COLaF : Corpus et Outils pour les Langues de France et variétés de français

Benoît Sagot[1, *] Slim Ouni[2, *]
Sam Bigeard[2] Lucence Ing[1] Rasul Dent[1] Juliette Janès[1] Thibault Clérice[1]
Rachel Bawden[1] Emmanuel Vincent[2] Oriane Nédey[1] Malek Yaich[2]
Panagiotis Tsolakis[2] Vincent Colotte[2] Mostafa Sadeghi[2]
(1) Inria, Paris, France
(2) Université de Lorraine, CNRS, Inria, LORIA, France
prenom.nom@inria.fr

RÉSUMÉ

Nous présentons COLaF, un projet dédié au développement d'outils et de ressources de traitement automatique des langues (TAL) pour le français et les autres langues de France, avec une attention particulière sur les variétés moins dotées. Le projet concerne les données textuelles, audio et vidéo, et vise à fournir des corpus et des outils pour le langage écrit, parlé et signé. Il inclut la collecte, la normalisation et la documentation de données préexistantes, y compris des données actuellement non accessibles ou non exploitables à des fins de recherche, ainsi que le développement d'outils de TAL adaptés à ces langues, comme des outils pour l'annotation linguistique et pour la traduction automatique. Cet article présente les principaux défis posés par le projet et des premiers résultats.

ABSTRACT

COLaF: Corpus and Tools for Languages of France and Varieties of French

We introduce COLaF, a project dedicated to the collection and development of natural language processing (NLP) tools and resources for languages of France and varieties of French, with a particular focus on less-resourced languages and varieties. The project covers text, audio and video processing in order to provide corpora and tools for written, spoken and signed language. The project involves the collection, standardisation and documentation of pre-existing data, including data that is not currently accessible or exploitable for research purposes, and the development of NLP tools adapted to these languages, i.e. tools for linguistic annotations and machine translation. In this article, we describe the main challenges set out for the project and preliminary contributions.

MOTS-CLÉS : Français, Langues régionales, Langues d'outre-mer, Langues non territoriales, Langue des signes française, Langues peu dotées, Corpus, Normalisation, Open source.

KEYWORDS: French, Regional languages, Overseas languages, Non-territorial languages, French sign language, Low-resource languages, Corpora, Standardisation, Open-source.

∗. Co-porteurs du projet.

1 Introduction

Les domaines du traitement automatique des langues (TAL) et de la parole ont aujourd'hui un très fort impact économique et social. Des outils tels que les assistants vocaux et la traduction automatique font désormais partie du quotidien. Cependant, les produits des grands acteurs privés ne répondent pas nécessairement aux attentes et aux besoins de tous les locuteurs du français, en France ou ailleurs, ni à ceux des autres langues de France, dans toute leur diversité (Leixa *et al.*, 2014).

En effet, ces technologies voient leur performance baisser fortement dès lors que l'utilisateur s'éloigne d'un idéal standard, par son âge, accent ou variété de français liée à une région d'origine, à un niveau de connaissance du français, à une pathologie, etc. Quant aux autres « langues de France », presque toutes sont délaissées par les grands acteurs privés, laissant leurs locuteurs à l'écart de ces technologies.

Ces langues, moins standardisées (graphies non stabilisées, variantes locales, manque de langage de spécialité, etc.), présentent peu de corpus et autres ressources prêtes à l'emploi. Le développement de technologies de TAL pour ces langues présente de forts enjeux liés à la francophonie, à l'inclusion et à la promotion du patrimoine linguistique de la France dans toute sa diversité.

Le projet COLaF [1], financé par le Plan National de Recherche en Intelligence Artificielle, a pour objectif de développer et mettre à disposition des technologies numériques linguistiques ouvertes pour la francophonie et les langues de France, en contribuant à la création de corpus de données inclusifs, de modèles et de briques logicielles en partenariat avec les chercheurs, les associations de promotion des langues de France et autres fournisseurs de données. Il œuvre dans deux directions complémentaires :

— la création de jeux de données ouverts, notamment des corpus textuels non annotés, des lexiques morphologiques, des corpus audio et des corpus de LSF,
— l'utilisation de ces jeux de données pour développer des modèles et des outils de TAL adaptés aux langues et locuteurs de France.

Après avoir présenté les langues d'études et l'état de l'art, nous exposerons les volets du projet liés aux données et aux outils, en nous attachant à décrire, pour chacun d'eux, une étude déjà réalisée.

2 Les langues d'étude

COLaF se donne pour champ d'action l'ensemble des langues parlées en France, dans toute leur diversité. Plus précisément, nous prévoyons de travailler sur les langues et variétés suivantes [2] :

— français, dans toute sa diversité géographique, thématique (français scientifique, français juridique), diachronique, mais aussi diastratique et diaphasique (par exemple, français des réseaux sociaux, qui connaît en plus parfois une alternance codique avec d'autres langues), et dans ses différents niveaux de maîtrise (y compris les apprenants L2), etc. ;
— autres langues métropolitaines, que nous appellerons de façon impropre « langues régionales » : alsacien, breton, occitan, picard, corse, basque...
— langues créoles à base française, sans restriction géographique (créoles guadeloupéen, martiniquais, guyanais, réunionnais, louisianais, haïtien, mauricien, seychellois) ;

1. Des informations sur le projet peuvent être trouvées sur son site web : `https://colaf.huma-num.fr/`.
2. Pour une discussion plus approfondie des langues de France autres que le français, on pourra par exemple se reporter à (Cerquiglini, 1999; Collectif & Cerquiglini, 2003).

— langues locales parlées dans les territoires français extra-européens (langues polynésiennes, certaines langues mélanésiennes, langues amérindiennes de Guyane, langues kanak, shimaoré et shibishi à Mayotte) ;
— langues d'immigration parlées sur le territoire français (variétés d'arabe dialectal, arménien occidental, variétés de berbère, judéo-espagnol, romani, yiddish...) ;
— langue des signes française (LSF).

Nous nous concentrons sur certaines de ces variétés de français et certaines de ces autres langues seulement, en fonction d'un certain nombre de facteurs : le nombre de locuteurs, le manque de ressources aisément disponibles et utilisables, l'existence de ressources à mettre en valeur, l'intérêt des partenaires spécialisés dans une langue, et, de façon opportuniste, les compétences et appétences des personnes qui participent au projet. Nous souhaitons, à la fin du projet, avoir fait avancer significativement la situation d'au moins une langue relevant de chacune des catégories définies ci-dessus, en faisant varier également les types de données (documents non contemporains, corpus de spécialité, textes extraits des réseaux sociaux, données de paroles de média radio-télévisuels, etc.).

3 Travaux connexes

Une quantité considérable de travaux de recherche ont été consacrés au traitement automatique des langues (TAL) pour le français. La situation pour les autres langues de France, en termes de disponibilité de corpus et d'outils, est bien moins favorable, certaines langues n'étant pas couvertes du tout (Mariani *et al.*, 2012; Soria *et al.*, 2013; Adda *et al.*, 2023). Nous avons pour objectif ici de donner un aperçu des travaux antérieurs en matière de développement d'outils et de ressources pour les langues d'intérêt du projet COLaF. Nous nous concentrons notamment sur la littérature sur les langues minoritaires du projet.

COLaF n'est pas le premier projet destiné à la collecte de ressources pour les langues minoritaires de France. On peut citer par exemple, le projet ANR RESTAURE [•] (RESsources informatisées et Traitement AUtomatique pour les langues REgionales) (Bernhard & Vergez-Couret, 2015; Bernhard, 2019), qui s'est concentré sur le développement de ressources et outils pour l'alsacien, le picard et l'occitan, et le projet DIVITAL, qui a pour objectif d'améliorer la visibilité numérique des langues de France pour l'alsacien, le corse, l'occitan et le poitevin-saintongeais. D'autres projets sont d'une envergure plus large, par exemple le Digital Language Diversity Project [•] (Soria *et al.*, 2016), qui vise à améliorer la maturité numérique (en anglais, *digital readiness*) de langues minoritaires d'Europe, y compris le breton (Hicks, 2017).

Pour notre recherche, nous nous appuyons sur des inventaires existants, par exemple ceux d'Ethnologue [•], d'ELDA soutenu par la DGLFLF (Grouas *et al.*, 2016), et sur d'autres ressources existantes mais difficiles d'accès ou d'utilisation, ainsi que sur de nouvelles ressources et annotations produites par nos soins. Parmi les ressources déjà existantes, nous pouvons citer les grandes bases de textes et de la parole multilingues, par exemple ceux sur Opus (Tiedemann, 2016) pour des textes parallèles, ainsi que les données monolingues et parallèles issues du Web tel que CCMatrix (Schwenk *et al.*, 2021b), Wikimatrix (Schwenk *et al.*, 2021a) et Oscar (Ortiz Suárez *et al.*, 2019). La couverture des langues de France autres que le français y est cependant pour le moment limitée. Des bases multilingues spécialisées dans la parole et focalisées sur les langues d'intérêt existent aussi, notamment Cocoon [•] (COllections de COrpus Oraux Numériques) (Wion, 2023) et l'atlas sonore des langues régionales de France [•] (Boula de Mareüil *et al.*, 2017). Il existe également des bases de données spécifiques à une

langue, qui sont particulièrement importantes pour couvrir les langues à faibles ressources du projet. Par exemple, pour le picard, il existe PICARTEXT (Eloy *et al.*, 2015), qui couvre à la fois corpus et lexiques, pour l'occitan, BaTelÒc (Bras & Vergez-Couret, 2013), OcWikiDisc (Miletic & Scherrer, 2022) et un *treebank* UD (Miletic *et al.*, 2020), ou pour l'alsacien, MeThAL (Fabo, 2023).

Un des outils essentiels pour la création de corpus de bonne qualité issus du web est un détecteur de langue, qui permette d'identifier les langues les moins dotées, mais aussi de classer des textes par dialecte. Cette tâche est rendue compliquée à cause de la variation et de l'absence de standardisation de leurs graphies et de la présence d'alternance codique (Ferguson, 1959; Fishman, 1967; Tabouret-Keller, 2021). Nous nous appuyons sur des classifications de langues existantes (Cerquiglini, 1999) pour essayer de couvrir un éventail de langues aussi large que possible.

Nous cherchons également à couvrir d'autres aspects minorisés des langues, tels que le domaine, le genre, le sociolecte et même l'état historique de la langue. Il est important de collecter à la fois des données historiques (à des fins patrimoniales) et des données contemporaines (pour encourager la vitalité numérique et à des fins éducatives). Les ressources telles que celles collectées à partir des médias sociaux sont d'une grande importance, comme celles produites dans le cadre du projet SoSweet pour le français (Icard *et al.*, 2023), mais aussi les ressources provenant de textes anciens, comme ceux numérisés dans le cadre de Gallic(orpor)a [•] (Sagot *et al.*, 2022).

Enfin, le projet souhaite développer des outils pour la création et l'exploitation de corpus. Il s'agit d'outils performants d'identification des langues, d'étiqueteurs et de parseurs pour l'analyse linguistique, de modèles de langue et de systèmes de traduction automatique. Le développement de ces outils s'appuie sur les travaux antérieurs concernant les langues d'intérêt, par exemple des modèles de langue tels que CamemBERT-v2 (Antoun *et al.*, 2024), des modèles d'identification de la langue tels que GlotLID (Kargaran *et al.*, 2023), et des ressources auxiliaires telles que des listes de mots (Caswell *et al.*, 2020), des étiqueteurs morphologiques (Millour *et al.*, 2020), et des systèmes de traduction automatique tels qu'Apertium (Forcada *et al.*, 2011), qui sont particulièrement bien adaptés à la traduction entre langues peu dotées et similaires (Armentano i Oller *et al.*, 2006). Un dernier ensemble de ressources à ne pas négliger est celui des modèles multilingues, entraînés sur des langues similaires aux langues d'étude, par exemple le catalan et l'espagnol pour l'occitan. Concernant la reconnaissance de la parole, le modèle Whisper (Radford *et al.*, 2023), entraîné sur 97 langues, ouvre de nouvelles portes. Ce modèle semble prometteur dans le cas des langues peu dotées, où les données disponibles sont insuffisantes seules pour entraîner un modèle classique, et où les connaissances déjà apprises par le modèle pourraient être transférées, en le spécialisant sur une langue inconnue. Pour le texte, il existe plusieurs modèles candidats possibles, notamment les derniers modèles de langue génératifs tels que Llama3 (Grattafiori *et al.*, 2024) et Mistral (Jiang *et al.*, 2023) et des systèmes de traduction automatique multilingues tels que NLLB (Costa-jussà *et al.*, 2022).

4 Création et mise à disposition de jeux de données

Le projet cherche à enrichir l'offre de jeux de données disponibles. Dans le cadre de cette publication, nous ne détaillons pas l'ensemble du travail effectué pendant le projet : nous présentons d'abord un tour d'horizon des tâches accomplies et à accomplir, puis un exemple de corpus que nous avons construit et publié.

4.1 Identification des tâches et travaux réalisés

Nous identifions les tâches suivantes pour structurer notre travail :

- inventaire des sources de données (corpus librement accessibles existant sous une forme directement exploitable ; données recueillies par les linguistes ; collections d'institutions patrimoniales, dont certaines sont déjà exploitables et d'autres ne le sont pas, notamment celles de la BNF pour l'écrit, celles de l'INA pour l'oral ; données issues des associations et organisations de défense, promotion ou étude des langues de France) ;
- collecte des données, qui inclut la fouille de corpus en dehors des répertoires déjà connus ;
- clarification des contraintes juridiques et des contreparties financières ;
- normalisation, documentation et structuration des données ;
- prétraitement semi-automatique des données ;
- mise à disposition et indexation des corpus.

Nos créations incluent un schéma XML-TEI pour la description fine des langues, une pipeline de production de données allant de l'image à l'encodage XML-TEI – dont LADaS, un jeu de données pour l'analyse de la mise en page [•] –, et des scripts de conversion de documents XML et HTML en XML-TEI. Ces réalisations ont permis de créer des jeux de données variés couvrant différentes périodes, langues, genres et types de sources tels que : le corpus Molyé, le forum de discussion en occitan *Forum Occitania*, des textes de fiction en picard issus d'un concours d'écriture organisé par l'Agence régionale de la langue picarde, et des monographies et documents administratifs de l'époque moderne fournies par la Bibliothèque numérique de la Sorbonne.

4.2 Molyé, un corpus pour l'étude des créoles à base française

Parmi les corpus produits dans le cadre de COLaF, le corpus Molyé permet l'analyse du contact entre les langues en France et dans ses anciennes colonies (Dent *et al.*, 2024) [•], en particulier l'émergence et la diffusion des langues de contact, et principalement des créoles à base française.

Recherche de documents Pour produire ce corpus, nous avons employé des méthodes d'identification de langues par listes de mots. Cette stratégie permet de récupérer une grande quantité de données brutes efficacement. Cette approche facilite l'identification de passages courts dans les langues cibles (créoles) lorsqu'ils sont intégrés dans des documents écrits majoritairement en français ou dans une autre langue répandue. Les deux moteurs de recherche principaux utilisés pour effectuer ces recherches sont Google Books et Gallica. De cette manière, nous avons pu identifier des centaines de pièces de théâtre, récits de voyage, romans et autres types de document contenant des indications linguistiques ou métalinguistiques sur le contact des langues en France et dans ses colonies du début de l'époque moderne au XX^{e} siècle.

Classification des variétés Bien que certains traits linguistiques soient significatifs et peuvent donc servir à identifier facilement une langue, la classification de plusieurs des textes nous a posé question. En effet, pour les langues créoles (et leur prédécesseurs immédiats) aux XVIII^{e} et XIX^{e} siècles, l'usage de certains marqueurs préverbaux au lieu de la conjugaison flexionnelle est un des shibboleths les plus importants pour les distinguer du soit-disant français simplifié. Cependant, dans certains textes, il est très difficile d'établir de quel créole il s'agit. Par exemple, dans le texte de *Le duel singulier* (Dorvigny, 1800), la variété en question est simplement étiquetée "patois". De plus, l'histoire se

déroule à Boston. Si certains éléments du syntagme verbal, particulièrement l'usage de *fini* pour exprimer l'aspect perfectif, et le choix de *assisé* pour 's'asseoir', nous amènent à croire que l'écrivain des passages (qui n'est pas forcément Dorvigny) a envisagé le parler de l'île de France à la fin du XVIII[e] s. (actuellement l'île Maurice), d'autres éléments compliquent cette hypothèse. On trouve notamment le possessif postposé (*maître à moi*, *billet à vous*), typique des Antilles et à l'époque déjà devenu stéréotype à Paris, à côté du possessif préposé (e.g. *son zépée*, *son canon*) du mauricien actuel.

Dans d'autres cas, les données détaillées dont nous disposons sur la vie de l'écrivain nous permettent de faire des hypothèses plus exactes. Par exemple, bien que *Ferragan, chef de brigands* (Lorquet, 1827) ne mentionne pas l'île Maurice, l'auteur y a vécu longtemps, et la variété utilisée par le personnage Francisco est vraisemblablement le créole mauricien du début du XIX[e] siècle. De plus, la différence principale entre cette variété et le créole mauricien de nos jours, qui se trouve dans l'usage du pronom de la troisième personne, *li*, est corroborée par un autre témoin contemporain.

Les représentations stéréotypées des différentes sous-populations sont généralement fondées sur un nombre assez restreint de jeux de mots et sont donc plus faciles à regrouper. Néanmoins, il semble manquer d'étiquettes standard pour certaines variétés identifiables. Par exemple, nous avons retrouvé une étude ancienne rédigée en allemand qui parle du "deutsch-französische Jargon" (Damm, 1911) ainsi qu'un article français qui décrit "le langage des 'Suisses' de Molière" (Haas, 2015). Nous avons regroupé ces deux appellations sous l'étiquette de "baragouin germanique", pour résumer l'idée qu'il est d'abord et principalement utilisé pour se moquer des erreurs morphosyntaxiques et phonétiques des locuteurs de langues germaniques qui parlent le français, plutôt que des locuteurs des langues romanes ou non-européennes. Au fil du temps, et surtout au XIX[e] siècle, d'autres conventions se développent pour représenter ces derniers. Pour les éditions suivantes du corpus, nous envisageons aussi d'inclure ces variétés, afin de rendre compte de leur évolution.

Construction du corpus Notre corpus est constitué de documents créoles qui nous sont parvenus par hasard et d'œuvres littéraires manifestement artificieuses, et n'est donc pas représentatif des variétés étudiées. Nous visons à schématiser les distributions de certaines variables pour faciliter le développement d'un modèle de divergence grammaticale fondée sur des règles. Pour cette raison, pour la première version du corpus, nous avons retenu une soixantaine de textes sur les plus de 250 identifiés, en fonction de leur intérêt linguistique et du degré de traitement nécessaire à leur conversion en XML-TEI (TEI Consortium, 2023), le format pivot pour les textes au sein de COLaF.

Une partie importante des pièces de théâtre du corpus provient du corpus *Théâtre classique* (Fièvre, 2007), qui contient des pièces déjà encodées en XML, ne nécessitant donc qu'une adaptation de schéma. Certaines œuvres – comme *Le duel singulier* (Dorvigny, 1800), semblent avoir échappé aux études antérieures. Leurs données-texte n'étant pas directement accessibles, nous les avons transcrites, soit automatiquement sur l'interface graphique eScriptorium (Kiessling *et al.*, 2019), qui utilise le moteur d'OCR Kraken (Kiessling, 2019), en employant le modèle d'OCR CATMuS Print (Gabay *et al.*, 2024), soit manuellement, dans le cas des citations les plus intéressantes. Pour les textes en créoles anciens, nous avons aussi privilégié des documents numérisés absents des corpus existants.

Les documents ont ainsi été produits en XML-TEI, selon le schéma spécifié pour les besoins de COLaF, dans l'objectif d'assurer un cadre commun – pour la structure des textes comme pour leurs métadonnées – aux corpus textuels produits au sein du projet. Ce schéma de spécialisation [•] a été développé en début de projet et est progressivement mis à jour au fil des corpus traités et des cas rencontrés. Ce schéma [•] vise à structurer les documents traités sur trois axes :

— la structuration des données : encodage de la structure intellectuelle (paragraphes, vers, titres...) et sémantique (informations linguistiques et étiquettes grammaticales) du document ;
— l'identification des langues : finesse de description dans le corps du texte et dans les métadonnées (informations géographiques, temporelles et de locuteurs) ;
— les métadonnées : informations bibliographiques, chronologiques et géographiques associées à la production du document.

Résultats Cette première version du corpus [3] se compose de 63 documents différents de genres variés : pièces de théâtre, poèmes, documents administratifs et récits de voyages, représentant environ 172K tokens. Parmi ces 63 documents, 41 contiennent des représentations stéréotypées, dont 33 touchent les locuteurs de l'allemand, de l'anglais ou du néerlandais, tandis que 15 se moquent des paysans français et quatre représentent un accent gascon. Parmi les 22 documents qui concernent les langues créoles, nous avons quelques attestations des créoles antillais, haïtien, louisianais, mauricien et réunionnais qui avaient déjà été identifiées par des travaux antérieurs, tels que Chaudenson (1981), Bollée (2007) et Hazaël-Massieux (2008), mais aussi certaines trouvailles comme *Le duel singulier*.

Nous avons mené sur ce corpus une étude sur les représentations stéréotypées du français utilisé par les étrangers européens au XVII^e^ siècle. Ces derniers avaient tendance à employer des pronoms de la troisième personne, notamment *li*, comme marqueurs de pré-verbaux dépourvus de contenu sémantique, généralisés à toutes les personnes grammaticales. Voici un exemple du début du XVIII^e^ siècle donné par de La Colonie (1737, 276) :

> ... mon cher quer, fous **ly** être trop honnête homme pour **ly** faire chamais l'oubliance des pelles promesses que fous **ly** afoir fait à moi si tentrement.

Plus récemment, nous avons trouvé d'autres documents qui nous aident à mieux comprendre les parallèles entre l'Europe, l'Amérique et l'océan Indien. Par exemple, les stratégies de pluralisation distinguent clairement les créoles (à base française) du français, ainsi que les parlers de l'océan Indien de ceux des Amériques. Plus précisément, les créoles d'Amérique utilisent les mots *yo* et *yé*, formes régionales dérivées de *eux*, comme des pronoms de la troisième personne du pluriel. Pour certaines langues, à savoir celles de la Louisiane, la Guyane et Haïti, ces pronoms peuvent servir encore de marqueurs de pluralité à la fin des groupes nominaux. En revanche, les créoles de l'océan Indien présentent le pronom *zot* qui vient d'une forme renforcée (soit *eux-autres*, soit *les autres*) ainsi que le marqueur *bann* qui précède le nom pluralisé. Compte tenu de ces différences structurales assez nettes entre les deux aires, certains chercheurs, notamment Chaudenson (1981), ont postulé que les parlers actuels sont les langues filles d'au moins deux proto-créoles distincts.

En élargissant le corpus, nous avons constaté que certains textes de l'océan Indien qui remontent au début du XIX^e^ siècle, nommément le roman *Ferragan, chef de brigands* (Lorquet, 1827), et les écritures créoles d'Auguste Le Duc (Kriegel, 2023) utilisent le pronom *li* pour exprimer la troisième personne du pluriel, en parallèle de l'usage de *zot*. Par exemple :

> Astore **zautes** dansé autour çà di feu-là : **li** santé, **li** crié, **li** faire son tapaze. Après çà **zautes** sisé , et à v'là ça vilains li siens-là , mo' dire vous, **li** manze son lé corps à ça pauvres zens-làque mo' ité parlé vous. (Lorquet, 1827, 282)

Cet usage suggère que l'ascendance de *zot* à l'île Maurice et ses dépendances n'est pas encore réalisée à l'époque. De plus, Robin (1807) indique que *li* est utilisé en Louisiane au pluriel, à la

3. Pour une liste des documents détaillée avec métadonnées, se référer à https://defi-colaf.github.io/Molye/dataset_HTML/index.html.

fois comme pronom et comme article. Puisque l'usage du *li* pluriel est aussi attesté dans *Monsieur de Pourceaugnac* (Molière, 1670), notre étude permet de soutenir l'hypothèse que les isoglosses observées de nos jours entre différents territoires créolophones, comme cette distinction entre les formes du pluriel, se sont consolidées au fil des siècles plutôt qu'au début de leurs colonisations respectives (Arends, 1992).

5 Création d'outils et modèles

La seconde partie du projet consiste à adapter des outils et modèles TAL existants ou en créer de nouveaux qui soient adaptés à la variabilité et à la quantité faible de données, éléments caractéristiques de nos langues peu dotées. De même que dans la section précédente, nous choisissons de présenter une des tâches effectuées : une chaîne de traitements pour la reconnaissance de la parole.

5.1 Identification des tâches et travaux réalisés

Nous organisons notre travail selon les tâches suivantes :

- identification des langues peu dotées, qui représentent à la fois un défi scientifique (identification fine de langues moins standardisées) et un défi technique de passage à l'échelle ;
- adaptation et développement d'outils de capture de l'information (OCR, ASR) ;
- exploration des méthodes multilingues, incluant : la mise en commun des ressources de plusieurs langues afin d'avoir de plus grands corpus d'entraînement ; la question de la traduction automatique vers un continuum dialectal ;
- développement d'outils résistants au bruit, car, qu'il s'agisse de retranscription audio, de réseaux sociaux ou de documents manuscrits retranscrits automatiquement, les données produites par ces processus sont généralement bruitées. La quantité de données étant généralement réduite, cela rend le bruit d'autant plus visible et problématique ;
- génération de la parole vers un continuum linguistique ;
- développement d'outils pour la génération de vidéo en langue des signes.

Un exemple de nos réalisations sur ces tâches est l'affinage et le pré-entraînement de modèles PIE (Manjavacas *et al.*, 2019), des modèles d'annotation linguistique, afin d'améliorer les performances de tels annotateurs sur des langues peu dotées et à variation dialectale, grâce à des jeux de données plus conséquents dans des langues similaires [4].

5.2 Reconnaissance automatique de la parole

La reconnaissance automatique de la parole (*automatic speech recognition* ou ASR) permet de transcrire automatiquement des données audio. Pour les langues les moins dotées, les seuls corpus disponibles sont des données audio non-transcrites, plus difficiles à utiliser tant comme corpus d'entraînement pour des applications TAL que pour un locuteur. Pour les linguistes de terrain, un outil d'ASR, même imparfait, serait un gain de temps considérable pour préparer et enrichir leurs données.

4. Documentation : https://github.com/DEFI-COLaF/modeles-papie/blob/main/occitan.md.

TABLE 1 – Résultats quantitatifs de l'affinage de Whisper.

Langue	Modèle	Données d'entraînement	Temps d'entraînement	WER	CER
Basque	base	sans affinage		97.5%	25.2%
	large	sans affinage		43.7%	8.1%
	base	116h20	1h	15.6%	3.6%
	large	116h20	26h06	8.6%	1.5%
Alsacien	base	sans affinage		96.1%	65.4%
	large	sans affinage		92.6%	62.9%
	base	5h06	7 min	61.0%	35.8%
	large	5h06	15 min	48.5%	26.7%
Shimaoré	base	sans affinage		98.9%	45.0%
	large	sans affinage		89.0%	41.0%
	base	1h28	4 min	76.9%	35.6%
	large	1h28	5 min	69.2%	29.1%

TABLE 2 – Résultats qualitatifs de l'affinage de Whisper.

Langue	Type	Phrase
Basque	vérité terrain	hezkuntzak prestatu zituen probak pisa eta antzekoak eredu
Basque	whisper large affiné	hazkuntzek prestatu zituen probak bisa eta antzekoak ere du
Alsacien	vérité terrain	Will mer so Ùstellige gemàcht han
Alsacien	whisper medium affiné	Weil wir so Ausstellungen gemacht haben, ohne
Shimaoré	vérité terrain	Tsingaliya tsirongoa mdridrimishio ini vanu comme d'habitudi riparawo kilakati.
Shimaoré	whisper large affiné	Tsi ganliya tsirongoa bon ini vanu kamudabitru idi dde iparawo kilakazi.

Les modèles d'ASR multilingues se construisent à partir d'une phase d'entraînement initial sur plusieurs centaines d'heures de données dans des langues mieux dotées, suivie par une étape d'affinage (*fine-tuning*) sur des langues moins dotées. Cette méthode permet d'obtenir de bons résultats avec relativement peu de données d'entraînement pour une langue donnée. Nous utilisons le modèle Whisper (Radford *et al.*, 2023). Notre objectif est d'évaluer la capacité du modèle à apprendre une langue inconnue à partir de très petits volumes de données, identifier les difficultés les plus fréquentes et établir une chaîne de traitements généralisable à n'importe quelle langue.

L'affinage de modèles Whisper requiert des segments audio de 30 secondes maximum. Un segment trop court ne fournit pas assez de contexte, et un segment trop long peut cumuler des décalages entre la prédiction et la transcription. Nous constatons qu'un modèle entraîné sur des segments trop courts tend à s'arrêter avant d'avoir terminé une transcription. Empiriquement, l'idéal semble être des segments de 10 à 30 secondes constitués d'une ou plusieurs phrases complètes.

Puisqu'il est nécessaire de segmenter les données vocales d'entraînement, et donc leur transcription, il est également nécessaire d'aligner cette transcription afin de pouvoir faire correspondre la segmentation. Cette étape d'alignement est non-triviale. Si les données d'entraînement sont issues d'émissions de télévision ou de radio, nous utilisons inaSpeechSegmenter (Doukhan *et al.*, 2018) pour en exclure les segments musicaux et les silences. Pour l'alignement de la transcription, nous avons testé plusieurs méthodes : Audapolis, Montreal Forced Aligner (McAuliffe *et al.*, 2017) (MFA) et l'alignement forcé avec Wav2Vec2 (Baevski *et al.*, 2020). MFA repose sur des modèles acoustiques spécifiques à chaque langue et nécessite un corpus d'entraînement avec des annotations phonétiques et textuelles pour effectuer un alignement précis. Cette méthode s'avère particulièrement performante lorsque la langue considérée (ou une langue suffisamment proche dans notre cas) est supportée par le modèle, permettant ainsi un alignement optimal. À l'inverse, Wav2Vec adopte une approche multilingue, capable de traiter diverses langues sans nécessiter de spécification préalable, offrant ainsi une flexibilité accrue, en particulier dans des contextes multilingues. En effet, nous avons rencontré des difficultés d'alignement avec MFA lorsque la langue considérée n'était pas supportée, comme c'est le cas pour le shimaoré, pour lequel Wav2Vec a montré de meilleures performances.

Nous avons fait des expériences sur trois langues de France, choisies pour représenter des situations différentes : le basque, l'alsacien et le shimaoré. Le basque représente une situation idéale pour l'apprentissage automatique : avec plus de 500 heures dans Common Voice, il constitue un large corpus cohérent déjà découpé en phrases. De plus, il est déjà couvert par Whisper, mais avec un taux d'erreur sur les caractères (*character error rate* ou CER) élevé, ce qui nous permet d'évaluer

facilement l'influence de l'affinage. L'alsacien nous intéresse pour sa forte variation dialectale et son absence d'orthographe standardisée. Autre aspect intéressant, les langues les plus proches de l'alsacien sont parmi les mieux couvertes par Whisper : l'allemand (5^{e} langue la mieux couverte, 5.7% CER), et le français (16^{e} langue, 10.8% CER). Cela nous permet de partir sur une base solide. En contraste, le shimaoré, langue native d'un territoire d'outre-mer, est philogénétiquement éloignée des langues européennes, qui sont les mieux couvertes par Whisper. En effet, la langue la plus proche présente dans Whisper est le swahili, sur lequel les modèles sont les moins performants, avec 51.2% de CER. Le shimaoré permet donc d'étudier une situation difficile pour l'apprentissage automatique.

Les résultats de l'affinage se trouvent dans la table 1. La différence entre les modèles *base*, *medium* et *large* réside dans le nombre de paramètres et dans la quantité de données d'entraînement initial, avant phase d'affinage. Pour les trois langues, le modèle non affiné obtient de mauvais résultats. L'affinage les améliore, mais avec de fortes disparités entre les langues sur le nombre de points gagnés.

Dans le cas du basque, nous obtenons une grande amélioration des résultats, ce qui valide l'hypothèse que l'affinage d'un modèle Whisper est une bonne méthode pour obtenir un système ASR performant lorsqu'on a à disposition un grand volume de données d'entraînement.

Dans tous les cas, on observe une grande disparité entre les mesures de taux d'erreur sur les mots (*word error rate* ou WER) et de CER. Les phrases dans le tableau 2 montrent que les mots produits sont phonétiquement proches de la réalité terrain. Cependant, le modèle n'a pas pu prendre connaissance d'un vocabulaire suffisamment varié dans la phase d'apprentissage, et privilégie donc des mots connus de la langue parente (par exemple allemand standard pour l'alsacien), des mots vus à l'étape d'affinage (qui deviennent sur-employés), ou inventent des mots inexistants dans la langue. Les mots produits étant proches de la vérité terrain, le CER est moins élevé que le WER. Par ailleurs, l'alsacien et le shimaoré présentent de nombreux cas d'alternance codique avec le français et, pour l'alsacien, l'allemand, qui résultent en mots très mal transcrits. Les résultats sur l'alsacien et sur le shimaoré étant assez mauvais, il est difficile d'établir des conclusions sur la différence entre les deux langues.

En prolongement de ce travail, nous songeons à la possibilité d'utiliser du texte écrit lors de la phase d'apprentissage, sans audio, pour aider à l'acquisition du vocabulaire.

6 Conclusion

Le projet COLaF vise à rendre disponibles des corpus et outils pour le français et pour les autres langues de France dans toute leur diversité, sous forme écrite, orale ou signée. Ses premières productions sont des corpus de textes structurés, l'établissement d'un schéma XML-TEI, un travail d'analyse automatique de la mise en page de documents anciens et la création d'un modèle d'annotation linguistique. À cela s'ajoute une chaîne de traitement pour l'entraînement de modèles d'ASR, testée sur des cas de figure variés, que nous continuons de perfectionner.

Notre projet reste ouvert aux collaborations : notre expertise en traitement automatique des langues, en numérisation et structuration de corpus (OCR, ASR, HTR, XML-TEI) peut contribuer à la valorisation de langues et de variétés linguistiques pour lesquelles d'autres institutions et laboratoires disposent de données et de questions de recherche que des études sur corpus peuvent contribuer à éclairer.

Références

ADDA G., VASILESCU I. & YVON F. (2023). *Language Report French*, In G. REHM & A. WAY, Éds., *European Language Equality : A Strategic Agenda for Digital Language Equality*, p. 139–142. Springer International Publishing : Cham. DOI : 10.1007/978-3-031-28819-7_16.

ANTOUN W., KULUMBA F., TOUCHENT R., ÉRIC DE LA CLERGERIE, SAGOT B. & SEDDAH D. (2024). Camembert 2.0 : A smarter french language model aged to perfection.

ARENDS (1992). Towards a Gradualist Model of Creolization. In F. BYRNE & J. HOLM, Éds., *Atlantic Meets Pacific : A Global View of Pidginization and Creolization*, volume 11 de *Creole Language Library*. John Benjamins Publishing Company.

ARMENTANO I OLLER C., FORCADA & L. M. (2006). Open-source machine translation between small languages : Catalan and Aranese Occitan. In *Proceedings of the 5th Workshop on Strategies for developing machine translation for minority languages*, p. 51–54, Genoa, Italy.

BAEVSKI A., ZHOU H., MOHAMED A. & AULI M. (2020). wav2vec 2.0 : a framework for self-supervised learning of speech representations. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA : Curran Associates Inc. DOI : 10.5555/3495724.3496768.

BERNHARD D. (2019). Natural language processing for regional languages of France : Lessons learned from the RESTAURE project. In *New Ways of Analyzing Dialectal Variation*.

BERNHARD D. & VERGEZ-COURET M. (2015). Le projet RESTAURE. In *Colloque sur les technologies pour les langues régionales de France (TLRF 2015)*, Les technologies pour les langues régionales de France, p. 96–100, Meudon, France : Délégation générale à la langue française et aux langues de France, laboratoire de recherche en informatique pluridisciplinaire (LIMSI) - Centre national de la recherche scientifique (CNRS), Institut des technologies multilingues et multimédias de l'information (IMMI) Ministère de la Culture et de la Communication - Délégation générale à la langue française et aux langues de France. HAL : hal-01297835.

BOLLÉE A. (2007). *Deux textes religieux de Bourbon du 18e siècle et l'histoire du créole réunionnais*. opus.

BOULA DE MAREÜIL P., VERNIER F. & RILLIARD A. (2017). Enregistrements et transcriptions pour un atlas sonore des langues régionales de France. *Géolinguistique*, **17**, 23–48. HAL : hal-01719532.

BRAS M. & VERGEZ-COURET M. (2013). BaTelÒc : a Text Base for the Occitan Language. In *Proceedings of the First International Conference on Endangered Languages in Europe*, Alcanena, Portugal.

CASWELL I., BREINER T., VAN ESCH D. & BAPNA A. (2020). Language ID in the Wild : Unexpected Challenges on the Path to a Thousand-Language Web Text Corpus. In D. SCOTT, N. BEL & C. ZONG, Éds., *Proceedings of the 28th International Conference on Computational Linguistics*, p. 6588–6608, Barcelona, Spain (Online) : International Committee on Computational Linguistics. DOI : 10.18653/v1/2020.coling-main.579.

CERQUIGLINI B. (1999). *Les Langues de France : 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*. Rapport interne, Ministère de l'éducation nationale, de la recherche et de la technologie.

CHAUDENSON R. (1981). *Textes Créoles Anciens : La Réunion et Île Maurice : Comparaison et Essai d'analyse*. H. Buske.

COLLECTIF & CERQUIGLINI B. (2003). *Les Langues de France*. Presses Universitaires de France - PUF.

COSTA-JUSSÀ M. R., CROSS J., ÇELEBI O., ELBAYAD M., HEAFIELD K., HEFFERNAN K., KALBASSI E., LAM J., LICHT D., MAILLARD J., SUN A. Y., WANG S., WENZEK G., YOUNGBLOOD A., AKULA B., BARRAULT L., GONZALEZ G. M., HANSANTI P., HOFFMAN J., JARRETT S., SADAGOPAN K. R., ROWE D., SPRUIT S., TRAN C., ANDREWS P., AYAN N. F., BHOSALE S., EDUNOV S., FAN A., GAO C., GOSWAMI V., GUZMÁN F., KOEHN P., MOURACHKO A., ROPERS C., SALEEM S., SCHWENK H. & WANG J. (2022). No language left behind : Scaling human-centered machine translation. *CoRR*, **abs/2207.04672**. DOI : 10.48550/ARXIV.2207.04672.

DAMM O. (1911). *Der deutsch-französische Jargon in der schönen französischen Literatur*. Emli Eberling.

DE LA COLONIE J.-M. (1737). *Memoires de Monsieur de La Colonie, marechal de camp des armées de l'électeur de Baviere...* Aux dépens de la compagnie.

DENT R., JANES J., CLERICE T., ORTIZ SUAREZ P. & SAGOT B. (2024). Molyé : A Corpus-based Approach to Language Contact in Colonial France. In M. HÄMÄLÄINEN, E. ÖHMAN, S. MIYAGAWA, K. ALNAJJAR & Y. BIZZONI, Éds., *Proceedings of the 4th International Conference on Natural Language Processing for Digital Humanities*, p. 189–199, Miami, USA : Association for Computational Linguistics.

DORVIGNY L. F. (1800). *Le duel singulier, comédie en un acte et en prose*. Au magasin de pièces de théâtre.

DOUKHAN D., CARRIVE J., VALLET F., LARCHER A. & MEIGNIER S. (2018). An open-source speaker gender detection framework for monitoring gender equality. In *2018 IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*.

ELOY J.-M., MARTIN F. & REY C. (2015). PICARTEXT : Une ressource informatisée pour la langue picarde. In *Proceedings of TALaRE 2015 - Traitement Automatique des Langues Regionales de France et d'Europe*, Caen, France.

FABO P. R. (2023). The methal alsatian theater corpus and related resources : Work done and perspectives. *Actes des Journées LIFT 2023*.

FERGUSON C. A. (1959). Diglossia. *WORD*, **15**(2), 325–340. DOI : 10.1080/00437956.1959.11659702.

FIÈVRE P. (2007). Théâtre classique.

FISHMAN J. A. (1967). Bilingualism with and without diglossia; diglossia with and without bilingualism. *Journal of Social Issues*, **23**(2), 29–38. _eprint : https ://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1540-4560.1967.tb00573.x, DOI : 10.1111/j.1540-4560.1967.tb00573.x.

FORCADA M. L., GINESTÍ-ROSELL M., NORDFALK J., O'REGAN J., ORTIZ-ROJAS S., PÉREZ-ORTIZ J. A., SÁNCHEZ-MARTÍNEZ F., RAMÍREZ-SÁNCHEZ G. & TYERS F. M. (2011). Apertium : a free/open-source platform for rule-based machine translation. *Machine translation : MT*, **25**(2), 127–144.

GABAY S., CLÉRICE T., JACSONT P., LEBLANC E., JEANNOT-TIROLE M., SOLFRINI S., DOLTO S., GOY F., LUJÁN C. C., ZAGLIO M., PERREGAUX M., JANES J., SAGOT B., BAWDEN R., DENT R., NÉDEY O. & CHAGUÉ A. (2024). Reconnaissance des écritures dans les imprimés. In *Humanistica 2024*, OCR, Meknès, Morocco : Association francophone des humanités numériques.

GRATTAFIORI A., DUBEY A. & AL (2024). The llama 3 herd of models.

GROUAS T., MAPELLI V. & SAMIER Q. (2016). Review on the existing language resources for languages of France. In N. CALZOLARI, K. CHOUKRI, T. DECLERCK, S. GOGGI, M. GROBELNIK, B. MAEGAARD, J. MARIANI, H. MAZO, A. MORENO, J. ODIJK & S. PIPERIDIS, Éds., *Procee-*

dings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), p. 4539–4542, Portorož, Slovenia : European Language Resources Association (ELRA).

HAAS W. (2015). « Déguisé en Suisse » : les « Suisses » de Molière et leur langage. *Littératures classiques*, **87**(2), 191–189. DOI : 10.3917/licla1.087.0191.

HAZAËL-MASSIEUX M.-C. (2008). *Textes anciens en créole français de la Caraïbe : Histoire et analyse*. Editions Publibook.

HICKS D. (2017). *Breton — a digital language ?* Rapport interne, Erasmus+ Programme.

ICARD B., CLAVEAU V., ATEMEZING G. & EGRÉ P. (2023). Un traitement hybride du vague textuel : du système expert VAGO à son clone neuronal. In C. SERVAN & A. VILNAT, Éds., *Actes de CORIA-TALN 2023. Actes de la 30e Conférence sur le Traitement Automatique des Langues Naturelles (TALN), volume 1 : travaux de recherche originaux – articles longs*, p. 151–163, Paris, France : ATALA.

JIANG A. Q., SABLAYROLLES A., MENSCH A., BAMFORD C., CHAPLOT D. S., DE LAS CASAS D., BRESSAND F., LENGYEL G., LAMPLE G., SAULNIER L., LAVAUD L. R., LACHAUX M., STOCK P., SCAO T. L., LAVRIL T., WANG T., LACROIX T. & SAYED W. E. (2023). Mistral 7b. *CoRR*, **abs/2310.06825**. DOI : 10.48550/ARXIV.2310.06825.

KARGARAN A., IMANI A., YVON F. & SCHUETZE H. (2023). GlotLID : Language identification for low-resource languages. In H. BOUAMOR, J. PINO & K. BALI, Éds., *Findings of the Association for Computational Linguistics : EMNLP 2023*, p. 6155–6218, Singapore : Association for Computational Linguistics. DOI : 10.18653/v1/2023.findings-emnlp.410.

KIESSLING B. (2019). Kraken - a universal text recognizer for the humanities. In *Digital Humanities 2019*, Utretch, Netherlands.

KIESSLING B., TISSOT R., STOKES P. & STÖKL BEN EZRA D. (2019). eScriptorium : An Open Source Platform for Historical Document Analysis. In *2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)*, volume 2, p. 19–19. DOI : 10.1109/ICDARW.2019.10032.

KRIEGEL S. (2023). Les allocutions en créole d'Auguste Le Duc, Galéga en 1835. In GRANGET, CYRILLE, REPISO, I. . F. SING & G. (EDS.)., Éds., *Language, creoles, varieties : From emergence to transmission*, EuroSLA Studies. Language Science Press. DOI : 10.5281/zenodo.10280493, HAL : hal-04425336.

LEIXA J., MAPELLI V. & CHOUKRI K. (2014). *Inventaire des ressources linguistiques des langues de France*. Rapport interne, ELDA.

LORQUET H.-L. (1827). *Ferragan, Chef de Brigands*. Pigoreau.

MANJAVACAS E., KÁDÁR Á. & KESTEMONT M. (2019). Improving Lemmatization of Non-Standard Languages with Joint Learning. In *Proceedings of the 2019 Conference of the North*, p. 1493–1503, Minneapolis, Minnesota : Association for Computational Linguistics. DOI : 10.18653/v1/N19-1153.

MARIANI J., PAROUBEK P., FRANCOPOULO G., MAX A., YVON F. & ZWEIGENBAUM P. (2012). *The French Language in the Digital Age / La langue française à l'ère numérique*. META-NET White Paper Series. Georg Rehm and Hans Uszkoreit (Series Editors). Springer.

MCAULIFFE M., SOCOLOF M., MIHUC S., WAGNER M. & SONDEREGGER M. (2017). Montreal forced aligner : Trainable text-speech alignment using kaldi. In *Interspeech*.

MILETIC A., BRAS M., VERGEZ-COURET M., ESHER L., POUJADE C. & SIBILLE J. (2020). Building a Universal Dependencies treebank for Occitan. In N. CALZOLARI, F. BÉCHET, P. BLACHE, K. CHOUKRI, C. CIERI, T. DECLERCK, S. GOGGI, H. ISAHARA, B. MAEGAARD, J.

MARIANI, H. MAZO, A. MORENO, J. ODIJK & S. PIPERIDIS, Éds., *Proceedings of the Twelfth Language Resources and Evaluation Conference*, p. 2932–2939, Marseille, France : European Language Resources Association.

MILETIC A. & SCHERRER Y. (2022). OcWikiDisc : a corpus of Wikipedia talk pages in Occitan. In Y. SCHERRER, T. JAUHIAINEN, N. LJUBEŠIĆ, P. NAKOV, J. TIEDEMANN & M. ZAMPIERI, Éds., *Proceedings of the Ninth Workshop on NLP for Similar Languages, Varieties and Dialects*, p. 70–79, Gyeongju, Republic of Korea : Association for Computational Linguistics.

MILLOUR A., FORT K. & MAGISTRY P. (2020). Répliquer et étendre pour l'alsacien "étiquetage en parties du discours de langues peu dotées par spécialisation des plongements lexicaux" (replicating and extending for Alsatian : "POS tagging for low-resource languages by adapting word embeddings"). In G. ADDA, M. AMBLARD & K. FORT, Éds., *Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 33e édition), Traitement Automatique des Langues Naturelles (TALN, 27e édition), Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition). 2e atelier Éthique et TRaitemeNt Automatique des Langues (ETeRNAL)*, p. 29–37, Nancy, France : ATALA et AFCP.

MOLIÈRE J.-B. P. (1670). *Monsieur de Pourceaugnac*. Jean Ribou.

ORTIZ SUÁREZ P. J., SAGOT B. & ROMARY L. (2019). Asynchronous pipelines for processing huge corpora on medium to low resource infrastructures. In P. BAŃSKI, A. BARBARESI, H. BIBER, E. BREITENEDER, S. CLEMATIDE, M. KUPIETZ, H. LÜNGEN & C. ILIADI, Éds., *Proceedings of the Workshop on Challenges in the Management of Large Corpora (CMLC-7)*., p. 9–16, Cardiff, UK : Leibniz-Institut für Deutsche Sprache. DOI : 10.14618/ids-pub-9021.

RADFORD A., KIM J. W., XU T., BROCKMAN G., MCLEAVEY C. & SUTSKEVER I. (2023). Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23 : JMLR.org.

ROBIN C. C. (1807). *Voyages dans l'interieur de la Louisiane ...*, volume 2. F. Buisson.

SAGOT B., ROMARY L., BAWDEN R., ORTIZ SUÁREZ P. J., CHRISTENSEN K., GABAY S., PINCHE A. & CAMPS J.-B. (2022). Gallic(orpor)a : Extraction, annotation et diffusion de l'information textuelle et visuelle en diachronie longue. In *DataLab de la BnF : Restitution des travaux 2022*, Paris, France : DataLab de la BnF. HAL : hal-03930542.

SCHWENK H., CHAUDHARY V., SUN S., GONG H. & GUZMÁN F. (2021a). WikiMatrix : Mining 135M parallel sentences in 1620 language pairs from Wikipedia. In P. MERLO, J. TIEDEMANN & R. TSARFATY, Éds., *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics : Main Volume*, p. 1351–1361, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2021.eacl-main.115.

SCHWENK H., WENZEK G., EDUNOV S., GRAVE E., JOULIN A. & FAN A. (2021b). CCMatrix : Mining billions of high-quality parallel sentences on the web. In C. ZONG, F. XIA, W. LI & R. NAVIGLI, Éds., *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1 : Long Papers)*, p. 6490–6500, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2021.acl-long.507.

SORIA C., MARIANI J. & ZOLI C. (2013). Dwarfs Sitting on Giants' Shoulders : How LTs for Regional and Minority Languages Can Benefit from Piggybacking Major Languages. *FEL XVII : Endangered Languages Beyond Boundaries*.

SORIA C., RUSSO I., QUOCHI V., HICKS D., GURRUTXAGA A., SARHIMAA A. & TUOMISTO M. (2016). Fostering digital representation of EU regional and minority languages : the digital language diversity project. In N. CALZOLARI, K. CHOUKRI, T. DECLERCK, S. GOGGI, M.

GROBELNIK, B. MAEGAARD, J. MARIANI, H. MAZO, A. MORENO, J. ODIJK & S. PIPERIDIS, Éds., *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, p. 3256–3260, Portorož, Slovenia : European Language Resources Association (ELRA).

TABOURET-KELLER A. (2021). Classification des langues et hiérarchisation des langues en alsace. *Cahiers du plurilinguisme européen*, (13). DOI : 10.57086/cpe.1370.

TEI CONSORTIUM (2023). *TEI P5 : Guidelines for Electronic Text Encoding and Interchange Version 4.7.0*. Standard.

TIEDEMANN J. (2016). OPUS – parallel corpora for everyone. In *Proceedings of the 19th Annual Conference of the European Association for Machine Translation : Projects/Products*, Riga, Latvia : Baltic Journal of Modern Computing.

WION A. (2023). Cocoon : A Platform for Documenting the World's Languages. HAL : halshs-04161812.