches a permis de mettre en valeur l'intérêt scientifique exceptionnel des parlers du Croissant qui, en tant que variétés de contact et du fait de leur haut degré de variation interne, constituent un objet d'étude de choix pour de nombreuses branches des Sciences du Langage.

4. Les réalisations récentes sur les parlers du Croissant

4.1. Publications scientifiques et de valorisation[14]

4.1.1. Publications collectives

Deux publications récentes illustrent de façon claire le travail produit en équipe par les chercheurs impliqués dans les différents projets mentionnés ci-dessus sur les parlers du Croissant :

- *Le croissant linguistique : entre oc, oïl et francoprovençal* (Esher, Guérin, Quint & Russo 2021). Ce volume rassemble 17 contributions scientifiques originales d'une vingtaine d'auteurs, toutes consacrées au Croissant ou à ses environs géographiques immédiats. C'est la première publication collective de cette ampleur portant essentiellement sur l'aire croissantine.

- le numéro 30 de la revue *Langues et Cité*[15], une publication produite par la DGLFLF et traitant de la diversité linguistique en France. Ce numéro, intitulé « Les parlers du Croissant » et qui regroupe plus de quinze contributions, fait le point sur le sujet dans un style à la fois rigoureux et accessible au grand public, et en particulier aux personnes qui vivent dans la région concernée ou qui en sont originaires.

4.1.2. Monographies, chapitres d'ouvrage et articles

Par ailleurs, plusieurs dizaines de publications scientifiques produites par un ou quelque(s) chercheur(s) parfois associé(s) à des locuteurs sont parues depuis le lancement des premiers projets. On peut regrouper ces productions en quatre grandes catégories :

[14] Pour une liste à jour de ces publications, cf. https://parlersducroissant.huma-num.fr/publications.html

[15] https://www.languesetcite.fr/148

(i) les monographies descriptives, consacrées à un parler donné dont elles présentent les principales caractéristiques grammaticales et lexicales. Quatre de ces monographies sont parues, consacrées respectivement aux parlers de Dompierre-les-Églises (Haute-Vienne, Guérin 2019), Oradour-Saint-Genest (Haute-Vienne, Guérin 2020b), La Celle-Dunoise (Creuse, Maurer-Cecchini 2021) et Châtel-Montagne (Allier, Maurer-Cecchini 2023).

(ii) un atlas linguistique de la Creuse (Brun-Trigaud 2020), qui explore la variation diatopique dans le département, y compris dans sa partie croissantine.

(iii) des éditions critiques de textes produits par des locuteurs natifs (trois sont parues à ce jour).

(iv) environ 20 chapitres d'ouvrage et articles, rédigés en anglais ou en français, portant sur des sujets transversaux (dialectologie, morphologie, traitement automatique des langues…) et mettant généralement en jeu plusieurs variétés croissantines.

4.1.3. Publications de valorisation

Les membres de l'équipe ont aussi été actifs dans ce domaine, à trois niveaux au moins.

(i) L'action la plus spectaculaire consiste en la traduction systématique du *Petit Prince* d'Antoine de Saint-Exupéry dans un grand nombre de variétés croissantines. En effet, du fait du caractère désormais quasi-exclusivement francophone de la société française contemporaine (en particulier dans des zones rurales comme le Croissant), les locuteurs croissantins restants n'ont généralement que très peu d'occasions de pratiquer leur langue maternelle et, s'il leur est assez aisé de répondre à des questionnaires lexicaux ou morphologiques, il leur est souvent nettement plus difficile de produire spontanément des textes oraux d'une certaine ampleur. L'exercice de la traduction d'un texte littéraire connu (*le Petit Prince*) s'est avéré un bon moyen d'illustrer les traits syntaxiques de chaque variété employée : en effet, comme les locuteurs avaient le temps de se concentrer sur le texte et de se remettre mentalement en situation dans la langue de leur enfance, ils étaient plus à même d'en restituer les particularités que si nous leur avions tendu un micro pour qu'ils nous racontent quelque chose en « patois » (dénomination la plus couramment usitée par les locuteurs croissantins pour désigner leurs

parlers). *Le Petit Prince*, qui est l'œuvre la plus traduite au monde après la Bible, présente des attraits certains pour une adaptation dans une langue à tradition orale (cas de figure de l'ensemble des variétés du Croissant) : c'est un texte relativement court (environ 100 pages), écrit dans une langue assez simple et contenant peu de mots étrangers à la civilisation rurale dans laquelle l'ensemble des locuteurs actuels ont acquis leurs parlers respectifs[16]. De plus, d'un point de vue linguistique, le texte est riche :

- l'ensemble des temps verbaux du français (y compris le prétérit et l'imparfait du subjonctif) y apparaissent et ces temps ont généralement des équivalent fonctionnels dans la plupart des parlers du Croissant, ce qui permet de les observer en contexte ;

- on y trouve un grand nombre de phrases complexes avec différents types de subordonnées ;

- toutes les personnes du discours y sont aussi attestées.

Une traduction du *Petit Prince* soigneusement préparée par un locuteur natif (assisté par un linguiste pour assurer la cohérence phonologique et morphologique du texte) fournit donc un tableau assez complet du fonctionnement d'une variété croissantine donnée.

En outre, le texte du *Petit Prince* présente un avantage qui dépasse tous ceux qui viennent d'être énumérés : de nombreux locuteurs du Croissant ont adhéré à son contenu et ont pris plaisir à le traduire dans leur variété maternelle. C'est ainsi que, au moment où j'écris ces lignes, 26 traductions distinctes du Petit Prince ont été produites par des locuteurs croissantins, dont 20 sont aujourd'hui publiées en version papier[17] et quatre autres ont déjà été révisées conjointement avec un linguiste. Cet effort considérable de traduction a donc permis de produire des textes suivis dans un grand nombre de parlers de l'ensemble du Croissant et par là même de constituer une précieuse base textuelle (entièrement numérisée), qui pourra dans le futur être exploitée par différents spécialistes (p.ex. romanistes, typologues, spécialistes de TAL).

---

[16] Il y a quelques exceptions telles que *astronome* et *télescope* (termes techniques récents) ou encore *baobab* et *boa* (désignant des espèces vivantes absentes de l'environnement croissantin), mais ces termes sont en nombre réduit dans l'œuvre.

[17] Six de ces traductions sont aussi disponibles en version audio : https://parlers-ducroissant.huma-num.fr/livres-audios/#book0

En sus de leur intérêt linguistique patent, les versions publiées de ces traductions sont également appréciées au niveau local, où le grand public ne dispose généralement que de peu d'informations sur les langues pratiquées traditionnellement dans la région : l'expérience a montré que les acteurs culturels locaux et les traducteurs eux-mêmes étaient souvent surpris par le succès rencontré dans les communes et microrégions concernées par les livres produits.

Les traductions du *Petit Prince* dans des parlers croissantins constituent donc une activité de valorisation au sens positif du terme : elles fournissent aux scientifiques une grande quantité d'informations sur la langue des traducteurs et permettent dans le même temps un retour des connaissances engrangées sur cette langue vers la population de l'aire linguistique considérée.

(ii) Un imagier, *Mes mille premiers mots en bas-marchois (Guérin & Dupeux 2020), a également été produit pour le bas-marchois, un ensemble dialectal croissantin relativement homogène comprenant la majeure partie du Croissant de Haute-Vienne ainsi que plusieurs communes creusoises (autour de La Souterraine). Cette réalisation présente la particularité d'être fondamentalement destinée à un public enfantin, susceptible de redécouvrir ainsi la langue autrefois pratiquée par la majorité des habitants de la Basse-Marche.*

(iii) Les membres des projets consacrés aux parlers du Croissant ont été également actifs dans les médias (plusieurs dizaines d'interventions dans la presse écrite, à la radio[18] et à la télévision), ce qui a contribué à mieux faire connaître le Croissant au public local mais aussi national.

4.2. Collecte de matériaux linguistiques

En sus des publications, il convient de souligner l'énorme activité de collecte de matériaux linguistiques effectuée par les membres du projet. Des questionnaires adaptés aux caractéristiques culturelles et

[18] Dans ce domaine on mentionnera la chronique de podcasts à destination du grand public produite par les chercheurs impliqués dans les projets sur les parlers du Croissant et diffusée par la radio locale Ici L'Onde - la webradio du Haut Limousin, implantée dans le Croissant (https://parlersducroissant.huma-num.fr/audiovisuel.html) et par Canal-U (https://www.canal-u.tv/chaines/cnrs-service-audiovisuel-d-ardis-uar2259/les-podcast-du-croissant).

typologiques des parlers du Croissant ont été développés (Brun-Trigaud, Guérin & Quint 2018) et soumis dans plus de 30 communes de toute la zone considérée (soit environ une commune sur dix)[19].

Grâce à ces collectes, il n'existe plus à proprement parler de « trou noir » dans l'aire linguistique croissantine : certes, ces relevés sont loin d'être exhaustifs et il est encore possible d'observer de nombreux phénomènes linguistiques méconnus dans le Croissant. Cependant, nous avons désormais une idée assez complète des principales caractéristiques typologiques de l'ensemble du Croissant, d'autant plus que les données des questionnaires sont complétées par un certain nombre de textes oraux et par les plus de 20 traductions du Petit Prince déjà complétées par les locuteurs (cf. **4.1.3.**).

Notons enfin que l'ensemble des questionnaires relevés ont été enregistrés au moyen d'enregistreurs numériques, ce qui fait que ces collectes répondent aux critères actuels de documentation linguistique et permettent la divulgation de données fiables, de première main et aisément exploitables par les chercheurs actuels ou du futur.

4.3. L'audiovisuel

L'un des engagements pris lors du dépôt du projet ANR était la production d'un film portant sur les parlers du Croissant. Cet engagement a été tenu et un documentaire a été produit, avec la collaboration d'ARDIS (**Appui à la Recherche et Diffusion des Savoirs – UAR 2259**)[20]. Ce film, d'une durée de 40 minutes, a été réalisé par Franck Guillemain (2020). Le tournage s'est déroulé principalement dans le Croissant : il a impliqué la participation de trois linguistes (Amélie Deparis, Maximilien Guérin et Nicolas Quint) et de plusieurs dizaines de locuteurs. Ces nombreuses rencontres ont aussi stimulé les interactions entre chercheurs et locuteurs et le format audiovisuel, accessible au plus grand nombre, a contribué à faire connaître les travaux de notre équipe dans le Croissant et auprès du grand public en général.

---

[19] https://www.google.com/maps/d/u/0/viewer?mid=15sF_lH--rndKshFW9Ws8Uk4bVQJfxfNC&ll=46.16158130055626%2C1.9859560000000176&z=9

[20] https://www.canal-u.tv/chaines/cnrs-service-audiovisuel-d-ardis-uar2259/les-parlers-du-croissant

4.4. Internet et l'accès aux données en ligne

Les recherches contemporaines consacrées aux parlers du Croissant ont aussi pu profiter du développement rapide des humanités numériques.

Ainsi, chacun des deux grands projets financés sur le thème du Croissant dispose de son propre site, hébergé par HumaNum :

- ANR Croissant : https://parlersducroissant.huma-num.fr/
- Émergence(s) Oc/Oïl : https://oc-oil.huma-num.fr/?fbclid=IwAR2aWXv6SqQkm0dohLotOvlSo812T2W7S2nkPc7oRNsXlEGYcUD-L0yoeYk

Les données collectées (soit plusieurs milliers de fichiers son et divers documents écrits numérisés) sont hébergées sur ces espaces de stockage et rendues ainsi accessibles à la communauté scientifique et au grand public : https://parlersducroissant.huma-num.fr/corpus/

Enfin, une page Facebook, animée par des membres de ces projets[21], favorise les échanges avec les internautes intéressés par les parlers du Croissant. Les locuteurs y sont en particulier sollicités chaque semaine pour fournir l'équivalent d'un terme français donné dans leurs variétés respectives. Cet exercice de traduction rencontre régulièrement un franc succès (avec parfois plus de 20 participants par item proposé) et a permis à notre équipe d'entrer en contact avec des locuteurs pratiquant des variétés croissantines que nous n'avions pas encore pu étudier.

5. Et demain ? Perspectives sur le patrimoine linguistique croissantin

Pour clore ce panorama, il est bien naturel de se poser la question de l'avenir des études croissantines et de celui des parlers du Croissant.

5.1. La disparition inéluctable des variétés locales

L'extinction des variétés croissantines, comme celle de nombreuses autres langues dans le monde, revêt un caractère inéluctable. En effet, il est quasi-certain que, d'ici quelques décennies, avec la disparition des derniers locuteurs, les parlers du

[21] https://www.facebook.com/ParlersCroissant/

Croissant auront cessé d'être des langues vivantes et il est d'ores et déjà trop tard pour enrayer ce mouvement : la transmission familiale de la langue aux enfants ayant généralement cessé il y a au moins 60 à 70 ans (cf. **2.4.** et Guérin 2022), on ne voit pas trop ce qui pourrait renverser la tendance.

De plus, à compter que les parlers croissantins soient demain pris en compte dans le cadre scolaire (ce qui n'a quasiment jamais été le cas), cela supposerait la production de matériel pédagogique adapté. Or, en raison de l'énorme diversité interne du Croissant, de l'absence d'une variété locale reconnue comme plus prestigieuse et du faible nombre de spécialistes de ces variétés, il semble peu probable qu'un standard pan-croissantin se développe dans un horizon proche.

Enfin il faut souligner que, indépendamment de la tendance déjà ancienne au remplacement des parlers croissantins par le français, l'exode rural massif qui a frappé la plus grande partie du Croissant au cours du 20e siècle (où certaines communes ont vu leur population divisée par trois) a aussi largement contribué à fragiliser lesdits parlers. Que cela plaise ou non, les parlers du Croissant, tels que les linguistes d'aujourd'hui peuvent encore les observer, appartiendront bientôt au domaine des souvenirs.

5.2. L'enjeu patrimonial

Malgré ces sombres pronostics quant à leur transmission en tant que langues parlées, les parlers du Croissant constituent un enjeu patrimonial aussi réel qu'actuel pour les régions concernées, et ce à plusieurs niveaux :

- tout d'abord, sur le plan scientifique, le patrimoine linguistique croissantin, malgré les efforts actuels, est loin d'avoir été complètement inventorié ni exploré. Le Croissant recèle encore assez d'attraits pour justifier de nombreuses recherches en linguistique et dans d'autres sciences humaines (anthropologie, histoire…).

- ensuite, les parlers croissantins représentent un des éléments les plus originaux du patrimoine culturel de l'aire où ils sont traditionnellement pratiqués. Que ces parlers soient ou non encore parlés à l'avenir, ils constituent un bien possédé en propre par les habitants du Croissant et susceptible de contribuer à un développement positif du sentiment identitaire local.

- enfin, à une époque où la société française s'interroge régulièrement sur la valeur et les vertus de ses langues régionales, les parlers du Croissant, pratiqués à la limite entre occitan, langues d'oïl et francoprovençal, aident à mieux comprendre ce qui fait la spécificité de chacune de ces trois zones linguistiques qui, prises ensemble, recouvrent plus de 80 % du territoire français métropolitain.

Le Croissant et ses multiples parlers méritent donc qu'on s'intéresse davantage à eux, peut-être plus encore demain qu'aujourd'hui.

6. Bibliographie

6.1. Références sur les parlers du Croissant

BALDIT Jean-Pierre (1980). *Les parlers creusois*. Guéret : UFOLEA (Fédération des Œuvres Laïques de la Creuse / Institut d'Études Occitanes Marche - Combraille).

BALDIT Jean-Pierre, Jeanine BERDUCAT, Guylaine BRUN-TRIGAUD, Gérard GUILLAUME & Christophe MATHO (2010). *Patois et chansons de nos grands-pères Marchois : Haute-Vienne, Creuse, Pays de Montluçon*. Romorantin : CPE (Communication-Presse-Édition).

BEC Pierre (1986). *La langue occitane*. Paris : Presses Universitaires de France (coll. *Que sais-je ?*).

BOUCHARD Edmond (2009). *Le patois, tel que je l'ai pratiqué de 1930 à 1946*, ISBN : 978-1445214825 [pas de lieu d'édition ni de nom d'éditeur].

BRUN-TRIGAUD Guylaine (1990). *Le croissant : le concept et le mot*. Lyon : Centre d'études linguistiques Jacques Goudet.

BRUN-TRIGAUD Guylaine (1992). Les enquêtes dialectologiques sur les parlers du Croissant : corpus et témoins. *Langue française* 93. 23-52.

BRUN-TRIGAUD Guylaine (1993). *Le parler de Lourdoueix-Saint-Michel (Indre)*. Guéret : Mémoires de la Société des Sciences Naturelles et Archéologiques de la Creuse.

BRUN-TRIGAUD Guylaine (2020). *Les parlers de la Creuse. Frontière et carrefour*, Études creusoises XXIV. Guéret : Société des Sciences Naturelles, Archéologiques et Historiques de la Creuse.

BRUN-TRIGAUD Guylaine, Maximilien GUÉRIN & Nicolas QUINT (2018). Questionnaire « Parlers du Croissant » — Conjugaison ; lexique fondamental ; lexique complémentaire. Projet ANR « Les Parlers du Croissant », http://parlersducroissant.huma-num.fr/participer.html (consulté le 07/04/2022) ou http://tulquest.huma-num.fr/en/search/node/croissant (consulté le 07/04/2022).

DEPARIS Amélie (à paraître). *Le Croissant linguistique, contact entre langue d'oc et langue d'oïl dans l'aire gallo-romane : étude des traits linguistiques significatifs et de leurs représentations en cartographie*. Paris : INALCO.

DUBUISSON Pierrette (1971-1982). *Atlas linguistique et ethnographique du Centre* (ALCe). Paris : CNRS, 3 vol.

DUPEUX Michel (2014 [2013]). *Le patois de la Basse-Marche*, 3e édition. Édité par l'auteur. ISBN 978-2-7466-6921-5.

ESCOFFIER Simone (1958). *La rencontre de la langue d'oïl, de la langue d'oc et du francoprovençal entre Loire et Allier : limites phonétiques et morphologiques* [thèse de doctorat]. Lyon : Université de Lyon.

ESHER Louise, Maximilien GUÉRIN, Nicolas QUINT & Michela RUSSO (dir.) (2021). *Le Croissant linguistique entre oc, oïl et franco-provençal, Des mots à la grammaire, des parlers aux aires*. Paris : L'Harmattan.

GARDETTE Pierre (1950-1976). *Atlas linguistique et ethnographique du Lyonnais* (ALLyl). Paris : CNRS, 5 vol.

GUÉRIN Maximilien & Michel DUPEUX (2020). *Mes mille premiers mots en bas-marchois*. Neckarsteinach / La Crèche : Tintenfaß / La Geste Éditions.

GUÉRIN Maximilien & Nicolas QUINT (dir.) (2021). Les parlers du Croissant. *Langues et Cité* 30, https://www.languesetcite.fr/148 (consulté le 14/08/2021).

GUÉRIN Maximilien (2019). *Grammaire du parler marchois de Dompierre-les-Églises (Haute-Vienne)*. Paris : L'Harmattan.

GUÉRIN Maximilien (2020a). Les parlers du Croissant : des parlers de transition au cœur de l'aire gallo-romane. *Feuille de Philologie Comparée Lituanienne et Française* XI. 13-33.

GUÉRIN Maximilien (2020b). *Le parler marchois d'Oradour-Saint-Genest (Haute-Vienne) : Abrégé grammatical et lexique thématique*. Paris : L'Harmattan.

GUÉRIN Maximilien (2022). Les parlers du Croissant : des parlers minorisés et marginalisés. In Stéphanie NOIRARD (dir.), *Transmettre les langues minorisées : entre promotion et relégation*. Rennes : Presses Universitaires de Rennes. 129-141.

LAGUEUNIÈRE France (1983). *Études de géographie linguistique dans l'arrondissement de Bellac (Haute-Vienne) : phonétique historique et phonologie* [thèse de doctorat]. Paris : Université Paris-Sorbonne.

MASSIGNON Geneviève & Brigitte HORIOT (1971-1983). *Atlas linguistique et ethnographique de l'Ouest* (ALO), Paris : CNRS, 3 vol.

MAURER-CECCHINI Philippe (2021). *Grammaire descriptive du parler croissantin de La Celle-Dunoise (Creuse) — Avec quelques données sur le parler de Saint-Sulpice-le-Dunois*. Paris : L'Harmattan.

MAURER-CECCHINI Philippe (2023). *Grammaire descriptive du parler croissantin de Châtel-Montagne (Montagne bourbonnaise, Allier)*. Paris : L'Harmattan.

MIETZKE Stephan. 2000. *Isoglossenverschiebungen im Croissant. Von der monodimensionalen Sprachgeographie zur pluridimensionalen Mikrodialektologie*. Kiel : Westensee Verlag.

PASTY Gilbert (1999). *Glossaire des dialectes marchois et haut limousin de la Creuse*. Édité par l'auteur. ISBN 2-9513615-0-5.

POTTE Jean-Claude (1975-1987). *Atlas linguistique et ethnographique du Limousin et de l'Auvergne* (ALAL). Paris : CNRS, 2 vol.

QUINT Nicolas (1991). *Le parler marchois de Saint-Priest-la-Feuille (Creuse)*. Limoges : La Clau Lemosina.

QUINT Nicolas (1996). *Grammaire du parler occitan nord-limousin marchois de Gartempe et de Saint-Sylvain-Montaigut (Creuse) : Étude phonétique, morphologique et lexicale*. Limoges : La Clau Lemosina.

QUINT Nicolas (2021). Le parler traditionnel de Saint-Amant-de-Boixe et l'extrémité orientale du Croissant (1ère partie). *Jadis [le canton de Saint Amant de Boixe et ses environs]* 20, 92-104.

QUINT Nicolas (2022). Le parler traditionnel de Saint-Amant-de-Boixe et l'extrémité orientale du Croissant (2e partie). *Jadis [le canton de Saint Amant de Boixe et ses environs]* 21, 88-96.

QUINT Nicolas, GUÉRIN Maximilien & BRUN-TRIGAUD Guylaine (à paraître). Les parlers croissantins d'Indre : un patrimoine linguistique original et méconnu. *Revue de l'Académie du Centre*.

REICHEL Karl-Heinz (2012). *Études et recherches sur les parlers arverno-bourbonnais*. Chamalières : Cercle Terre d'Auvergne.

RONJAT Jules (1913). *Essai de syntaxe des parlers provençaux modernes*. Mâcon : Protat, 4 vol.

ROUSSELOT Jean-Pierre (1891). *Les modifications phonétiques du langage étudiées dans le patois d'une famille de Cellefrouin (Charente)* [thèse de doctorat]. Paris : Welter.

TERRACHER Adolphe-Louis (1914). Les aires morphologiques dans les parlers populaires de l'Angoumois (1800-1900) [thèse de doctorat]. Paris : Champion.

TOURTOULON Charles de & Octavien BRINGUIER (1876). *Étude sur la limite géographique de la langue d'oc et de la langue d'oïl*. Paris : Imprimerie Nationale. Carte accessible en ligne : https://upload.wikimedia.org/wikipedia/commons/a/a7/Tourtoulon.jpg (consultée le 11/08/2021).

VIGNAUD Jean-François & Michel MANVILLE (2007). *Langue & mémoire du pays de Guéret*. Guéret : Conseil Général de la Creuse.

YVERNAULT Edith (2013). *Le Petit Yvernault illustré - Patois d'Archignat*. Édité par l'autrice, https://docs.google.com/viewerng/viewer?url=http://ekladata.com/GoECO4oG1E4THpWpPtltXpCMUMg/Le-patois-d-Archignat.pdf (consulté le 07/04/2022).

6.2. Traductions du Petit Prince

BARBIER Pierre (traducteur) & Nicolas QUINT (éd.) (2021). *Le P'ti Prince* [traduction en allousien (Alloue, Charente) du *Petit Prince* d'Antoine de Saint-Exupéry]. Neckarsteinach : Tintenfaß.

CONTARIN-PENNEROUX Madeleine (traductrice) & Nicolas QUINT (éd.) (2022). *Le P'chot Prince* [traduction en nouzerinois (Nouzerines, Creuse) du *Petit Prince* d'Antoine de Saint-Exupéry]. Neckarsteinach : Tintenfaß.

DUBAC Gérard (traducteur) & Nicolas QUINT (éd.) (2021). *Le P'ti Prince* [traduction en parler de Saint-Pierre-le-Bost (Creuse) du *Petit Prince* d'Antoine de Saint-Exupéry]. Neckarsteinach : Edition Tintenfaß.

GROBOST Henri (traducteur), Maximilien GUÉRIN & Nicolas QUINT (éds) (2020). *Le P'tit Princ'* [traduction en navois (Naves, Allier) du *Petit Prince* d'Antoine de Saint-Exupéry]. Neckarsteinach : Tintenfaß.

GUY Daniel, Marie-Claire RAZET-SALESSETTE (traducteurs) & Nicolas Quint (éd.) (2022). *Le P'tsë Prince* [traduction en nouzerinois (Nouzerines, Creuse) du *Petit Prince* d'Antoine de Saint-Exupéry]. Neckarsteinach : Tintenfaß.

LAROSE Pierre (traducteur), Michel BIDAUD, Maximilien GUÉRIN & Nicolas QUINT (éds) (2021). *Le P'tit Prince* [traduction en castro-vicecontamlien (La Châtre-Langlin, Indre) du *Petit Prince* d'Antoine de Saint-Exupéry]. Neckarsteinach : Tintenfaß.

MOUTET Henri (traducteur) & Nicolas QUINT (éd.) (2021). *Le P'tit Prince* [traduction en châtelois (Châtel-Montagne, Allier) du *Petit Prince* d'Antoine de Saint-Exupéry]. Neckarsteinach : Tintenfaß.

PRADEAU Guy (trad.) & Nicolas QUINT (éd.) (2021). *Le P'ti Prinsse* [traduction en nothois (Noth, Creuse) du *Petit Prince* d'Antoine de Saint-Exupéry]. Neckarsteinach : Tintenfaß.

SAINT-EXUPÉRY Antoine, de (2007 [1946]). *Le Petit Prince [original français]*. Paris : Gallimard.

6.3. Filmographie

GUILLEMAIN Franck (2020). Les parlers du Croissant [film documentaire] https://www.canal-u.tv/video/cnrs_ups2259/les_parlers_du_croissant.57913 (consulté le 14/08/2021). Paris / Villejuif : LLACAN (CNRS) / ARDIS.

Quint Nicolas
LLACAN – UMR 8135
CNRS / INALCO / EPHE
nicolas.quint@cnrs.fr

Chapitre 7

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

Philippe Boula de Mareüil, Marcel Courthiade[1], Frédéric Vernier

Université Paris-Saclay & CNRS, LISN

Le spectacle de cette uniformité universelle m'attriste et me glace

Alexis de Tocqueville (1848), *De la démocratie en Amérique*,
Tome 4/Quatrième partie/Chapitre 8.

[1] Marcel Courthiade (1953–2021) était professeur de rromani à l'INALCO au moment de la rédaction d'une grande partie de ce texte. Le 4 mars 2021, il est décédé à Tirana, quelques jours seulement après avoir enregistré les locuteurs de kajnas et de macédonien de Golo Brdo dont il est question dans ce chapitre ; quelques jours seulement après avoir rédigé quelques notes concernant cette dernière variété. Il devait, dans les semaines qui suivent la rédaction des lignes de la section 6.1, compléter l'analyse des langues balkaniques qu'il parlait pratiquement toutes — en plus de l'occitan et bien d'autres langues. Sa perte est grande pour la connaissance et fait cruellement défaut.

Abstract

To promote linguistic diversity, we propose to present a speaking atlas of regional or minority languages which, starting from Metropolitan France, has been extended to the French overseas territories, to languages without compact territory such as Rromani, as well as to other countries in the immediate vicinity of France. This online atlas <https://atlas.limsi.fr> allows visitors to listen to the same Aesop fable (and to read it) in over 900 versions. This work is the result of numerous field surveys and the development of an attractive interface. As confinement hardly lends itself to field linguistics, during the quarantine of 2020 we undertook to collect around forty translations of this fable, via the Internet, in minority languages /dialects of Europe.

We will describe their mapping and we will focus on some of the endangered languages collected, from various linguistic areas: Romance (Occitan, Moeso-Romanian and Aromanian), Finno-Ugric (Sámi, Meänkieli and Kven), and Slavic (Ruthenian, Moravian and Bunjevac). We will see that these languages raise common, controversial questions, due to their obsolescence. The heterogeneity of these languages, almost consubstantial with their minority state, will thus fuel a purist, fixist and essentialist discourse: "we don't say it like that" or, between two varieties of these languages, "they [the others] don't speak like us". Under these conditions, the writing of minority languages, which is crucial for their documentation and survival, raises important questions, which are shared by the languages selected here. Different solutions proposed will be discussed, which continue to be debated, in the *Oc* Gallo-Romance area (Provençal and Oriental Languedocien, from which we will start and which we will analyze in detail) as well as in the Balkans especially.

1. Introduction

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. Ces premiers (mais également

ces derniers) cantonnent ces dialectes à la sphère privée, à un usage familial si ce n'est familier et informel. Presque par définition, les « patois » — comme les linguistes n'osent plus guère les appeler — sont depuis les XIII^e^–XVII^e^ siècles considérés comme des langages bestiaux ou enfantins, inintelligibles et incorrects, rustiques et grossiers (Courouau, 2005). Pour Dauzat (1927), ce sont des parlers ruraux, socialement déclassés, qui ne disposeraient que d'un seul registre, celui de l'immédiateté et du quotidien. Ils constituent la langue du peuple, puisque les élites les ont abandonnés au profit de variétés plus prestigieuses. Rappelons la formule de Sainte-Beuve (1851) : « Je définis un patois une ancienne langue qui a eu des malheurs, ou encore une langue toute jeune et qui n'a pas fait fortune. » Les dialectes — terme moins investi péjorativement — seraient donc étrangers à l'idée de progrès ; ils seraient incapables de tout dire, d'exprimer des abstractions, de s'adapter au monde moderne.

Ces arguments, qui peuvent se manifester par une sorte de haine de soi (Fanon, 1952), rappellent le colonialisme du XIX^e^ siècle : « les races supérieures […] ont le devoir de civiliser les races inférieures » (Ferry, 1885). Ils sont aujourd'hui repris par les adversaires politiques des langues régionales, de droite et de gauche, qui consentiront seulement à certains éléments de langage mettant en avant la richesse de notre patrimoine linguistique (Martel & Verny, 2020). La patrimonialisation est un moyen de muséifier ces langues, de les renvoyer à un passé révolu et de ne rien entreprendre pour leur transmission (Martel, 2019). À cet argument s'ajoute celui de l'éclatement dialectal, de l'hétérogénéité et du manque de standardisation des langues minoritaires, qui les rendraient donc non-fonctionnelles. Pour sortir de ce qui apparaît comme un « cercle vicieux », on pourrait mettre des moyens, afin de doter ces langues de systèmes d'écriture unifiés, de grammaires, de manuels d'enseignement (Viaut, 2020). Cependant, la volonté politico-économique n'est pas là, le plus souvent.

En France, on se heurte aux sacrosaints deux premiers articles de la Constitution de la V^e^ République pour promouvoir la diversité linguistique, au nom de l'indivisibilité de la société française (Eysseric, 2005 ; Viaut & Pascaud, 2017)) — nous ne reviendrons

pas sur les récentes péripéties constitutionnelles. 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 de notre paysage linguistique, nous avons mis au point un atlas sonore qui, partant de la France hexagonale, a été étendu aux Outre-mer (Caraïbe, Océan Indien et Pacifique), aux « langues non-territoriales de France » comme le rromani, ainsi qu'à des pays comme la Belgique, la Suisse, l'Italie et la Péninsule ibérique, dans le voisinage immédiat de la France (Boula de Mareüil *et a*l., 2017, 2019a, 2019b. 2020). Cet atlas en ligne permet d'écouter et de lire une même fable d'Ésope dans plus de 900 versions, résultats de nombreuses enquêtes sur le terrain et du développement informatique d'une interface attractive.

C'est cet atlas linguistique que nous nous proposons de présenter dans cet article. Accessible en ligne à l'adresse https://atlas.limsi.fr, il nous a valu des centaines de courriers électroniques de réactions le plus souvent positives, parfois négatives, auxquels nous nous sommes efforcés de répondre. Pour qualifier ces commentaires, nous avons choisi le terme « épilinguistiques » (*i.e.* relevant du discours ordinaire, spontané, sur le langage, du type « le français est une belle langue », « sa forme la plus pure est celle de Tourraine » ou « l'occitan n'est pas une (vraie) langue ») plutôt que celui de « métalinguistiques », qui fait plutôt référence à des savoirs issus d'une réflexion sur le langage (Canut, 1998). Nous nous livrerons à une brève analyse du discours de critiques reçues, en nous limitant à un sous-domaine occitan (un territoire qui, chose exceptionnelle, porte un nom de langue) : provençal et languedocien oriental. Nous partirons de cette aire, dont nous décrirons les contours géographiques ainsi que des traits de prononciation, de grammaire et de vocabulaire observés dans une douzaine de points d'enquête, avant d'élargir la discussion aux langues minoritaires d'Europe.

Le confinement ne se prêtant guère à la linguistique de terrain, en effet, nous avons pendant la (première) quarantaine de 2020

entrepris de recueillir, via Internet, une quarantaine de traductions de la fable d'Ésope, dans des langues/dialectes minoritaires d'Europe, dont des langues sans territoire compact comme l'aroumain (Courthiade & Karamagkiola, 2013). Nous verrons que ces langues, objets de stigmatisation à l'instar du rromani (Courthiade, 2000, 2004, 2007, 2013, 2020 ; Canut, 2011), soulèvent des questions non moins polémiques que l'occitan (provençal). Le constat de l'obsolescence des langues minoritaires accroît les problèmes pour l'atlantographie linguistique et nous incite à repenser le statut des langues sans territoire compact. L'hétérogénéité de ces langues, presque consubstantielle à leur état minoritaire, va ainsi alimenter un discours commun, puriste, fixiste et essentialiste : « on ne dit pas ça comme ça » (Sallabank & Marquis, 2018) ou, entre deux variétés de ces langues, « ils [les autres] ne parlent pas comme nous ».

Les prochaines sections (2–4) sont consacrées à notre atlas sonore des langues régionales de France, qui fait suite à des travaux portant sur la Norvège (Almberg & Skarbø, 2002) et l'Italie (Romano, 2016) : plus précisément, nous nous centrerons sur les enregistrements en occitan (provençal et languedocien oriental), leur analyse épilinguistique et linguistique. Les section suivantes (5–6) ébaucheront une extension de cet atlas aux langues minoritaires d'Europe. Le travail relaté, empruntant initialement à l'anthropologie linguistique les enquêtes de terrain, s'est petit à petit rapproché des sciences participatives (en anglais, *citizen sciences*), utilisant les technologies de l'information et de la communication (TIC). La section 7 conclut et ouvre quelques perspectives.

2. Pour un atlas sonore des langues régionales de France : focus sur l'occitan

2.1. Protocole et points d'enquête

Depuis 2011 au moins, nous avons entrepris de faire traduire en langues régionales de France, entre autres choses, la fable d'Ésope « La bise et le soleil », utilisée par l'Association Phonétique Internationale (API) depuis plus d'un siècle pour décrire nombre

de langues du monde. Complémentaire de listes de mots, le protocole a été éprouvé par une longue tradition linguistique. Comme avant nous E. Edmont (Gilliéron & Edmont, 1902–1910), nous avons sillonné la France, le plus souvent en train, et avons trouvé un accueil extrêmement chaleureux : plus d'une fois, il est arrivé que nos informateurs viennent nous chercher à la gare et nous hébergent. Dans nos premières enquêtes (Corse, Provence, Languedoc), nous demandions à nos témoins de traduire à la volée cette fable, à partir du texte français sous les yeux. Petit à petit, les locuteurs ont commencé à écrire leurs traductions, et nous n'avons pas renâclé à récupérer leurs transcriptions. Même depuis 2014, année à partir de laquelle se sont multipliées nos enquêtes, certains n'ont pas jugé utile de le faire : ils n'ont pas démérité, selon nous, et il nous paraissait toujours intéressant d'éliciter de la parole qui sonne le plus naturellement possible.

En Provence et en Languedoc oriental, nous avons retenu une douzaine de points d'enquête, que l'on peut visualiser dans la Figure 1 et le Tableau 1. Nous ne nous attarderons pas sur Nice, représentée par les enregistrements de deux textes différents (l'un en graphie alibertine, l'autre en graphie mistralienne), le niçois s'éloignant à bien des égards, du reste du domaine provençal (Dalbera, 1994). Ces deux orthographes sont en effet concurrentes : la première, également dite « classique », est fondée sur les usages anciens des troubadours du Moyen-Âge, les travaux d'Alibert (1935) et du Conselh de la Lenga Occitana (CLO) ; la seconde, de type phonétisant et guidée par la prononciation rhodanienne (autour de Maillane), a été élaborée par Joseph Roumanille et Frédéric Mistral (Jouveau, 1980). Même si la graphie classique se veut plus englobante (Sumien, 2007) et est largement majoritaire, du moins hors de Provence (Lieutard, 2019), certains locuteurs sont fort attachés à la graphie mistralienne, et par respect pour eux, nous avons introduit une double graphie pour trois points d'enquête occitans de notre atlas : Sanary-sur-Mer, en Provence, Le Pont-de-Montvert et Domessargues dans les Cévennes.

2.2. Cartographie

Les locuteurs ont été cartographiés dans la ville dont ils estimaient parler la variété d'occitan, même si nous les avons

rencontrés ailleurs ou s'ils se sont enregistrés eux-mêmes : nous leur avons dès lors explicitement posé la question. Il s'agissait en Provence d'(anciens) enseignants d'occitan-langue d'oc ou de responsables associatifs ; pas de professeurs certifiés d'occitan-langue d'oc en Languedoc oriental, mais également des associatifs. Un consentement signé leur était demandé pour une libre diffusion de leur voix, dans lequel les témoins étaient invités à donner quelques renseignements à caractère autobiographique. Ceci nous a permis de nous assurer que nos locuteurs étaient natifs de leur lieu de résidence (comme celui de Marseille) ; ou bien, en cas de déménagement, nous les avons épinglés, sur la carte, à la ville dans laquelle ils ont grandi.

À ce problème de cartographie s'en ajoute un autre, dès lors que nous avons désiré dessiner les contours des aires dialectales (provençal, languedocien, etc.) comme apport d'information. Il y a des désaccords pour des zones de transition comme les départements du Gard et de Lozère (correspondant à quelques cantons près au Gévaudan) : pour nous en tenir au sous-domaine occitan qui nous intéresse ici, on peut utilement consulter les relevés du ThésOc (Oliviéri *et al.*, 2017). Sans clore le débat qui peut s'éterniser jusqu'à la mort des derniers combattants, nous avons fait passer la limite du languedocien au nord du Pont-de-Montvert, vers le domaine nord-occitan, et à l'est de Domessargues, vers le domaine provençal — pour ne citer que des localités où nous avons des points d'enquête (*cf.* Figure 1). Quant au domaine provençal, nous l'avons fait cesser à Roquebrune-Cap-Martin, où nous avons enregistré du dialecte monégasque (ligurien) : au-delà, au Nord-Est, on repasse au nord-occitan, à Menton, et au ligurien, à Saorge.

Figure 1 : Extrait de la carte https://atlas.limsi.fr pour le provençal et le languedocien oriental.

3. Analyse du discours en réaction à quelques enregistrements en occitan

3.1. De la magie de l'écriture

Avant d'aller plus loin dans l'analyse des données collectées, nous rapportons ci-dessous une douzaine d'extraits de courriers électroniques reçus depuis 2017, sans que leur numérotation ne corresponde nécessairement à l'ordre chronologique. Anonymisés ici (et reproduisant l'orthographe de leurs auteurs), les messages pouvaient être signés par des scientifiques, des militants politiques ou des associatifs. La plupart d'entre eux portent sur la graphie, objet de virulents conflits et de fétichisme s'il en est (Caubet *et al.*, 2002), mais pas seulement : il y a aussi une réification de ce qu'est un « bon locuteur » *vs* un « néo-locuteur ». Les conventions orthographiques, en effet, en même temps qu'elles visent la communication dans une langue, érigent également une barrière avec les autres langues. Elles circonscrivent en cela un territoire symbolique que la géographie humaine a bien analysé (Breton, 1974 ; Claval, 1992). Or les provençalistes les plus religieusement attachés à la graphie mistralienne, loyale envers les conventions françaises, n'ont pas pour ligne de mire le plus vaste espace occitan : leur territoire est la Provence, à l'intérieur de la France. L'occitan ne serait pour eux qu'une langue artificielle, et les

connotations politiques ne sont pas absentes de ces choix identitaires, tout aussi idéologiques que le projet occitaniste (Costa, 2012). Nous ne rapporterons pas le vif débat qui a tourné court avec un détracteur à qui nous avons proposé de contribuer à l'atlas sonore, qui assez vite nous a déclaré ne pas (assez bien) parler provençal.

(1) je trouve que transcrire la fable en graphie « occitane » dite classique en écoutant le parler de Maillane, pays de Mistral est une faute, faute que j'attribue à votre méconnaissance de la langue provençale. […] Les provençaux ne sont pas des occitans, la Provence n'est pas l'Occitanie et leur langue s'appelle le provençal et non l'occitan.

(2) nous remarquons que la transcription est occitane pour Marseille, Aix, Forcalquier et même pour Maillane

(3) A Maillane, patrie de Mistral, transcrire en graphie occitane recomposée ne vous choque pas ! Vous êtes chercheurs alors chercher ne vous laissez pas enfumer par l'idéologie occitaniste.

(4) je suis très étonné qu'à Maillane (pays de Frédéric Mistral) vous osiez mettre uniquement la graphie et la prononciation dite classique., alors que vous faites la différence dans les formes maritimes. Cela frise l'outrage, surtout qu'il a été prouvé que la modernisation de la graphie par Mistral s'est essentiellement appuyée sur le parlé rhodanien. Je ne doute pas que vous saurez corriger cette erreur.

(5) Le Provençal est une langue à part entière et la pression de l'éducation nationale et des enseignants à nous apprendre l'Occitan ne nous plaît pas du tout.

(6a) Le choix de la graphie alibertine (ou occitane) pour la retranscription de la langue provençale est inadapté à la lecture de la langue et incomprise des provençaux. La graphie utilisée en Provence à 95% est la graphie dite "mistralienne". Elle convient à l'ensemble des dialectes provençaux , sa qualité et sa beauté récompensées par un prix Nobel en 1904 en témoignent.

(6b) Je ne vous apprendrai pas que la graphie mistralienne peut être utilisée pour toute les langues d'oc (cf le Trésor du Félibrige) et permet donc elle aussi les comparaisons... Le choix d'utiliser la graphie languedocienne en Provence n'est qu'un choix politique

(7) Pour des chercheurs du CNRS, il est étonnant que vous ne sachiez pas que le seul Prix Nobel de Littérature a été donné en 1904 à Frédéric Mistral pour la totalité de son œuvre en Provençal et en particulier bien sûr son Trésor du Félibrige.

(8) Pourquoi ne pas avoir fait confiance aux structures compétentes en place comme le Félibrige….?

(9) Le provençal de Maillane […] est parfaitement lu mais la transcription est claie de fautes s'en est honteux . je ne peux pas faire confiance à votre site et je le ferai savoir!!!

(10) ça sent la connerie occitaniste

(11) la voix de Marseille ne dit pas le parler de Marseille ; les autres voix provençales sont bien approximatives (sauf Forcalquier, Sanary et Nice). Sans compter les fautes multiples : *bufar* et *bofar* dans le même texte montpelliérain (il faut choisir) ; […] emploi du passé composé « à la française » au lieu du prétérit (Maillane, Caromb, ce qui est une énorme faute)

(12) [Tel locuteur] huguenot!) préfèrerait ne pas apparaître dans l'Atlas plutôt que de voir son texte écrit en occitan!

(13) Donnerait-on comme exemple de langue française, surtout quand on la localise territorialement, le parler d'un étranger qui la maîtrise imparfaitement, même s'il parvient à bien communiquer dans la population locale ? […] Ne serait-ce pas confondre hybridation naturelle dans un milieu communiquant et dégradation de la langue dans la perte d'un usage de communication réelle ? Faut-il alors exclure du modèle la jeunesse ? Faut-il que les modèles de la langue s'arrêtent aux derniers locuteurs naturels, quitte à se cantonner à des voix d'outre-tombe ou de *grabataires?* On peut en débattre à l'infini.

On notera que la graphie alibertine est dénommée « occitane », implicitement donc rejetée comme « étrangère » dans les premières réactions polémiques, partisanes de la graphie mistralienne (1) (2) (3) : de fait, contrairement à cette dernière, la graphie classique ne fait pas l'objet d'un culte de la personnalité envers son instigateur — convaincu de collaboration après la Seconde Guerre mondiale. En contrepoint, la graphie mistralienne est dotée de la qualité de « moderne » par le rédacteur du commentaire (4)[2]. Le commentaire (6a) précise la chose, avançant un chiffre de 95 % (qu'on pourrait descendre à 50 %) difficile à prouver, pour le pourcentage d'utilisateurs de cette graphie. L'argument du prix Nobel décerné par l'Académie suédoise revient également, même s'il n'a pas récompensé une graphie (la seule à l'époque) mais l'auteur d'une œuvre littéraire — de même que le prix Goncourt, qui par deux fois a été attribué à Romain Gary, un néo-locuteur du français. Quant au commentaire du même auteur

[2] Suite à un échange avec l'auteur de ce commentaire et avec son concours, en 2021, nous avons ajouté une double graphie classique-mistralienne pour les deux enregistrements de Maillane et de Caromb.

(6b) et aux extraits suivants (7) (8), ils mentionnent le Félibrige, noble institution fondée par Mistral dont nous avons été en contact avec quelques majoraux, sans beaucoup de succès.

3.2. Haro sur les néo-locuteurs

On retrouve dans plusieurs commentaires le même tropisme pour les « locuteurs naturels » qui parleraient une langue pure (même si on ne sait pas de qui ils sont les enfants naturels) opposés aux maléfiques néo-locuteurs qui la travestiraient honteusement, affreux occitanistes pétris d'idéologie. Il n'est pas rare que des gens qui ne se classent pas à gauche de l'échiquier politique, pour ne pas dire plus, prétendent être pragmatiques et nient avoir une idéologie, contrairement à ceux à qui ils s'opposent (Alain, 1925). La suspicion d'illégitimité, l'essentialisation de la langue, la nostalgie d'un passé idéalisé, la quête romantique d'authenticité ne sont pas nouvelles et ont bien été étudiées (Bucholtz, 2003 ; Costa, 2010)[3]. Le délitement du monde rural, la perte de l'usage social de l'occitan, la généralisation de l'enseignement du français et l'alphabétisation dans cette langue, l'explosion des médias de masse accessibles à tous, tout concourt à ce que, nécessairement, les locuteurs d'occitan nés en France après les années 1950 sont au minimum bilingues français/occitan, et il n'y a aucune raison pour que cela ne change au fil du temps. L'influence de la langue dominante (ou *Dachsprache*, « langue toît ») est inévitable dans une situation de diglossie : on peut le regretter, mais on ne bâtit pas une reconquête de l'usage de la langue sur des remords ou de la rancœur.

Évidemment, une personne de plus de 90 ans parle un provençal différent ; sa voix est plus éraillée par les années : on percevra chez elle un fort accent, qui s'est forgé à une époque où l'on entendait encore parler cette langue partout autour de soi, où toute personne âgée de plus de 50 ans baragouinait un français approximatif voire ne le parlait pas du tout. Mais la comparaison vaut pour un grand

[3] « Concernant [...] le bout d'occitan enregistré au Pont de Montvert... [...] il y avait une façon de raconter les histoires [...] ; mon grand-père (locuteur natif) avait cette intonation particulière, [...] c'est un bon exemple de cette langue régionale qui, pour moi, est un peu une madeleine de Proust et qui, quand je l'entends [...], ne résonne pas tout à fait avec les souvenirs de mon enfance... »

nombre de langues du monde qui ont vu la société se transformer à une vitesse vertigineuse depuis un siècle. Toutes les langues évoluent ; personne ne parlera sans doute plus la langue d'oc comme il y a deux générations, tout comme à l'époque on ne parlait pas comme deux générations auparavant. Victor Gelu (1856) le pointait déjà, lui qui se lamentai : « L'idiome provençal se meurt. » *Nihil novi sub sole,* comme on disait en patois latin.

On pourra donc répondre ceci aux gardiens de la langue inaltérée de leurs glorieux devanciers et à certains arguments spécieux qui ne font que conforter l'hégémonie du français : il est dommage qu'au lieu d'essayer de développer l'usage de sa langue, on aille stigmatiser ceux qui font l'effort de la parler, qu'au lieu de se réjouir d'une certaine reconquête de la langue minorisée on vienne rechigner sur tel ou tel trait linguistique. Par exemple, l'usage du passé composé plutôt que du passé simple, considéré comme une « énorme faute » dans le commentaire (11) est une évolution que l'on retrouve dans tout le nord et le centre de l'Italie : on ne saurait donc l'imputer à la seule influence du français — langue dans laquelle on ne songerait pas à parler de « faute ».

L'assignation d'un enregistrement à un point géographique (avec une présentation ville par ville en France hexagonale ou des glossonymes au-delà) suggère que le locuteur affiché est « représentatif » de sa variété, même si nulle part nous ne le laissons croire explicitement. Depuis Labov (1976), on sait que le concept de locuteur représentatif est scientifiquement mal défini pour de grandes villes qui brassent nombres de différences et tendent à escamoter certaines spécificités traditionnelles. Gage de spontanéité, les locuteurs que nous avons enregistrés à Montpellier, Maillane, Caromb, Marseille et Aix-en-Provence ont eu le mérite de traduire directement, à partir du texte français qu'ils avaient sous les yeux, la fable « La bise et le soleil », contrairement aux autres locuteurs qui ont écrit leurs traductions. Ils parlent couramment, si ce n'est quotidiennement, l'occitan et, malgré tous nos efforts, nous n'avons pas trouvé de « meilleurs » locuteurs dans leurs villes. Ajoutons que les locuteurs de ces trois dernières villes, encore actifs, sont plus jeunes que la moyenne de nos informateurs (70 ans). Nous n'étions pas mécontents non plus de montrer que des « jeunes » continuent à faire vivre la langue,

chose qu'il ne nous a guère été permise de faire dans le domaine d'oïl.

Si la magnification hyperlocaliste de la différence plutôt que de ce qu'on a en commun est fréquente chez les locuteurs de langues minorisées, la gêne de certains anciens à écouter des jeunes parler leur langue régionale peut trouver une autre source : la violence pas uniquement symbolique dont ils ont été victimes. Ayant grandi à l'époque où toutes sortes de « symboles » ou autres objets pervers d'humiliationet de délation étaient mis en place à l'école pour dissuader les élèves de patoiser (Walter, 1988), ceux qui n'ont pas transmis la langue peuvent sentir de la culpabilité quand le flambeau est repris par leurs petits-enfants. D'où certaines réactions véhémentes. Évidemment et dans les cas extrêmes, si l'on n'interagit plus dans ces langues sur les marchés, lors des foires ou dans les bals, la posture selon laquelle « on ne se comprend pas [d'un village à l'autre ou entre générations] » gagne en pertinence. Elle demande toutefois une analyse plus proprement linguistique.

4. Analyse linguistique de quelques points d'enquête en occitan

4.1. Provençal

Le commentaire (13) soulève de bonnes questions autour de la norme de fait et de l'appropriation du modèle par des apprenants qui n'ont pas communiqué au quotidien avec les générations précédentes sans discontinuité. Sans jugement de valeur, il ne fait pas non plus l'économie de conférer un statut référentiel spécifique à la dernière génération de locuteurs dits « naturels », qui auraient commencé à grandir en milieu rural avant la fin de la guerre. Le cadre de la carte ne s'accommode pas (facilement), cependant, de ces exigences, formulées sous le nom de « Non-mobile Old Rural Males » (NORMs) par Chambers et Trudgill (2004). Aux problèmes pratiques qui se posent pour trouver de tels locuteurs acceptant d'accomplir une tâche de traduction enregistrée s'ajoutent en effet des contraintes juridiques, interdisant ou compliquant la diffusion d'informations qui aideraient à identifier lesdits locuteurs. Les différences que l'on peut observer entre les versions recueillies de

la fable d'Ésope peuvent ainsi légitimement être interprétées comme relevant de la variation diatopique.

Au niveau phonétique/phonologique, la prononciation du [r] intervocalique semi-roulé avec un seul battement de la langue, très caractéristique du provençal hors du rhodanien moyen et de Nice (Bouvier & Martel, 1975–1986), y est bien présente, comme pour les locuteurs de Marseille et d'Aix-en-Provence, mais de façon parfois irrégulière. Le [r] semi-roulé semble ignoré du locuteur de Maillane, mais on est près de l'aire où ce phénomène a disparu depuis longtemps. Ce locuteur représente une langue bien maîtrisée, comme la locutrice de Caromb, mais là nous sommes dans une zone sans [r] apical. Par ailleurs, la diphtongaison du /ɔ/ accentué (*fòrt, fòrça*) qui est si typique de la Provence non-rhodanienne jusqu'à Nice est absente de l'extrait du locuteur aixois, et irrégulièrement perceptible chez celui de Marseille, tandis qu'elle est très nette chez ceux de Sanary-sur-Mer, Antibes et Nice.

Au niveau morphologique, certaines variantes répondent à de réelles différences majoritaires entre sous-domaines, mais il n'en est pas moins vrai que d'autres relèvent de choix personnels qu'il serait imprudent d'expliquer par un ancrage local. Relèvent du premier cas la distinction pour « chacun » entre *chascun* (plutôt intérieur) et *cadun* (plutôt maritime) qui se constate effectivement dans les enregistrements ; le choix plutôt alpin, pour « mettre », de *botar* (Forcalquier) par opposition à *metre* ailleurs ; la distinction [e]~[i] entre les formes de pronoms personnels inaccentués *se* (intérieur) et *si* (maritime), très caractéristique, mais variable à Marseille. Dans l'extrait aixois figure en outre une forme *vegèron* (« ils virent »), qui a pu surprendre car il semble que partout en Provence on dise *veguèron* (Barthélémy-Vigouroux & Martin, 2017). Enfin, des tournures relèvent d'une volonté de donner un tour plus affectif ou populaire à telle expression : *lo soleu te comença a lusir* (littéralement « le soleil te commence à briller », Forcalquier) ; *rescaufat que rescaufat* (« réchauffé », Sanary-sur-Mer), ce qui ne marque aucune spécificité locale, mais seulement une interprétation personnelle du registre du texte à traduire. Et si le locuteur d'Antibes a recours au passé composé (comme celui de Maillane), présenté à nous comme un « Antibois de souche » par un félibre majoral, il n'est pas suspect d'être un néo-locuteur.

Au niveau lexical, le rapport au territoire des formes observées n'est pas aussi motivé qu'au niveau de la prononciation ou de la grammaire, et c'est là que la cartographie peut paraître sans objet. Voyons quelques exemples :

Aire	Commune	"La bise et le soleil se disputaient" : traduction
Provençal	Caromb (84)	La bisa e lo soleu se garrolhavan
	Forcalquier (04)	Se disputavan l'aura e lo soleu
	Maillane (13)	La bisa e lo solèu s'escharpavan
	Aix-en-Provence (13)	La bisa e lo soleu se garrolhavan
	Marseille (13)	L'aura e lo soleu se disputavan
	Sanary-sur-Mer (83)	Lo mistrau e lo soleu si debequignavon *Lou mistrau e lou soulèu si debequignavon*
	Antibes (06)	Lo ventolet e lo soleu si chamalhavan
Languedocien oriental	Le Pont-de-Montvert (48)	L'aura e lo sorelh se carcanhavan *L'auro e lou sourel si carquignàbou*
	Sainte-Croix-Vallée-Française (48)	L'ura e lo sorelh se carcanhavan
	Domessargues (30)	La rispa e lo sorelh se carcanhavan *La rispo e lou sourel se carcagnàvou*
	Montpellier(34)	La cisampa e lo solelh s'atissavan
	Sète (34)	La cisampa e lo sorelh se fasián "au pus fòrt la pelha"

Tableau 1 : Début de la fable « La bise et le soleil » dans 7 points d'enquête en provençal et 5 points d'enquête en languedocien oriental (avec les codes des départements correspondants). Les locuteurs sont 10 hommes et 2 femmes (Caromb et Le Pont-de-Montvert). Les transcriptions en graphie mistralienne sont précisées en italiques.

- « se disputaient » est traduit par *s'escharpavan* (Maillane), *se disputavan* (Marseille, Forcalquier), *se chamalhavan* (Antibes), *se garrolhavan* (Caromb, Aix-en-Provence), *se debequinhavan* (Sanary-sur-Mer). Il n'y a rien de territorial dans tout cela : les trois premiers verbes sont soutenus par le français, le quatrième est la forme courante du provençal dans un registre neutre, le dernier est aussi connu partout, avec une nuance plus agressive ;
- « assurer » est dit *afortir* à Maillane, *assegurar* à Caromb, Aix-

en-Provence, Sanary-sur-Mer et Marseille ; mais les deux verbes sont courants partout, l'un reposant sur la métaphore d'une affirmation renforcée, l'autre sur la sûreté et la garantie de celui qui affirme ;

- « voyageur » devient *voiatjor* (Maillane), *vo(a)iatjaire* (Caromb et Antibes), *viatjaire* (ailleurs) : le premier est évidemment un francisme bien acclimaté, le seul en usage aujourd'hui, les autres des reconstitutions plus ou moins puristes avec hésitation sur le radical qui témoigne de leur caractère artificiel ;
- « enveloppé » voit défiler six verbes différents : *agolopat* (Maillane et Caromb), *engolopat* (Antibes), *envertolhat* (Sanary-sur-Mer), *plegat* (Forcalquier), *envolopat* (Aix-en-Provence) et et *enviroutat* (Marseille). Tous sont valables et en usage général, chacun donnant une nuance différente ; on en retrouve deux ensuite pour exprimer le geste du voyageur qui resserre son vêtement (*agolopat* à Sanary-sur-Mer, *s'envertolhar* à Forcalquier), avec un croisement de locuteurs, qui montre bien leur caractère globalement interchangeable quand on n'approfondit pas les nuances ;
- le soleil qui brille va *brilhar* à Maillane, Aix-en-Provence, Marseille — c'est le mot courant et ancien ; il va *lusir* à Sanary-sur-Mer, Antibes et Forcalquier : le premier évoquant des éclats intermittents qui se succèdent rapidement (on le rattache au radical de « pirouette »), le second une source de lumière égale et continue ; enfin *dardalhar* (que l'on trouve à Caromb) exprime plutôt les agressions d'un soleil qui pique la peau comme un ardillon. Naturellement, les trois mots sont répandus sur tout le territoire provençal (et largement ailleurs) ;
- « tomber d'accord » est généralement traduit par *tombar d'acòrd*, avec la variante un peu archaïsante *d'acòrdi*, hormis à Sanary-sur-Mer, dont le locuteur emploie *fèron pache* « conclurent un marché », choix qui doit à la volonté de privilégier une autre tournure tout aussi répandue et un peu plus populaire ;

Passons *bisa*, *aura* et autre *mistrau*, discutés dans Boula de Mareüil

et al. (2017), dont on peut dire autant, pour nous arrêter sur *bofar* (forme universelle du provençal) et la forme *bufar* des locuteurs marseillais et aixois. Cette forme caractérise le languedocien et des parlers septentrionaux : nous y reviendrons.

4.2. Languedocien oriental

Au niveau phonético-phonologique, on a affaire au Pont-de-Montvert à un parler « languedocien en *cha* » (Ronjat, 1913). L'aboutissement du CA- latin est bien *cha* dans *chaminaire* (« chemineau ») et *reschaufat* (« réchauffé »), alors qu'on relève également *carcanhar* (« quereller »). Plus au Sud, on a déjà (*r)escaufat.* à Sainte-Croix-Vallée-Française et à Domessargues. Même le locuteur de Domessargues mentionné dans le commentaire (12), qui nous a été présenté comme « très représentatif » du parler allésien, du pays de Lédignan et du piémont gardois, restitue une prononciation partiellement francisée — ce qui accentue la ressemblance avec le provençal rhodanien lorsque celui-ci aussi subit l'influence du français. Les francismes les plus manifestes, dans l'enregistrement du Gard (près du domaine provençal) sont la neutralisation des deux rhotiques intervocaliques /r/ et /ʀ/ en un seul type /ʀ/, l'amuïssement de certaines consonnes finales (ex. [mɛ] pour traduire le français *mais*) et le /e/ fermé qui passe à [ə].

Le /r/ apical est au reste quasiment absent chez tous nos locuteurs de languedocien oriental, y compris dans les cas de rhotacisme tels que *sorelh* (« soleil » < SOLICULU(M) à comparer au *soleu* provençal) chez les trois Cévenols. À noter pour ce mot, au siècle dernier, que les formes [sulel] (au Pont-de-Montvert), [suʁel] et [surel] avaient été relevées en Lozère (Boisgontier, 1981–1986 : 48). Quant à l'aboutissement du groupe latin -ELLU(M) il est *-èl* et non *-èu* : on a *mantèl* dans les Cévennes, face à *mantèu* à Montpellier (et en Provence), pour traduire le français « manteau ». Le locuteur de Montpellier prononce également [ɔ] pour la finale post-tonique héritée du A latin et [y] pour le graphème <u>, contrairement à celui de Sète, qui maintient la finale latine [a] et prononce [ø] pour /y/ (ex. *bufar* [bøfa] « souffler »). Ceci a été reproché violemment à ce locuteur montpelliérain ; mais précisons que, pour avoir enregistré une

dizaine de locuteurs à Montpellier, il ne nous a pas été donné d'entendre ces prononciations sétoises, qui appartiennent également, historiquement, au « vrai parler montpelliérain » (Boisgontier, 1981–1986 : 34).

Au niveau morphologique, on a dans nos 5 points d'enquête en languedocien oriental des pluriels en *–s* pour d*els* — voire *das* (« des ») à Domessargues et à Sète — et non en *dei* comme en provençal. Comme en provençal, en revanche, on a l'amalgame *dau* pour « du » à Montpellier et à Sète, ainsi que les conditionnels (et imparfaits des 2^{e} et 3^{e} groupes) en [je] à Domessargues et à Sète (ex. *capitariá* [kapitarje] « arriverait », transcrit *capitariè* en graphie mistralienne). De façon intéressante aussi, « ils ont vu » est traduit par le prétérit *veguèron* partout dans nos points d'enquête du Languedoc oriental sauf à Montpellier, où l'on a *vegèron*, une forme que l'on retrouve non loin en languedocien occidental.

Au niveau lexical, nous serons plus brefs que pour le provençal. Relevons en passant deux nouveaux mots, *rispa* et *saile,* pour désigner respectivement le vent de « bise » et le « manteau » ou la cape du voyageur dans les Cévennes. À Montpellier et à Sète, le nom du vent est *cisampa*, que l'on retrouve plus à l'ouest dans cette sous-région : un vent froid et sec qui va *bufar* (« souffler ») dans ces deux points d'enquête, certes à côté de *bofava* (« soufflait ») à Montpellier. Cette forme critiquée dans le commentaire (11) notamment se retrouve dans les Cévennes, où l'on a *bufar*/*bufava* à Sainte-Croix-Vallée-Française, *bofar*/*bofava* ailleurs. L'hésitation peut donc être simplement le témoin d'une variation qui est le lot de toutes les langues minorisées.

5. Vers un atlas sonore des langues minoritaires d'Europe

5.1. Langues collectées et transcriptions

Plus récemment, nous avons élargi ce travail aux langues minoritaires d'Europe, dont nous avons cherché à contacter des locuteurs par Internet, les invitant à s'enregistrer eux-mêmes, les téléphones portables, notamment, donnant aujourd'hui de très

bons résultats. Non seulement la crise sanitaire nous a empêchés de faire des enquêtes sur le terrain, mais encore aller des Îles Féroé au Grand Nord scandinave, des Pays baltes aux Balkans, aurait été dispendieux et coûteux en temps. Le confinement nous a également permis de rajeunir la moyenne d'âge de nos témoins, des collègues linguistes nous proposant d'enregistrer leurs filles, locutrices natives des langues/dialectes qui nous intéressaient. Envoyant une dizaine de courriers électroniques par jour, de fil en aiguille, à des relations de relations, ce fut plus rapide que prévu. À partir de langues dominantes comme l'anglais, l'allemand, etc., nous avons ainsi reçu des traductions de la fable d'Ésope dans :

- 5 langues celtiques : gallois, cornique, gaéliques écossais, mannois et irlandais ;
- 5 langues finno-ougriennes :sámi (anciennement nommé lapon), võro, meänkieli et kven (langues fenniques respectivement d'Estonie, de Suède et de Norvège) et hongrois sicule (székely de Transylvanie) ;
- une langue balte de Lettonie, le latgalien ;

une langue turcique de Moldavie, le gagaouze ;

- une prononciation restituée du grec ancien d'après les préconisations d'Allen (1987), clin d'œil à Ésope (que nous avons cartographiée en Attique) ;
- 20 langues ou dialectes germaniques : féroïen, gutnisk (vieille langue de l'Île de Gotland), scots (deux variétés des Îles Shetland et de la frontière écossaise), northumbrien (également au Royaume-Uni), frison, limbourgeois, groningois et zélandais (tous les quatre aux Pays-Bas), saxon, berlinois, bas saxon (Plattdüütsch), palatin et hessois (deux variétés franciques), souabe et vorarlbergeois (deux variétés alémaniques), tyrolien, bavarois et styrien stoan (ces quatre dernières variétés cartographiées en Autriche), et vilamovien (variété silésienne de la petite ville de Wilamowice, au sud de la Pologne) ;
- 11 langues slaves : sorabe (Allemagne), cachoube (Pologne), morave (Tchéquie), ruthène (Slovaquie), biélorusse (langue officielle mais minoritaire de Belarus, comme l'est le gaélique en Irlande), trasianka (variété mixte russe-biélorusse), surzhyk (variété mixte russe-ukrainienne), bunjevac (Voïvodine), goran (Kosovo), kajnas et macédonien de Golo Brdo (Albanie) ;

- 5 variétés de langues balkano-romanes : aroumain, istro-roumain (dans les deux variétés žejanski et vlaški), moéso-roumain (de l'ancienne province romaine de Moesia superior, dans l'actuelle Serbie) et roumain de Transylvanie, qui viennent s'ajouter à des enregistrements réalisés et transcrits antérieurement, dans :
- 7 (autres) langues sans territoire compact : rromani, yiddish, judéo-espagnol (3 variétés), arménien occidental et même espéranto (cartographié à Białystok, ville natale de son créateurr Zamenhof, et indiqué par une étoile verte, symbole de cette langue). Des contacts ont enfin été pris pour des variétés albanaises et turciques, notamment chez les Ahkalis (Courthiade, 2000).

Même si cela ne vaut pas la richesse des rencontres personnelles, chose importante quand on se pique de faire des sciences humaines, les échanges que nous avons eus à distance ont été l'objet de belles anecdotes, telle cette locutrice de meänkieli qui s'est répandue en excuses pour le retard de sa réponse : c'était que, dans son troupeau de 48 rennes, les femelles avaient mis bas et qu'il fallait bien s'en occuper. Des publications scientifiques pourront par ailleurs être consultées sur les langues sám (Picard, 2000) et fenniques (Léonard, 2013 ; Ridanpää, 2018; Keränen, 2018). Pour les langues slaves, Mladenović (2001) explicite les choix orthographiques du goran, dans la double graphie cyrillique et latine, alors que pour le kajnas (parler « comme nous » d'Albanie également apparenté au macédonien), nous avons opté pour l'alphabet latin comme dans Courthiade (1988). Une étude, qui plus est, a été publiée à partir de la fable d'Ésope « Severák a Slunce » en tchèque prononcé par des locuteurs de Bohême et de Moravie (Šimáčková *et al.*, 2012), dont nous nous sommes inspirés.

Les langues balkano-romanes, quant à elles, ont été transcrites en alphabet latin, les mots serbes du moéso-roumain étant laissés en alphabet cyrillique, afin de rendre immédiatement visible à l'œil la part des emprunts — cette proportion s'élevant à 13 % dans la traduction de la fable « Северни vânt şĭ soareli ». Pour l'aroumain, une autre codification a été adoptée (Cunia, 1999). Pour l'istro-roumain, que nous avons reçu dans deux variétés (žejanski et vlaški), l'orthographe est celle qui a été promue dans un projet d'aménagement linguistique en Croatie <https://www.vlaski-

zejanski.com/en/jezicne-lekcije/learn-language-lessons-writing-3>. Pour le dialecte transylvain que nous donnons en comparaison avec les variétés balkano-romanes hors de Roumanie, a été utilisée l'orthographe étymologique de l'auteur, basée sur les orthographes de 1835–1918 (Staelens, 2019).

5.2. Discours épilinguistiques : focus sur quelques langues slaves

Les discours épilinguistiques ont été tout aussi saisissants qu'en France, où l'on a vu les outrances d'un certain sécessionnisme linguistique provençaliste. Qu'il s'agisse de la dépréciation de variétés minoritaires (*cf.* introduction), de leur essentialisation ou de leur hiérarchisation, on retrouve les mêmes ressorts : ainsi, en (14), une réaction à propos du morave ; en (15), une réaction toute autre à propos du bunjevac, parler ikavien d'un groupe ethnique catholique de Serbie ; ou encore, en (16) et (17a), concernant le rapport entre le ruthène et l'ukrainien ou en (17b) et (17c) concernant une traduction reçue en ruthène de Roumanie (non enregistrée ni cartographiée).

(14) Il n'y a pas un dialecte morave, mais plusieurs, quoiqu'il y ait des traits communs. On distingue quatre groupes. Traduire la fable en "morave", cependant, ne me semble pas possible. Il s'agit de variétés qui existent dans leur forme parlée, et ne remplissent pas toutes les fonctions que remplit une langue officielle. On peut trouver des textes "traduits" dans un parler local, mais c'est plutôt pour amuser.

(15) We are very happy that you have put the Bunjevac language on the mat of the European language. For centuries, we have fought to make it equal to other languages

(16) l'explle ruthène et l'ukrainien sont deux langues différentes. Ce sont deux peuples différents.

(17a) vous savez que notre nation sont autonome et autochtonne, seulement pendant le régime socialistique, nous, des Routhenes, nous sommes été interdit comme la nation. Apres le 1990 nouv pouvons exister librement - en Slovaquie

(17b) ce text en ruthene, ce n´est pas un text ruthene, c´est mélange de toutte

(17c) A oui, c´est la traduction de Roumanie? C´ est interessant enfin pour moi, Je ne comprend toutte de cette traduction, mais c´est interessante que en Roumanie vit quelq´un qui sait traduit un text en routhene

Le visiteur de l'atlas sonore pourra apprécier qu'en remplaçant les <i> du bunjevac par des <e>, on est très proche du serbe standard (en alphabet latin), mais il ne nous appartient pas de juger les sentiments identitaires des locuteurs bunjevci, s'inscrivant dans un processus d'individuation (Djordjević, 2013). C'est tout le débat entre catégorisations *émique* (intérieure au groupe) et *étique* (extérieure, provenant du chercheur) avec, face à un continuum linguistique, des arguments scientifiques qui peuvent s'opposer aux idéologies nationalistes. De même pour le ruthène : si la traduction de la variété de Slovaquie est assez différente de l'ukrainien standard (lequel ne connaît pas les lettres cyrilliques <ы>, <ё> et <етц>), des habitants d'un même village, originaires de la même région (la Galicie orientale, à l'ouest de l'Ukraine actuelle), vont s'auto-dénommer Ruthènes ou Ukrainiens, selon l'époque à laquelle leurs ancêtres se sont déplacés — l'individuation de la communauté ruthène, avec une langue littéraire propre, s'est faite essentiellement durant le siècle dernier (Djordjević-Léonard, 2014). Le mot « Ukraine » est relié étymologiquement à 'kraj' « extrémité ». Or, comme l'écrit V. Saïdi (2009) à propos de la nomination de la langue ukrainienne (terme apparu à la fin du XIX^e^ siècle, alors qu'on parlait à l'époque de « petit russien »),

> " la langue ne peut pas exister sans nom, car le nom « détermine l'appartenance de cette langue à telle ou telle nation ; si le nom n'existait pas, on l'inventerait tôt ou tard », « la langue sans nom appartient à un petit peuple, qui est peu important et ne représente pas une nation ». Une telle conception n'est pas propre qu'au monde slave. "

On ne saurait mieux dire, puisque nommer, c'est faire exister : que l'on songe au provençal *vs* occitan, au moldave *vs* roumain (Bochmann, 2022)… Sur les querelles d'alphabet et les questions de possible ethnogénèse, le lecteur pourra se reporter au très bon site « L'aménagement linguistique dans le monde » de l'Université Laval <http://www.axl.cefan.ulaval.ca/>. Dans la page dédiée aux langues minoritaires d'Europe de notre site, nous avons indiqué des glossonymes (en anglais) et non des localités comme nous l'avons fait en France et dans son voisinage immédiat. On sait que le terme même de macédonien, par exemple, est encore contesté

par certains, bien qu'il renvoie à une langue littéraire qui a vu le jour au milieu du XIX^e^ siècle, à côté du bulgare (Garde, 2004). Même si des étiquettes comme cachoube, ruthène ou macédonien de Golo Brdo peuvent suggérer une certaine atomisation de tel ou tel diasystème slave occidental, oriental ou méridional qui peut être instrumentalisée, nous en sommes conscients, notre but dans cet atlas sonore est avant tout illustratif. L'Atlas UNESCO des langues en danger dans le monde donne d'autres éléments sur le degré de vulnérabilité de ces langues (Moseley, 2010).

5.3. Cartographie

La cartographie a fusionné deux sources d'informations : (i) un découpage assez fin de la France et des pays limitrophes (Italie, Belgique, Suisse, Espagne, etc.) en aires dialectales ; (ii) le découpage du reste de l'Europe, jusqu'aux frontières de la Russie et jusqu'aux bords du Bosphore en Turquie, selon des aires linguistiques plus vastes. Les plus de 700 points d'enquête du premier ensemble (i) sont uniquement accessibles en ouvrant les cartes des pays correspondants — ce que permet un clic à l'intérieur des cadres prévus à cet effet ; mais la carte de l'Europe permet de visualiser l'étendue d'aires linguistiques transnationales comme celle du francoprovençal, éclaté entre la France, l'Italie et la Suisse. Pour le deuxième ensemble (ii), nous avons retenu et tracé les contours des familles, branches ou groupes de langues rapportés dans la Figure 2 et (en français) dans le Tableau 2. Des étiquettes sur trois lettres et des jeux de couleurs ont été associés : dans les jaunes pour les langues germaniques (jusqu'à l'orange pour le groupe anglo-frison), vert pour les langues celtiques (y compris le breton), turquoise pour les langues (balkano)-romanes, rouge pour les langues slaves, bleu pour les langues finno-ougriennes, ce qui rend immédiatement visible la discontinuité de ces derniers domaines. D'épais trais noirs soulignent de plus les contours des domaines qui ne sont pas indo-européens, qu'ils soient finno-ougriens, turciques ou basque.

Une signalétique particulière est prévue pour les poches linguistiques enclavées dans des aires distinctes. À la différence du sorabe en Lusace (Allemagne) et du goran au Kosovo, certains îlots linguistiques sont en effet trop petits pour que leurs surfaces soient

visible, à l'échelle de l'Europe. Il en va ainsi du kajnas — dont sans doute le dernier locuteur a été enregistré (Courthiade, 1988) — et du macédonien de Golo Brdo (Albanie), que nous avons fait figurer par des carrés rouges rappelant l'aire slave. Il en va ainsi, également, du vilamovien ou *wymysorys* (Anderson & Król, 2016) et du yiddish (cartographié à Cracovie), marqués de carrés jaunes en Pologne. Il en va ainsi, encore, des deux villages istro-roumains et de l'aroumain à l'intérieur du domaine slave — indiqués par des carrés dont la couleur turquoise rappelle celle de la Roumanie, sur notre carte. Il en va ainsi, enfin, du rromani, qui mérite pleinement son statut de langue sans territoire compact, dans la mesure où il se trouve éclaté dans trois aires linguistiques différentes : hellénique, balkanique (albanaise) et slave.

Code	Langues	Code	Variétés germaniques
cel	Celtiques	sca	scandinaves
fin	finno-ougriennes	anf	anglo-frisonnes
pal	Baltiques	pla	Plattdüütsch
hel	Helléniques	ndl	néerlandiques
sla	Slaves	lim	limbourgeoises
rom	Romanes	fra	franciques
alb	balkaniques	mde	germaniques centrales
tur	Turciques	als	alémaniques
ind	autres langues indo-européennes	bai	austro-bavaroises

Tableau 2 : Familles, branches ou groupes linguistiques dessinés sur la carte de la Figure 2, avec leur abréviation (code sur 3 lettres).
Dans « autres langues indo-européennes » entrent le rromani et l'arménien occidental. À droite sont consignées les variétés germaniques

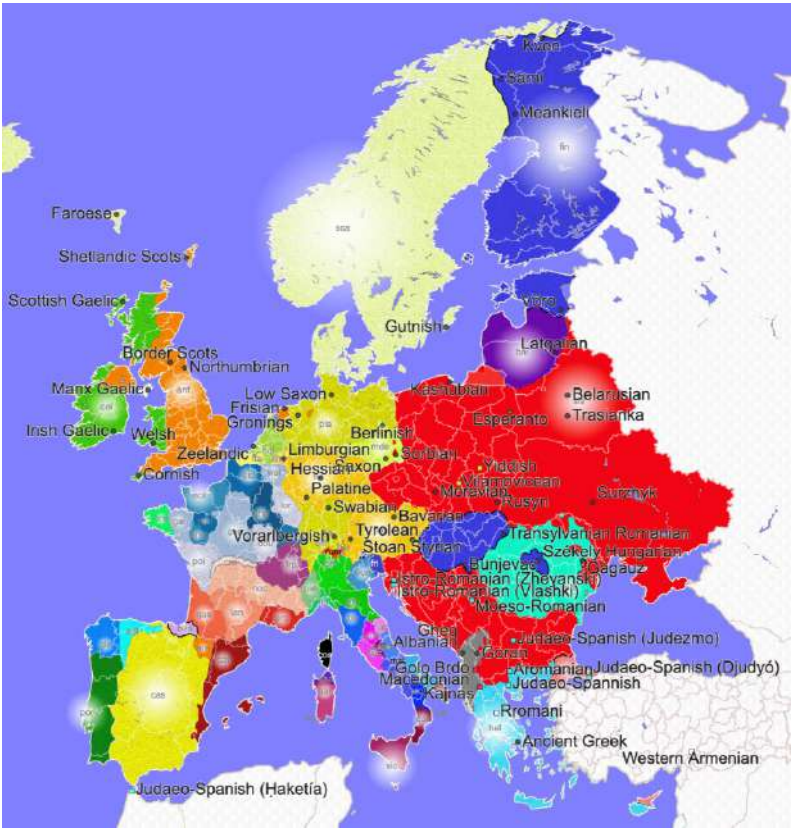


Figure 2 : Carte des langues minoritaires d'Europe (sans l'option qui fait apparaître les cadres autour de pays de l'Ouest).

6. Analyse linguistique de quelques langues des Balkans

Il serait intéressant d'analyser les enregistrements collectés dans ces langues, du point de vue de la phonétique/phonologie, de la morpho-syntaxe et du lexique, comme nous l'avons fait pour le provençal et le languedocien oriental, et de même que nous avons rapproché les discours bunjevac et ruthène de ceux de certains Provençaux. La pandémie, cependant, ne nous en a pas laissé la

possibilité[4]. Nous nous contenterons de quelques notes concernant le macédonien de Golo Brdo (région de l'est de l'Albanie majoritairement peuplée de slaves musulmans), le goran et le roumain de Transylvanie, parmi les derniers enregistrements recueillis dans les Balkans. Soulignons d'emblée que ces trois variétés ont développé une marque morphologique de la définitude (détermination) que par commodité nous appellerons « article (postposé) », même si son emploi se raréfie en goran et n'a pas été jugé nécessaire par le locuteur enregistré.

6.1. Macédonien de Golo Brdo

En macédonien de Golo Brdo, il y a chute de la voyelle épenthétique qui est ici [a], contre [e]/[a] en macédonien, devant l'article postposé (ici *ветар* 'vetar' → *ветрот* 'vetrot', « vent » → « le vent »). Le vieux-slave Ѳ a donné [o] et non [a] comme en macéédonien, [u] en goran et dans d'autres langues slaves (ex. *потот* 'potot' « le chemin », *потник* 'potnik' « le voyageur »). Concernant l'usage des temps, la traduction utilise de façon expressive l'aoriste (*видоа* 'vidoa' « ils ont vu »), l'imparfait (ex. *вееше* 'veeše' « il soufflait »). De plus, on a à plusieurs reprises la forme en *л* 'l' du passé non-testimonial (Friedman, 2001), qui serait gauche en macédonien standard (ex. *се расправале* 'se raspravale' « ils auraient été en train de se disputer »), ainsi que la terminaison en *-ит* '-it' des verbes à la 3e personne du singulier du présent, complétant des verbes semi-auxiliaires (ex. *сакаше да покажит* 'sakaše da pokažit' « voulait qu'il montre » = « voulait montrer »).

Au niveau syntaxique, la phrase *сега човекот го соблечит палдесӱто* 'sega čovekot go sobl̀ečit' peut sembler erronée, signifiant littéralement « maintenant l'homme enlève son manteau », au lieu de « si bien que… » : c'est qu'en golobrdski comme dans quasiment tous les dialectes macédoniens de l'ouest, le mot *сега* 'sega' est employé indifféremment dans le sens de « si bien que » ou « alors ». Au niveau lexical, on note enfin l'emprunt au français *палдесӱ* 'paldesü', avec le tréma d'origine turco-albanaise, tandis que, pour traduire « souffler », l'usage de *веит* 'veit' est à comparer au macédonien *вее* 'vee', un peu marqué

[4] *C*f. note 1.

stylistiquement et signifiant plutôt aujourd'hui « enneiger » — on utiliserait davantage *дува* 'duba', *дувне* 'dubne' en macédonien standard (Kostov, communication personnelle).

Il y a donc bel et bien des nuances, notamment cet archaïsme de la terminaison verbale. On peut comparer le début de la fable avec d'autres langues slaves des Balkans dans le Tableau 3.

Cat	Langue	"Le vent du Nord et le soleil se disputaient (un jour)"
Slaves	bunjevac	Siverni vitar i sunce su se svađali
	gGoran	Северни-ветер и сәнце се расправале 'Severni-veter i sənce se raspravale'
	Kajnas	Severo i sallcjëto fatije en den
	macédonien de Golo Brdo	*Северниот ветар и сонцето се расправале еден ден* 'Severniot vetar i sonceto se raspravale eden den'
Romanes	roumain de Transylvanie	S'ău încăĭeratŭ vîntułŭ de mĭéḑă-nópte şi cu sórile
	moéso-roumain	Северни vânt şï soareli ântr-o zï s-a certat
	aroumain	Aratsili shi Soarli s-ancãcea

Tableau 3 : début de la fable dans quelques langues slaves et romanes des Balkans, en alphabet cyrillique (en italique) et en alphabet roman (entre guillemets simples, comme dans le texte, quand il s'agit de notre translitération)[5].

6.2. Goran

En goran, on remarque *note* des sonantes légèrement palatalisées, notées ici avec l'apostrophe (ex. *се расправал'е* 'se raspraval'e' « se disputaient » ; *н'его* 'n'ego' « de lui », moins palatalisé que le serbe *њега* 'njega'). Comme en serbe, en revanche, les anciennes occlusives <т'> et <д'> se sont affriquées en <ћ> /t͡ɕ/ et <ђ> /d͡ʑ/ respectivement (ex. *ће* 'će' « veut »), l'opposition avec les affriquées non-palatalisées étant maintenue (Mladenović, 2001). Comme en macédonien, en revanche, le schwa a un statut de phonème (ici transcrit par le graphème <ə>, par exemple dans *сәнце* 'sənce' « soleil »). Comme dans les autres parlers de la

[5] Ailleurs et de façon conventionnelle, les graphèmes sont notés entre chevrons, tandis que les crochets (resp. les barres obliques) indiquent les transcriptions phonétiques (resp. phonologiques), suivant l'alphabet de l'API.

montagne Šar (entre Kosovo, Macédoine et Albanie), le *-л* '-l' en fin de syllabe a donné *–в* '-v', un /v/ qui s'assourdit en [f] en fin de mot, notamment dans la désinence du parfait (ex. *наишоф* 'naišof' « est venu »). Comme dans les parlers macédoniens occidentaux, l'accent se trouve sur l'antépénultième,, et on remarque la présence d'un accent syntagmatique sur le /i/ de *северни-ветер* 'severni-veter' « vent du nord » (d'où la transcription avec un trait d'union). Dans ce dernier mot, à comparer au serbe *ветар* 'vetar', la vocalisation issue du <ъ> '*jer* fort' protoslave est également de type macédonien.

Au niveau morphologique, le parfait est utilisé là où autrefois on employait l'aoriste ou l'imparfait. Et il n'existe au passé qu'une seule terminaison pour les trois genres , au pluriel : *-л'е* '-l'e' (ex. *се дoговорил'е* 'se dogovoril'e' « sont convenus »), ce qui constitue une caractéristique des parlers macédoniens — que l'on retrouve dans la variété de Golo Brdo. Sous l'influence des parlers macédoniens, également, le génitif masculin singulier des mots à flexion adjectivale conserve l'ancienne forme *-го* 'go' (notamment dans le pronom de 3e personne masculin). De plus, la comparaison des adjectifs et des adverbes se fait de manière analytique (ex. *појак* 'pojak', *појако* 'pojako', « plus fort »). Au niveau lexical, enfin, on remarque pour traduire « manteau » le mot *кожуф* 'kožuf', rare cas où le /x/ du proto-slave a été remplacé par /f/.

6.3. Roumain de Transylvanie

En Transylvanie, la prononciation est très différente de celle du roumain standard parlé à Bucarest : palatalisations en abondance ; consonnes supplémentaires (/c/, /ɟ/, /ɲ/, /x/), voyelles supplémentaires (/ɔ/, /ɛ/, /y/, /œ/), absence de la consonne /dʒ/ (remplacée par /ʒ/) et absence de la diphtongue /wa/ (Staelens, 2019). Alors que le roumain standard possède la diphtongue /e̯a/, <é> note ici la palatalisation de la consonne qui précède, suivie de la voyelle /a/; si la consonne ne peut être palatalisée, la semi-voyelle /j/ prend place. La diphtongue /əw/ devient souvent [oː], par exemple dans *ău* (« ont », dans la construction « ils ont vu »), et le digramme <ea> se prononce [ɛː], par exemple dans *aşea* (« donc », là où en roumain standard on a une diphtongue). Au niveau morpho-phonologique, l'article défini masculin, postposé,

obéit à l'harmonie vocalique : *le* après *e*, *łŭ* après *u* (ex. *vîntułŭ* « le vent »). Dans ce dernier cas, il fait au besoin réapparaître en surface le *ŭ* caduc du substantif (en fin de mot, *-tŭ* pouvant être voisé ou ne pas être prononcé, en fonction du mot qui suit), mais en aucun cas le *l* ne saurait être prononcé. Dans la pratique, un substantif finissant en /u/ se prononce de la même manière avec ou sans article, tndis qu'un mot en *ŭ* muet aura ce *u* prononcé. Cet article défini vient du latin ILLE, de sorte qu'on a **SOLE ILLE > **soreĭle > sór**i**le* (« le soleil », *soar**e**le* en roumain standard, où la voyelle semble s'être perdue plus tôt).

Au niveau de la morphologie verbale, « il est » se traduit *ĕĭ/îi/'ĭ* [ɯj] — la forme *este,* que l'on retrouve en roumain standard, est exclusivement prédicative dans la variété de Transylvanie, avec l'acception « il y a » ; elle n'est jamais copulative. Quant au subjonctif, il doit toujours comporter plusieurs syllabes à la 3e personne : d'où *deŭe* (« donne », face à *dea* en roumain standard). Au niveau morpho-syntaxique, sinon, l'infinitif est utilisé plus volontiers que le subjonctif du roumain standard (ex. *a luci şi străluci* « à briller et briller »). Au niveau véritablement syntaxique, en outre, le prédicat est en règle générale placé avant le sujet : ex. *s'ău încăĭeratŭ vîntułŭ…* « se disputaient le vent… », alors qu'ailleurs en roumain, l'ordre des mots peut être plus flexible (Nicolae, 2019). Au niveau lexical, enfin, on note :

- *mĭéḑă-nópte* (littéralement « minuit ») pour « nord » ;
- à plusieurs reprises l'interjection *nó* ([nɔ], commune au hongrois et aux langues slaves septentrionales, mais intraduisible en français) ;
- *preumbla* (« marchait », à comparer au roumain standard *plimba* < PERAMBULA(BA)T) ;
- *gubă, şubă.* (« limousine », « manteau », probablement de la même origine arabe que le français « jupe ») ;
- *acmu* (« à présent», à comparer au roumain standard *acum* < *ECCUM (HUC) MODO) ;
- *tătŭ* (< TANTU-, « tant »), là où en roumain standard on aurait plutôt *tot* (< TOTU-, « tout ») ;
- *s'ău urîtŭ* « s'est lassée » et *ţîpatŭ* « jeté » (apparenté au français « tapé »), alors que *s-a urît* signifierait plutôt « s'est détesté » et *tîpat* « crié » en roumain standard ;

- *atuncine* « alors, à ce moment-là », différencié en roumain de Transylvanie de *atuncĭ* « alors, donc » ;
- *s'ău hăţîtŭ* « a commencé », avec des nuances à côté du roumain standard *a începe*.

7. Conclusion

À l'heure de la mondialisation qui, au lieu d'élargir nos champs de visions du monde, a plutôt tendance à imposer une façon particulière de penser (*cf.* exergue), nous nous sommes faits dans cet article les avocats de la pluralité des langues. À l'heure des *big data*, le corpus que nous avons présenté fait figure de miniature, par sa petite taille, mais aussi dans le sens étymologique de décoration colorée. Même une minute de parole, nous l'avons vu, est cependant riche d'enseignement quant à la variation, sur une base comparable. L'objectif était d'illustrer la diversité linguistique de l'Europe, pour la promouvoir à travers une carte parlante et préserver un témoignage sonore de certaines langues menacées d'extinction (Hagège, 2000).

En nous focalisant sur quelques langues et les commentaires épilinguistiques qu'elles suscitent, nous avons montré combien importante est la question du passage à l'écrit, non seulement pour la documentation mais encore pour l'enseignement et la survie des langues minoritaires — le numérique étant également un des enjeux de leur revitalisation (Soria *et al.*, 2013). Le système orthographique d'une langue, outil supradialectal qui affirme l'unité de cette langue et en vient à s'identifier à elle, dessine en même temps les contours de ce qui sépare celle-ci d'autres langues — on pense par exemple, dans les Balkans, au choix de l'alphabet latin ou cyrillique. Le discours sur le sentiment identitaire, dont nous avons rapporté quelques échantillons, est donc fondamental et une analyse plus poussée demanderait d'autres enquêtes de terrain.

Nous espérons reprendre celles-ci et enregistrer plusieurs versions d'une même langue, comme cela a été fait pour le Schleswig-Holstein, dont récemment une carte a été intégrée à notre site <https://atlas.limsi.fr/?tab=sh, avec plus de 80 enregistrements transcrits en collaboration avec l'Université de

Kiel. Il s'agit là sans doute de dialectes plus clairement que dans le cas du moéso-roumain, même si cette dernière langue n'est pas parlée sur un territoire compact. Cependant, des activistes du Plattdüütsch et des nationalistes roumains ne tiendraient pas ce discours : les réactions suscitées par le travail de cartographie linguistique présenté ici et en particulier notre analyse des variétés — somme toute assez proches — de provençal et de languedocien oriental nous obligent à une certaine prudence. Même les résultats de la dialectométrie — également appliqués à des traductions de la fable « La bise et le soleil » (Boula de Mareüil *et al.*, 2021) — doivent être pris avec précaution.

La question de ce qui différencie un dialecte d'une langue est une vieille question et en grande partie une mauvaise question : tout dépend de l'angle d'observation. Un dialecte est souvent interprété en négatif par le sens commun comme non-langue, non-moderne, non-écrit, n'obéissant à aucune grammaire. En *folk linguistics* (Preston, 2005), il s'agirait d'un parler oral, régional, trop peu différencié par rapport à une langue de plus vaste champ, et dénué de règles. Si l'écrit et l'écrivain confèrent au dialecte le statut de langue, on comprend les crispations autour de la graphie — ceci, dès les débuts des Félibres en Provence (Costa, 2012). En termes purement linguistiques, cependant, il est impossible de trancher entre langue et dialecte : la distinction est d'un autre ordre, sociolinguistique, politique, culturel, historique. On connaît la boutade attribuée au maréchal Lyautey et au sociolinguiste Weinreich : une langue est un dialecte qui a une armée et une marine. En réaction à une longue période de mépris pour les dialectes et du fait que l'Union européenne ne reconnaît que les langues minoritaires ou régionales moins utilisées, certains affirment que le provençal est une véritable langue (Blanchet, 2004) et non un dialecte de l'occitan. Le débat n'est pas clos, l'Occitanie ne disposant ni d'une armée ni d'une marine propres. Dans les Balkans, où les armes ont parlé il n'y a pas si longtemps, comme en Ukraine, où la guerre continue à tuer, émettons simplement le vœu pieux que perdure la richesse linguistique illustrée par cet atlas sonore.

8. Remerciements

Nous tenons à exprimer notre chaleureuse gratitude envers les locuteurs qui nous ont accordé de leur temps pour traduire dans leur langue la fable « La bise et le soleil » : sans eux, ce travail aurait été impossible et n'aurait eu aucun sens. Nous sommes également redevables à Mathieu Castel, à Alain Barthélémy-Vigouroux et à Jean Léo Léonard pour leurs commentaires. Nous remercions enfin vivement Radivoje Mladenovićc et Georges Staelens pour leurs analyses linguistiques. Les erreurs d'interprétation restent les nôtres.

BIBLIOGRAPHIE

ALAIN (1925) *Éléments d'une doctrine radicale.* Paris : Gallimard.

ALMBERG Jøn & Kristian SKARBØ (2002) « Nordavinden og sola. Ein norsk dialect database pånettet http://www.ling.hf.ntnu.no/nos ». In MOEN Inger, Hanne Gram SIMONSEN, Arne TORP & Kjell Ivar VANNEBO (eds.), *Utvalgte artikler fra Det niende møtet om norsk språk.* Oslo : Novus Forlag.

ALLEN W. Sidney (1987) *Vox Graeca: The Pronunciation of Classical Greek.* Cambridge : Cambridge University Press.

ANDERSON Alexander & Tymotesz KRÓL, (2006) *A grammar of Wymysorys.* Durham : SEERLC,.

BARTHELEMY-VIGOUROUX Alain & Guy MARTIN (2017) *Manuel pratique de provençal contemporain.* Aix-en-Provence : Edisud.

BLANCHET Philippe (2004) « L'identification sociolinguistique des langues et des variétés linguistiques : pour une analyse complexe du processus de catégorisation fonctionnelle », *Workshop MIDL*, Paris. 31–36.

BOCHMANN Klaus (2022) « Langues minoritaires et conflits linguistiques ». In NOIRARD Stéphanie (Dir.), *Transmettre les langues minorisées. Entre promotion et relégation.* Rennes : Presses Universitaires de Rennes. 25–36.

BOISGONTIER Jacques (1981–1986) *Atlas linguistique et ethnographique du Languedoc oriental.* Paris : Éditions du CNRS.

BOULA DE MAREÜIL Philippe, Frédéric VERNIER & Albert RILLIARD (2017) « Enregistrements et transcriptions pour un atlas sonore des langues régionales de France ». *Géolinguistique* 17. 23–48.

BOULA DE MAREÜIL Philippe, Valentina DE IACOVO, Antonio ROMANO & Frédéric VERNIER (2019a) « Un atlante sonoro delle lingue di Francia e d'Italia: focus sulle parlate liguri ». In TOSO Fiorenzo (a cura di), *Il patrimoniolinguistico storico della Liguria. Attualità e future*. Savona : Insedicesimo. 33–46.

BOULA DE MAREÜIL Philippe., Gilles ADDA, Lori LAMEL, Albert RILLIARD & Frédéric VERNIER (2019b) « A speaking atlas of minority languages of France: collection and analyses of dialectal data ». *19th International Congress of Phonetic Sciences*, Melbourne. 1709–1713.

BOULA DE MAREÜIL Philippe, Lucien MAHIN & Frédéric VERNIER (2020) « Les parlers romans dans l'atlas sonore des langues et dialectes de Belgique ». *Bien Dire et Bien Aprandre* 35. 85–108.

BOULA DE MAREÜIL Philippe, Gilles ADDA & Lori LAMEL (2021) « Comparaison dialectométriques de parlers du Croissant avec d'autres parlers d'oc et d'oïl ». In ESHER Louise, Maximilien GUERIN, Nicolas QUINT & Michela RUSSO (éds.), *Le Croissant linguistique entre oc, oil et francoprovencal : des mots a la grammaire, des parlers aux aires*. Paris : L'Harmattan. 159–17Z.

BOUVIER Jean-Claude & Claude MARTEL (1975–1986) *Atlas linguistique et ethnographique de la Provence*. Paris : Éditions du CNRS.

BRETON Roland (1974) *Géographie des langues*. Paris : Presses Universitaires de France.

BUCHOLTZ Mary (2003) « Sociolinguistic nostalgia and authentication of identity ». *Journal of Sociolinguistics* 7(3). 398–416.

CANUT Cécile (1998) « Pour une analyse des productions épilinguistiques ». *Cahiers de praxématique* 31. 69–90.

CANUT Cécile. (2011) « La langue romani : une fiction historique ». *Langage et société* 136. 55–80.

CAUBET Dominique, Salem CHAKER & Jean SIBILLE (2001) *Codification des langues de France*. Paris : L'Harmattan.

CERQUIGLINI Bernard (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>.

CHAMBERS Jack & Peter TRUDGILL (2004) *Dialectology*. Cambridge : Cambridge University Press.

COUROUAU Jean-François (2005) « L'invention du patois ou la progressive

émergence d'un marqueur sociolinguistique français XIII^e–XVII^e siècles ». *Revue de linguistique romane* 69. 273–274.

COSTA James (2010) *Revitalisation linguistique : discours, mythes et idéologies : approche critique de mouvements de revitalisation en Provence et en Ecosse*. Thèse de doctorat. Grenoble : Université Stendhal.

COSTA James (2012) « De l'hygiène verbale dans le sud de la France ou Occitanie ». *Lengas* 72. 83–112.

COURTHIADE Marcel (1988) « Les derniers vestiges du parler slave de Boboščica et de Drenovthne (Albanie) ». *Revue des Études Slaves* 60(1). 139–157.

COURTHIADE Marcel (2000) « Les Rroms, Ashkalis et Gorans de Dardanie (Kosovo) ». *Annales de l'autre Islam* 7. 255–280.

COURTHIADE Marcel (2004) *Les Rroms dans le contexte des peuples européens sans territoire compact*. Paris : INALCO.

COURTHIADE Marcel (2007) « Jeu dialectes-langues ». *Langues et cité* 9. 6–7.

COURTHIADE Marcel (2013) *A succint history of the Rromani language*. Paris : INALCO.

COURTHIADE Marcel (2020) *De la « tsiganologie » à la « rromologie » : études en langue, littérature, culture et société du peuple rrom en France et dans le monde*. Thèse d'habilitation à diriger des recherches. Paris : INALCO.

COURTHIADE Marcel & Stella KARAMAGKIOLA (2013) « Attitudes comparées de deux minorités européennes sans territoire compact vis-à-vis de la langue maternelle : les Rroms et les Aroumains ». In ALEN GARABATO Carmen (Dir.), *Gestion des minorités linguistiques dans l'Europe du XXI^e siècle*. Limoges : Éditions Lambert-Lucas. 193–215.

CUNIA Tiberius (1999) « On the Standardization of the Aromunian System of Writing », *The Bituli-Macedonia Symposium*. Syracuse. 1–16.

DALBERA Jean-Philippe (1994) *Les parlers des Alpes-Maritimes : étude comparative, essai de reconstruction*. Londres : AIEO.

DAUZAT Albert (1927) *Les patois. Évolution, classification, étude*. Paris : Librairie Delagrave.

DJORDJEVIC Ksenija (2013) « Bunjevci : une identité collatérale discutée dans le triangle de Baja ». *Carnets d'Atelier de Sociolinguistique* 7. 117–134.

DJORDJEVIC-LEONARD Ksenija (2014) « Le ruthène vs l'ukrainien : les paradoxes d'une individuation ». In NADAL FARRERAS Josep Maria, Anne-Marie CHABROLLE-CERRETINI & Olga FULLANA NOELL (éds.), *L'espace des langues*. Paris : L'Harmattan. 275–288.

EYSSERIC Violaine (2005) *Le corpus juridique des langues de France*, Rapport de la Délégation Générale à la langue française et aux langues de France.

FANON Frantz (1952) *Peau noire, masques blancs*. Paris : Éditions du Seuil.

FERRY Jules (1885) Débat à la Chambre des députés, Paris.

FRIEDMAN Victor (2001) *Macedonian*. Durham : SEERLC.

GARDE Paul (2004) *Le discours balkanique. Des mots et des hommes* Paris : Fayard.

GELU Victor (1856) *Chansons provençales*. Marseille : Camoin.

GILLIERON Jules & Edmond EDMONT (1902–1910) *Atlas linguistique de la France*. Paris : Champion.

HAGEGE Claude (2000) *Halte à la mort des langues*. Paris : Odile Jacob.

JOUVEAU René (1980) « Les hésitations orthographiques de Frédéric Mistral ». *5e Colloque de langues dialectales*, Monaco. 55–67.

KERÄNEN Mari (2018) « Language maintenance through corpus planning – the case of Kven ». *Acta Borelia* 35(2). 1–16.

LABOV William (1976) *Sociolinguistique*. Paris : Éditions de Minuit.

LIEUTARD Hervé (2019) « Les systèmes graphiques de l'occitan : un kaléidoscope des représentations et des changements linguistiques ». *Lengas* 86, en ligne.

LEONARD Jean Léo (2013) « Mulgi, kihnu, võro, seto : langues collatérales d'Estonie et pluralisme de proximité ». In ALEN GARABATO Carmen (Dir.), *Gestion des minorités linguistiques dans l'Europe du XXIe siècle*. Limoges : Éditions Lambert-Lucas. 35–48.

MARTEL Philippe (2019) « Langues de France et textes officiels : un peu d'histoire ». *Journée d'étude & débat sur les langues régionales : situation et perspectives*, Paris. 99–110.

MARTEL Philippe & Marie-Jeanne VERNY (2020). « Les langues régionales au Parlement, ou l'éternel retour ». *Glottopol* 34. 69–90.

MLADENOVIĆ Radivoje (2001) *Govor Šarplaninske Župe Gora*. Belgrade : SANU.

MOSELEY Christopher (2010) *Atlas of the World's Languages in Danger*. Paris : UNESCO.

NICOLAE Alexandru (2019) *Word order and parameter change in Romanian: A comparative Romance perspective*. Oxford : Oxford University Press.

OLIVIERI Michèle, Sylvain,CASAGRANDE., Guylaine BRUN-TRIGAUD & Pierre-Aurélien GEORGES (2017) « Le *Thesaurus Occitan* dans tous ses états ». *Revue française de linguistique appliquée* 22. 89–102.

PICARD Flore (2000) « À l'intersection entre complexité flexionnelle et complexité diasystémique : modélisation du verbe same ». *Verbum* 42(1–2). 63–84.

PRESTON Denis R. (2005) « What is folk linguistics? What should you care? ». *Lingua Posnaniensis* 47. 143–162.

RIDANPÄÄ Juha (2018) « Why save a minority language? Meänkieli and rationales of language revitalization ». *Fennia – International Journal of Geography* 196(2). 187–203.

ROMANO Antonio (2016) « La *BD AMPER*, *La tramontana e il sole* e altri dati su lingue, dialetti, socioletti, etnoletti e interletti del Laboratorio di Fonetica Sperimentale "Arturo Genre" ». *Quaderni del Museo delle Genti d'Abruzzo* 41. 225–240.

SAÏDI Victoriya (2009) « Le problème de la nomination de la langue ukrainienne ». *Études de lettres* 4. 101–114. Ó

SAINTE-BEUVE Charles Augustin (1851) *Causeries du lundi*. Paris : Garnier.

´SALLABANK Julia & Yan MARQUIS. (2018) « 'We Don't Say It Like That': Language Ownership and (De)Legitimising the New Speaker ». In SMITH-CHRISTMAS Cassandra, Noël Ó MURCHADHA, Michael HORNSBY & Mairead MORIARTY (eds.), *New Speakers of Minority Languages*. Londres : Palgrave Macmillan. 67-90

ŠIMÁČKOVÁ Šárka, Václav Jonáš PODLIPSKÝ & Kateřina CHLÁDKOVÁ (2012) « Czech spoken in Bohemia and Moravia ». *Journal of the International Phonetic Association* 42(2). 225–232.

STAELENS Georges (2019) *Étymologie roumaine revisitée. De la Rîmnŭ ne tragemŭ*. Wrocław : Amazon Fulfilment.

SORIA Claudia, Joseph MARIANI & Carlo ZOLI (2013) « Dwarfs sitting on the giants' shoulders – how LTs for regional and minority languages can benefit from piggybacking major languages ». *XVII Foundation for Endangered Languages Conference*, Ottawa. 73–79.

SUMIEN Doomergue (2007) « Preconizacions del Conselh de la Lenga Occitana ». *Lingüistica Occitana* 6. 1–157.

VIAUT Alain (2020) « De la relation entre variantes et standard dans les procédures de revitalisation des langues minoritaires ». *Les Cahiers du GEPE* 12, en ligne.
VIAUT Alain & Antoine PASCAUD (2017) « Pour une définition de la notion de "langue régionale" ». *Lengas* 82, en ligne.
WALTER Henriette (1988). *Le français dans tous les sens*. Paris : Robert Laffont.

Philippe Boula de Mareüil,
Marcel Courthiade, Frédéric Vernier
Université Paris-Saclay & CNRS, LISN
Philippe.boula.de.mareuil@limsi.fr

Table des matières

INTRODUCTION……1
Evolution du statut des langues régionales et de leur environnement numérique
Annie Rialland & Michela Russo

Section 1
Outils numériques, linguistique et revitalisation des langues (breton et picard)

CHAPITRE 1……37
Outils numériques et traitement automatique du breton
Mélanie Jouitteau & Reun Bideault

CHAPITRE 2……75
Peut-on revitaliser la langue picarde grâce aux nouvelles technologies
Christophe Rey

Section 2
Bases de données textuelles, lexique et syntaxe (occitan)

CHAPITRE 3……99
Nouvelles approches linguistiques et lexicographiques de l'occitan médiéval
Hervé Lieutard

CHAPITRE 4……121
Nouvelles perspectives pour la linguistique occitane à partir de la base textuelle BaTelÒc
Myriam Bras

Section 3

Bases de données orales, phonétique et diachronie (aires transitionnelles entre occitan et francoprovençal)

CHAPITRE 5……145
On vowel nasalization in transitional francoprovençal *and* occitan areas
Michela Russo & Jonathan Kasstan

Section 4

Outils numériques, bases de données orales, implication des locuteurs

CHAPITRE 6……213
Les parlers du Croissant : un aperçu des actions actuelles de documentation et de promotion d'un patrimoine linguistique menacé
Nicolas Quint

CHAPITRE 7……247
De la Provence aux Balkans : discours épilinguistiques autour d'un atlas sonore des langues régionales ou minoritaires d'Europe
Philippe Boula de Mareüil, Marcel Courthiade & Frédéric Vernier

Table des matières……285

Les études sur les langues régionales de France se situent dans un contexte qui évolue tant du point de vue du statut et de la vie de ces langues que de leur environnement numérique. Le but premier de cet ouvrage est de montrer comment ces études ont suivi cette évolution et ont pu en bénéficier et y contribuer. L'ouvrage offre aussi un regard sur la position de la France en matière de droits linguistiques des langues régionales et de leur enseignement, en particulier avec la 'loi relative à la protection patrimoniale des langues régionales et à leur promotion' dite loi Molac promulguée le 21 mai 2021.

Les travaux présentés ont tous eu recours à des ressources et outils numériques qui leur ont permis d'accroître le volume des données pouvant être prises en compte. Certains d'entre eux portent sur la création même de ces données et outils (corpus, traducteur automatique, analyseur syntaxique, atlas…) qui contribuent à l'équipement numérique de ces langues, lesquelles restent cependant sous-dotées, quoiqu'à des degrés divers. L'augmentation des données numériques et des possibilités de les interroger permet de renouveler des analyses linguistiques qu'elles soient phonétiques, lexicales, syntaxiques ou dialectales. En retour, les travaux des chercheurs influent sur la vie des langues régionales, qui sont souvent des langues en danger, en les documentant, en leur procurant des ressources qui leur permettront d'exister numériquement, en changeant le regard que les locuteurs et le public peuvent avoir sur elles. Les langues régionales considérées dans cet ouvrage sont principalement le breton, le picard, les parlers du Croissant, l'occitan, le francoprovençal et, dans une moindre mesure, le basque, l'alsacien ainsi que des 'langues sans territoire compact', telles que le rromani et des langues minoritaires d'Europe, en particulier des Balkans.

978-2-9570894-3-7