Au-delà des normes : identifier et documenter les langues minorisées pour le traitement automatique des langues

Delphine Bernhard, Marianne Vergez-Couret, Estèle Dupuy

Cahiers du plurilinguisme européen
ISSN : 2970-989X

16 | 2024
Acteurs et facteurs de la vitalité de quelques langues régionales de France

Au-delà des normes : identifier et documenter les langues minorisées pour le traitement automatique des langues

Delphine Bernhard, Marianne Vergez-Couret et Estèle Dupuy

https://www.ouvroir.fr/cpe/index.php?id=1710

DOI : 10.57086/cpe.1710

Référence électronique
Delphine Bernhard, Marianne Vergez-Couret et Estèle Dupuy, « Au-delà des normes : identifier et documenter les langues minorisées pour le traitement automatique des langues », *Cahiers du plurilinguisme européen* [En ligne], 16 | 2024, mis en ligne le 17 décembre 2024, consulté le 17 décembre 2024. URL : https://www.ouvroir.fr/cpe/index.php?id=1710

Au-delà des normes : identifier et documenter les langues minorisées pour le traitement automatique des langues

Delphine Bernhard, Marianne Vergez-Couret et Estèle Dupuy

PLAN

TEXTE

Ces travaux ont été réalisés dans le cadre du projet ANR-21-CE27-0004 DIVITAL *soutenu par l'Agence nationale de la recherche.*

Introduction

1 Les langues qui actuellement ne sont ni standardisées et/ou ni officielles se trouvent souvent dans une situation de minorisation[1] du point de vue de leurs pratiques. Dans cette situation, on observe généralement que ces langues ont de moins en moins de locuteurs, qu'elles perdent en visibilité, qu'elles ne sont pas ou peu soutenues par l'État, qu'elles sont peu ou pas standardisées, qu'elles sont peu enseignées et rarement objet de la recherche scientifique. Cela peut même aller jusqu'à la contestation de leur existence (langues dites contestées, cf. Tamburelli et Tosco, 2021). Et, en conséquence de cette minorisation, en France, la production d'écrits n'a pas été valorisée au cours du siècle dernier, ces langues présentant une plus grande variation interne (diatopique, graphique…) et les développeurs d'applications et de ressources numériques ne les ont dès lors pas prises en considération. C'est pourquoi on les nomme aussi « langues peu dotées ». Dans cet article, nous proposons une réflexion sur les défis de la documentation de ces langues dites minorisées à partir de

travaux réalisés dans le cadre du projet DIVITAL [2] (Accroître la vitalité et la visibilité numérique des langues de France : descriptions linguistiques et corpus annotés, ANR-21-CE27-0004). Le projet DIVITAL a pour objectif d'améliorer la vitalité et la visibilité de plusieurs langues de France qui sont précisément des langues minorisées et peu dotées : l'alsacien, le corse, l'occitan et le poitevin-saintongeais. Il se situe à l'intersection de la linguistique descriptive et de la linguistique de corpus. Le but principal est de créer des ressources, notamment des corpus bruts et annotés, avec plusieurs objectifs :

- la création de corpus monolingues et de corpus parallèles, à partir de la traduction de textes de genres diversifiés ;
- le développement de corpus annotés selon le cadre des *Universal Dependencies* [3] ;
- la production de descriptions linguistiques complètes et actualisées basées sur ces corpus et utilisées pour leur annotation ;
- la sensibilisation de la communauté du Traitement Automatique des Langues (TAL) aux défis des langues non standardisées et à l'importance de la variation linguistique dans les systèmes de TAL ;
- le partage et le transfert des expériences et outils entre les langues concernées.

2 Les premiers travaux du projet ont concerné la collecte d'un corpus parallèle (Stosic *et al.*, 2024) et la documentation des ressources collectées par des métadonnées à grain fin (Vergez-Couret *et al.*, 2024). Ces travaux ont mis en évidence deux enjeux sociolinguistiques majeurs dont il sera question dans cette contribution, en lien avec le développement actuel d'outils de traitement automatique pour les langues minorisées (section 1). Le premier concerne la caractérisation des langues lors de la documentation des ressources (section 2). Le second concerne la création de nouvelles ressources par traduction et la manière dont elles peuvent différer des pratiques des locuteurs des quatre langues du projet (section 3).

1. Méthodes actuelles pour le traitement automatique des langues minorisées

3 L'histoire du traitement automatique des langues est traversée de changements de paradigmes, au gré des évolutions techniques et algorithmiques. Dernière en date, l'arrivée des réseaux de neurones profonds et de l'architecture dite *transformer* (Vaswani *et al.*, 2017) a permis un bond qualitatif certain pour de nombreuses applications : traduction automatique, dialogue humain-machine, reconnaissance et génération de la parole. Ces derniers modèles[4] sont capables de prendre en compte des relations de dépendance linguistique entre mots distants, par exemple des contraintes d'accord. Ils reposent également sur un découpage en unités (la *tokénisation*) plus petites que les mots qui les rend mieux à même de gérer les mots nouveaux (appelés souvent « mots hors-vocabulaire ») qui étaient absents des données utilisées pour pré-entraîner ces modèles. Ce découpage rend ainsi les modèles moins sensibles à la variation : tout mot peut être reconstitué à partir d'une séquence de sous-mots connus du modèle (voire de caractères, dans le cas le plus extrême). Alors qu'auparavant les problématiques de normalisation se posaient de manière aiguë pour les données non standards (erreurs d'orthographe, langues non normées à l'écrit), elles peuvent sembler moins cruciales aujourd'hui au moins pour les tâches d'analyse (annotation automatique de textes). Elles restent toutefois centrales pour les tâches de génération de textes dans une langue minorisée comme la traduction automatique (voir section 2).

4 Les modèles de langues de type *transformer* sont produits par pré-entraînement sur des données de la ou les langues cibles pour ensuite être spécialisés sur des tâches particulières. Les quantités de données nécessaires sont généralement importantes et font partie des critères définitoires des grands modèles de langues : ainsi Rogers et Luccioni (2024) fixent un seuil de 1 milliard de tokens pour l'anglais. Cette nécessité est bien sûr un frein pour le développement de tels modèles pour les langues minorisées qui sont aussi souvent « peu dotées », dans la mesure où elles disposent de peu de ressources

linguistiques numériques, même si des travaux ont montré qu'il est possible d'avoir des résultats de bonne qualité avec 55 Mo (mégaoctets) de données pour l'ancien français, soit 10,5 millions de mots (Grobol *et al.*, 2022), tandis que Micheli et al. (2020) recommandent un corpus d'au moins 100 Mo. Ces indications de taille restent toutefois supérieures aux données disponibles pour des langues comme le corse : Millour *et al.* (2024) utilisent un corpus maximal d'environ 2,7 millions de mots.

5 Ainsi, pour réduire la spécialisation linguistique de modèles pré-entraînés sur des données d'une seule langue et accéder à une forme de compétence linguistique plus générique, des modèles multilingues ont vu le jour, avec une tendance à l'augmentation du nombre de langues données en entrée lors du pré-entraînement. On citera notamment les modèles *encodeurs* suivants : mBERT (104 langues ; Devlin *et al.*, 2019), XLM-R (100 langues ; Conneau *et al.*, 2020) ou Glot500 (511 langues ; Imani *et al.*, 2023).

6 Des travaux s'attachent à évaluer la capacité de tels modèles à être utilisés pour traiter de nouvelles langues, absentes des données de pré-entraînement, en exploitant la proximité linguistique entre langues. Ces travaux montrent que certains critères semblent influer sur les résultats qu'il est possible d'obtenir.

7 Ainsi, De Vries *et al.* (2022) ont étudié l'apprentissage par transfert interlinguistique à partir de modèles multilingues pré-entraînés pour la tâche d'étiquetage morphosyntaxique avec XLM-R utilisé comme modèle multilingue pré-entraîné. Leurs expériences montrent que l'inclusion de la langue cible (langue étiquetée) – et, dans une moindre mesure, de la langue source (langue dont les données sont utilisées pour l'entraînement de l'étiquetage) – dans l'ensemble de données de pré-entraînement pour le modèle multilingue revêt une importance particulière. Le fait d'appartenir à la même famille linguistique a également un effet sur l'exactitude, de même que le partage des systèmes d'écriture. D'autres travaux s'attachent à augmenter de manière artificielle la proximité linguistique de surface entre langues, de manière à améliorer les performances du modèle qui sera plus robuste face aux variantes orthographiques. Ainsi, Aepli et Sennrich (2022) et Blaschke *et al.* (2023) montrent qu'il est possible de simuler l'absence de norme orthographique à l'écrit pour des

langues peu dotées en injectant aléatoirement du bruit dans les données disponibles pour la langue mieux dotée utilisée pour l'affinage du modèle pré-entraîné.

8 Nous avons pu vérifier et exploiter les résultats de ces études en menant des expériences spécifiques dans le cadre du projet DIVITAL. Millour *et al.* (2024) font ainsi une étude détaillée de l'impact de l'utilisation de ressources en italien pour l'étiquetage morphosyntaxique du corse. Les expériences montrent que les meilleures performances sont atteintes lorsque l'on dispose de corpus bruts et annotés pour la langue cible (corse), mais que ces performances sont améliorées lorsqu'on y adjoint des ressources pour une langue proche (italien). Holgado et Vergez-Couret (2024) évaluent quant à elles l'effet de l'augmentation des données à l'aide d'un lexique bilingue poitevin-saintongeais/français pour faire de l'étiquetage morphosyntaxique. Là aussi, l'exploitation de ressources d'une langue proche, le français, est bénéfique. Enfin, Bernhard (2023) a montré que l'utilisation de lexiques bilingues alsacien-allemand, en particulier pour les classes fermées, permet de transformer les corpus à annoter en alsacien de manière à les rapprocher de l'allemand et ainsi augmenter les performances de modèles entraînés pour d'autres langues, sans nécessiter de ré-entraînement.

9 Dans tous les cas, l'exploitation de ressources d'une langue proche améliore les résultats, mais ne permet pas encore d'atteindre les niveaux de performance observés pour des langues bien dotées. La situation reste donc encore très inégalitaire entre les langues et l'absence de ressources linguistiques numériques reste un frein majeur au développement d'outils pour les langues minorisées. Au-delà de cela, Kreutzer *et al.* (2022) ont montré que les corpus multilingues collectés automatiquement à partir du web et régulièrement utilisés pour le pré-entraînement des grands modèles de langues avaient des niveaux de qualité très variables, notamment pour ce qui est de la caractérisation linguistique fine des données collectées (identification des langues). Dans la section suivante, nous allons revenir plus en détail sur l'importance et l'impact de cette caractérisation.

2. Caractérisation linguistique des langues minorisées et impact sur le TAL

10 Que ce soit en TAL ou en linguistique de corpus, les données langagières sont documentées à l'aide de métadonnées de divers types, avec un niveau de détail variable (Vergez-Couret *et al.*, 2024). Cela étant, la métadonnée fondamentale, qui permet de documenter *a minima* une ressource, est la dénomination de la langue représentée. De manière à s'abstraire de dénominations variables pour les langues et des cas d'homonymie, des codes normés ont vu le jour et notamment ISO 639-3 qui couvre les langues individuelles (langues vivantes, langues mortes, langues anciennes) et leur attribue un indicatif unique à trois lettres. ISO-639-3 est la plus étendue des quatre séries d'indicatifs d'ISO 639 et est utilisée très fréquemment pour documenter les corpus linguistiques. Cette utilisation a des impacts à divers niveaux :

- inventaire des langues et de leurs caractéristiques. Le développement d'outils de TAL nécessite d'avoir une connaissance des langues existantes et de certaines informations s'y rapportant : nombre de locuteurs, système d'écriture, pays ou régions dans lesquels une langue est parlée, ou encore état de vulnérabilité de la langue (van Esch *et al.*, 2022 ; Kargaran *et al.*, 2024 ; Ritchie *et al.*, 2024). D'une manière générale, les codes ISO 639-3 sont centraux dans ces inventaires et les langues qui ne bénéficient pas d'un tel code en sont donc souvent exclues. Le tableau 1 ci-dessous donne un aperçu de la présence des langues du projet DIVITAL dans deux inventaires récents, LinguaMeta (Ritchie *et al.*, 2024) et GlotScript (Kargaran *et al.*, 2024) ;
- identification automatique des langues dans les corpus de très grande taille. Comme nous l'avons expliqué dans la section 1, les modèles de TAL actuels nécessitent des corpus de très grande taille. Par conséquent, ces corpus sont collectés de manière automatisée sur le web : l'enjeu est alors de déterminer automatiquement la langue des documents collectés, afin de pouvoir filtrer les documents en fonction des langues cibles. Comme le montrent Kargaran *et al.* (2023), l'identification automatique de la langue incluant des langues peu dotées fait face à de

nombreux défis, dont la distinction entre des langues proches ou encore la coexistence d'indicatifs pour des macro-langues et des variétés particulières [5]. L'outil proposé, GlotLID, identifie les langues à l'aide de leur code ISO 639-3 associé au système d'écriture. La version 3 de l'outil gère 2102 langues [6]. Le tableau 1 indique la présence ou non des langues du projet DIVITAL dans la liste des langues gérées par GlotLID ;

- documentation et organisation des ressources. La documentation des ressources linguistiques par leur langue est nécessaire dans les travaux de recensement des ressources qui permettent de comparer la situation des langues et de dégager des axes prioritaires de développement. Ainsi, Giagkou *et al.* (2022) présentent un travail d'agrégation de méta-données pour lequel une standardisation des noms des langues a dû être réalisée. Dans le cas le plus général, l'identification se fait à l'aide du code ISO 639 auquel sont ajoutés des indicateurs pour la région, le système d'écriture et les variantes. Pour les langues et variétés absentes d'ISO 639, le code Glottolog (Hammarström *et al.*, 2023) est utilisé, ainsi que des dénominations non normées pour des variétés spécifiques. Dans le projet *Universal Dependencies* (De Marneffe *et al.*, 2021), les corpus annotés sont organisés par langue, selon les indicatifs ISO 639-3. Cela pose le problème des langues et variétés non couvertes par ISO 639-3, comme le poitevin-saintongeais qui se verrait classé parmi les autres corpus en français, ou encore les dialectes alsaciens qui seraient rangés sous le nom « suisse allemand » (*Swiss German*) alors même que la dénomination est impropre eu égard à la localisation des locuteurs. Une dénomination comme *Central Alemannic* telle qu'utilisée dans Glottolog serait moins problématique de ce point de vue. Autre conséquence : actuellement, l'absence de code ISO 639-3 [7] pour le poitevin-saintongeais invisibilise totalement cette langue et ses ressources linguistiques, qui ne figurent donc pas dans les inventaires mentionnés *supra*, et complique également le développement potentiel d'une Wikipédia (voir section 1), qui requiert l'existence préalable d'un code ISO 639 ou BCP 47 [8].

Tableau 1 : Présence des langues du projet DIVITAL dans divers inventaires

	dialectes alsaciens	**corse**	**occitan**	**poitevin-saintongeais**

ISO 639-3	gsw (*Swiss German - Alemannic - Alsatian*) pour les variétés alémaniques, pfl (*Pfaelzisch*) qui, suivant Glottolog, englobe le francique rhénan lorrain (*Lothringisch*) et le *Südrheinfränkisch/ Südfränkisch* parlé au nord de l'Alsace (francique rhénan méridional)	cos (Corsican)	oci (Occitan post 1500)	⌀
LinguaMeta	gsw, pfl	co (code bcp_47)	oc (code bcp_47)	⌀
GlotScript	gsw, pfl	cos	oci	⌀
GlotLID	gsw_Latn, pfl_Latn	cos_Latn	oci_Latn	⌀

11 Le tableau 1 montre bien la situation contrastée des langues du projet DIVITAL. S'il n'y a pas de problème d'indicatif pour le corse ou l'occitan, le poitevin-saintongeais est totalement absent. Le terme « alsacien » (*Alsatian*) figure dans le nom associé à l'indicatif gsw, mais ce terme englobe dans l'usage des parlers alémaniques et franciques, dont le francique rhénan lorrain et le francique rhénan méridional, qui sont exclus de l'indicatif gsw. Dans le cas du poitevin-saintongeais, l'absence d'indicatif à l'heure actuelle rend la langue inexistante dans l'espace numérique comme vu précédemment et implique, pour le moment, d'utiliser le code de la langue standard la plus proche, à savoir le français, ce qui porte à confusion car ce sont bien deux langues différentes, le premier écrit signalé en poitevin datant de la fin du XVII^{e} siècle. Les dialectes alsaciens représentent le cas d'une simplification de la situation linguistique, où une dénomination pouvant être considérée comme une « macro-langue », l'alsacien, recouvre en réalité les variétés alémaniques et franciques du territoire, associées à deux indicatifs différents. Or, seules les variétés alémaniques, qui dominent le territoire, sont associées à la dénomination « alsacien » via l'indicatif gsw. Ces observations confortent les deux idées (ou idéologies) largement à l'œuvre dans les technologies des langues et mises en évidence par Markl *et al.* (2024) : la première selon laquelle les langues peuvent et doivent être « standardisées » (*languages can and should be « standardised »*) et la seconde selon laquelle les langues sont des objets clairement délimités (*languages are clearly delineated objects*). La première implique qu'une variété particulière est mise en avant, ce que l'on peut voir à l'œuvre dans le choix primaire de la variété « suisse allemand » qui donne son nom à l'indicatif gsw et à laquelle l'alsacien (ou en tout cas les parlers dialectaux alémaniques parlés en Alsace) ont été rattachés

par la suite. La seconde est sous-jacente aux codifications elles-mêmes qui conduisent à une différenciation stricte en fonction de l'indicatif utilisé.

12 Comme nous venons de le voir avec ces divers exemples, la caractérisation linguistique des langues minorisées est loin d'être un problème résolu. Au-delà des séries d'indicatifs ISO 639, d'autres registres de codes existent comme ROLV[9] qui propose des identifiants pour l'alsacien, le poitevin, le saintongeais, les dialectes de l'occitan ou du corse ou encore le système IETF BCP 47. Mais ISO 639 reste central en TAL et en linguistique de corpus, avec pour conséquence d'invisibiliser certaines langues ou de simplifier le paysage linguistique de manière erronée. D'une manière générale, la prise en compte de la variation linguistique est assez complexe en TAL, en particulier pour les tâches qui impliquent de générer du texte. Ceci peut conduire à une prise en charge incomplète de la variation et au choix, explicite ou implicite, de variétés particulières. Ainsi, le traducteur automatique pour l'occitan développé par le Congrès permanent de la langue occitane, Revirada[10], propose de traduire au choix vers l'occitan gascon ou languedocien, mettant ainsi en évidence la variation dialectale, mais en ne couvrant pas l'ensemble des dialectes de l'occitan (approche lacunaire). À l'inverse, Google Translate[11] n'explicite jamais le dialecte particulier utilisé, et se contente de la dénomination « occitan », alors même que le choix a clairement été fait de privilégier le languedocien (approche opaque). Si l'initiative de proposer des langues minorisées dans un outil grand public comme Google Translate est très certainement louable, l'outil produit fait le choix de privilégier une variété, au risque de tromper les utilisatrices et utilisateurs et, à terme, de voir cette variété prendre le pas sur les autres par la multiplication des contenus traduits automatiquement. On peut ainsi parler d'une forme de normalisation artificielle, avec le choix d'un standard par les développeurs d'applications, non nécessairement locuteurs, qui ne fait pourtant pas consensus dans la communauté linguistique concernée[12]. Le choix de la graphie est lui aussi implicite, mettant généralement en avant une graphie dite « normalisée » tout en ignorant le fait que certaines de ces langues peuvent être écrites selon divers systèmes graphiques en cohabitation. Globalement, le manque d'information et de documentation au sujet des données utilisées pour produire ces outils les rend opaques,

en particulier pour les locutrices et locuteurs qui pourraient ne pas se reconnaître dans la variété sélectionnée de manière implicite.

13 La question des variétés représentées se pose également lors de la collecte et la création de ressources linguistiques pour les langues minorisées. Dans la section qui suit nous allons plus particulièrement discuter le choix opéré dans le projet DIVITAL, qui a consisté à traduire des textes de manière à obtenir un corpus disponible en cinq langues : alsacien, corse, occitan, poitevin-saintongeais et français. La constitution d'un tel corpus de traduction ne tendra pas nécessairement vers une standardisation déjà existante ou en cours de réflexion, l'objectif étant de tenir compte de la variation existante à ce jour.

3. Création de nouvelles ressources par traduction

14 Il existe un effort continu de constitution de ressources pour les langues du projet, l'objectif étant avant tout de rassembler des ressources pour la description en linguistique mais aussi pour les recherches en littérature, histoire, ethnologie, etc. On peut citer le corpus MeThAL pour une macro-analyse du théâtre alsacien (Ruiz Fabo *et al.*, 2021), TELPOS pour une base de textes électroniques en poitevin-saintongeais (Dourdet *et al.*, 2019), BaTelòc pour une base de texte en langue occitane (Bras et Vergez-Couret, 2016) ou les corpus de la Banque de données en langue corse (Kevers et Retali-Medori, 2020). Ces ressources ont également été constituées dans le respect des standards informatiques permettant leur exploitation en traitement automatique des langues (Bernhard *et al.*, 2019). Dans le cadre du projet DIVITAL, nous avons souhaité enrichir les ressources existantes avec des ressources d'un nouveau type. Nous avons constitué un corpus à partir de textes du français moderne à contemporain, qui est une langue dotée et connue de tous les locuteurs participants au projet. Cette langue source actuelle nous a permis de sélectionner des textes littéraires et non littéraires/non narratifs (textes juridiques et argumentatifs) portant notamment sur des thématiques choisies et contemporaines comme l'environnement, les droits humains, etc. À partir de ce corpus de textes en français langue source, nous avons constitué par traduction de ces derniers

un corpus parallèle dans les quatre langues du projet, dont les contenus sont détaillés dans le tableau 2 (le nombre de tokens est donné pour la partie en français), reprise de Stosic *et al.* (2024). Ainsi, le corpus complet pour les cinq langues (alsacien, corse, français, occitan et poitevin-saintongeais) contient environ 100 000 tokens.

Tableau 2 : Contenu du corpus parallèle DIVITAL

Titre	**Auteur**	**Date de publication de l'œuvre originale**	**Date de la version française utilisée pour la traduction**	**Domaine**	**Genre**	**Partie incluse**	**Nombre de tokens**
Le Petit Prince	Antoine de Saint-Exupéry	1943	_	Littéraire	Roman court	Chapitre 1 et 2	1150
L'homme qui plantait des arbres	Jean Giono	1953	_	Littéraire	Conte allégorique	Entier	3800
Lettres de mon moulin	Alphonse Daudet	1869	_	Littéraire	Nouvelle	2 nouvelles	4100
Contes du lundi	Alphonse Daudet	1873	_	Littéraire	Nouvelle	2 nouvelles	3300
Déclaration universelle des droits de l'homme	Inconnu	1948	_	Légal	Charte officielle	Entier	2070
Décaméron	Boccace	1349-1353	1884	Littéraire	Nouvelle	1 nouvelle	300
Pierre et le loup	Sergueï Prokofiev	1936	_	Littéraire	Conte symphonique	Entier	780
Le fils prodigue	Luc	IIe siècle	1831, 1879	Religion	Parabole	Entier	510
La bise et le soleil	Esope	VIe siècle av. J.-C.	2018	Littéraire	Fable	Entier	130
Chroniques sur les langues régionales de France	Michel Feltin-Palas	2021-2023	-	Journalisme	Chronique	4 chroniques	4050
TOTAL							**20190**

Cette démarche est particulièrement intéressante à plusieurs niveaux :

- cela a permis que tous les textes des langues concernées puissent être mis en regard avec une langue largement dotée ;
- cela a également permis d'enrichir les données existantes dans les langues de traduction (occitan, corse, poitevin-saintongeais et alsacien) en couvrant de nouveaux genres discursifs non narratifs. C'est une forme d'ouverture vers d'autres pratiques linguistiques et, en ce sens, la

traduction vers ces langues minorisées est un outil important de revitalisation. Ndjitat Tatchou (2016 : 27) utilise le terme écotraduction : « Si la traduction est utilisée en soutien à la préservation de cette glottodiversité, à la survie des langues, à l'entretien de ce que Calvet (1999) appelle des niches écolinguistiques, il s'agira alors d'une traduction au service de l'écologie des langues : nous la baptisons écotraduction » ;

- l'intérêt va au-delà du projet immédiat, dans la mesure où ces ressources parallèles pourront être exploitées pour analyser les différences syntaxiques entre les langues qui sont en situation de contact avec le français sur le modèle du corpus « Parallel Universal Dependencies » (PUD) (Nikolaev *et al*, 2020).

15 La traduction en langue minorisée implique finalement, comme pour nos prédécesseurs aux XV^e^ et XVI^e^ siècles qui ont eu à traduire des textes antiques (en grec et latin) vers une langue vernaculaire non normée, deux grandes missions : enrichir la littérature et augmenter les potentialités de la langue (Mencé-Caster, 2023). Néanmoins, notre démarche se distingue par sa nature scientifique et par son respect de la diversité dialectale et graphique des langues de notre projet.

16 Cette démarche de traduction pour la constitution de nouvelles données dans le cadre d'une activité scientifique n'est pas nouvelle : elle remonte au XIX^e^ siècle avec les traductions de la *Parabole de l'enfant prodigue* en langues locales (Coquebert de Montbret et Labouderie, 1831 ; Favre, 1879) ou plus récemment celles de la fable d'Ésope, *La bise et le soleil* (Boula de Mareüil *et al.*, 2017). Pour autant, les rapports entre traduction et documentation sont peu explicités comme le signalent Brunel (2022) et Cristinoi (2022). Cristinoi (2022) distingue deux types de traduction, le premier étant une traduction brute, littérale, proche du texte source (Woodbury, 2007) et le deuxième étant une traduction anthropologique ou ethnologique dans laquelle le texte produit prend les libertés nécessaires pour représenter la culture de la langue cible. Les textes produits ici relèvent du premier type. Cette démarche de traduction présente certains biais qui ne sont toutefois pas rédhibitoires au regard de l'apport qu'elle offre, mais que nous soulignerons ci-après pour mieux exploiter le corpus produit.

17 L'exercice actuel de traduction consiste à la fois à produire une traduction aussi fidèle que possible sur le plan morphosyntactico-sémantique et informationnel du texte source, et la plus idiomatisée

possible dans la langue cible de traduction. Il s'agit indéniablement d'un savant équilibre à trouver. En effet, il faut considérer que la langue source peut avoir une certaine influence sur la langue de traduction (lexique, morphologie, syntaxe, idiomatismes, etc.) et c'est notamment pourquoi Mounin (1963) considère la traduction comme un contact de langues.

18 En plus des biais évoqués ci-dessus, nous avons été confrontés à des problématiques directement liées au statut de langues minorisées. En l'absence de standardisation, traduire implique nécessairement de respecter la variété dialectale ou graphique. Mais il existe aussi des contraintes externes liées aux traducteurs eux-mêmes. Dans certains cas, le faible nombre de traducteurs professionnels disponibles a orienté la sélection des textes en évitant les textes dont les fonctions seraient majoritairement métalinguistiques ou poétiques.

19 Concernant la graphie, certaines langues minorisées ont déjà un système graphique qui peut être plus ou moins normé mais pour certaines d'entre elles, la mise en écriture n'existe pas ou est balbutiante ; parfois encore, elle existe mais les discussions liées à une homogénéisation sont encore vives. Ainsi, notamment, la variation touche-t-elle aussi les graphies qui peuvent être débattues. Or, dès lors que l'on tente d'accroître les ressources pour une langue, cette question entre dans le débat : garde-t-on la variation (modèle riche) ou la lisse-t-on (modèle en apparence stable mais moins représentatif de la diversité interne de la langue) ?

20 Les biais proviennent également des traducteurs et de leur connaissance intégrée de la langue minorisée, dite grammaire intériorisée qui est empreinte d'une plus ou moins grande variabilité diastratique, diamésique ou diatopique et peut-être même diachronique sur un faible empan (Schøsler, 2020) dans le cas des langues potentiellement non ou peu normées. Elle peut être due aussi au statut du traducteur (en lien avec le degré de minorisation de la langue de traduction) : professionnel-locuteur (formé selon les critères actuels) ou non-professionnel-locuteur de la langue minorisée. Pour ces derniers, dans un souci de « bien » faire/dire intégré à l'école laïque du XX^{e} siècle [13], ils auront peut-être, par exemple, du mal à se détacher de la consigne pour intégrer un idiomatisme dans leur traduction, par

souci de rester syntactico-sémantiquement proche du texte source, ce qui peut créer quelques biais de traduction.

21 Pour le projet, les biais ont été compensés au mieux par certains choix :

- le premier objectif a été de suivre la méthode actuelle de traductologie qui consiste à rester le plus fidèle possible au texte source (syntaxiquement et sémantiquement) tout en produisant dans la langue cible un texte représentatif de ses usages ;
- les traductions obtenues pour le corpus ont été faites par des traducteurs locuteurs ou par des traducteurs professionnels pour chaque langue sélectionnée. C'est en particulier le critère de locuteur qui a l'emporté sur le statut professionnel ou non du traducteur pour les traductions en poitevin-saintongeais. Les traducteurs de statut non-professionnel-locuteur de cette langue ont reçu des instructions visant à homogénéiser leur production au regard de celle de leurs homologues professionnels. Leurs traductions ont parfois été faites au sein de groupes sous la supervision d'un enseignant-chercheur locuteur formé à cet exercice et ne sont dès lors pas le fruit d'un seul traducteur ;
- pour les biais liés à l'influence potentielle de la langue source sur la langue de traduction, il est difficile de les éviter totalement et toute traduction a une part inévitable d'influences de la langue source sur la langue de traduction mais aussi de la grammaire intégrée du locuteur de la langue de traduction. Une formation académique peut lisser certains phénomènes mais certains restent inévitables. Une traduction en tant que texte n'est jamais un calque. Cet écart est donc à envisager comme acceptable dès lors que les critères de proximité syntactico-sémantique sont maintenus tout comme la possibilité d'injecter dans la traduction des idiomatismes adaptés ;
- pour le dialecte et la graphie (traits morphosyntaxiques et lexicaux du dialecte et représentations graphiques), nous avons conservé les choix proposés dans les différentes traductions existantes retenues ou par les traducteurs produisant ; nous avons fait de même pour les choix idiomatiques sans tenter de les uniformiser. Dans ce cas, si les traducteurs ne sont pas contraints par des instructions méthodologiques déjà existantes, leurs choix sont en revanche systématiquement consignés dans les métadonnées associées aux traductions.

22 Il est à noter de grands apports liés à ce travail de traduction pour les types de discours non narratifs sur des thématiques contemporaines. Il a requis, de la part des traductrices et traducteurs, un travail particulier sur la création lexicale comme le souligne Cristinoi (2022). Ainsi, cette exploration de nouveaux genres textuels, qui sont actuellement pas ou peu représentés dans le répertoire existant, constitue une véritable intervention sur le corpus de la langue, qui consiste non plus seulement à collecter des textes dans une visée de conservation à tendance patrimonialisante, mais également à participer à la création de nouvelles pratiques dans ces langues en proposant de nouvelles productions.

23 Pour enrichir encore davantage le corpus, il pourrait être envisagé de comparer les textes produits par différents traducteurs (graphiquement, morphologiquement, syntaxiquement, lexicalement) pour souligner une plus ou moins grande variabilité dans la langue de traduction si pour un texte source plusieurs traductions de plusieurs traducteurs sont disponibles dans la même langue minorisée. Mais, par manque de moyens et de forces vives disponibles, nous n'avons pu faire ce type de comparaison.

Conclusion

24 En guise de conclusion, nous pouvons formuler le constat qu'en plus de conditions défavorables auxquelles doivent faire face les typologues, les linguistes de terrain et, de manière générale, toutes les personnes qui travaillent à l'étude et à la préservation de la diversité des langues (manque de ressources humaines et financières – conditions elles-mêmes liées à la minorisation des langues étudiées), les travaux doivent également considérer l'absence de normes, essentiellement liée au degré de variation interne (notamment diatopique et graphique) plus important dans ces langues que celle des langues mieux dotées, ainsi que l'attachement des locuteurs à cette variation. Face à cette double particularité, les langues minorisées sont rarement prises en compte pour le développement d'applications et de ressources numériques, avec pour conséquences l'invisibilisation de certaines langues et la simplification du paysage linguistique.

25 Cependant, le travail mené dans le cadre du projet DIVITAL sur la création de ressources nouvelles par la traduction ainsi que l'explora-

tion de nouveaux genres textuels, qui sont actuellement peu ou pas représentés dans le répertoire existant, permettent d'enrichir ces langues en même temps qu'elles sont étudiées. En intervenant ainsi directement sur le corpus des langues minorisées et en créant de nouvelles ressources pour ces dernières, nous nous éloignons de la simple collecte et de la vision patrimonialisante qui y est généralement associée pour nous rapprocher d'une démarche glottopolitique (Guespin et Marcellesi, 1986) dans laquelle les chercheurs deviennent eux-mêmes acteurs de la visibilisation, voire de la vitalité, des langues minorisées.

26 En dernière analyse, le numérique apparaît comme un enjeu important pour créer de nouvelles formes ou donner de nouvelles fonctions aux langues minorisées, dans le but d'accroître leurs usages et leurs utilisateurs (King, 2001). Mais, comme nous avons pu le voir, pour répondre à ce défi, il reste encore du chemin à parcourir pour faire reconnaître ces langues dans cet espace.

BIBLIOGRAPHIE

AEPLI Noëmi et SENNRICH Rico, 2022, « Improving Zero-Shot Cross-lingual Transfer Between Closely Related Languages by Injecting Character-Level Noise », *Findings of the Association for Computational Linguistics: ACL 2022*, Dublin, Ireland, Association for Computational Linguistics, p. 4074-4083.

BERNHARD Delphine, 2023, « Transfert zero-shot pour l'étiquetage morphosyntaxique : analyse de l'impact de la transformation des données à étiqueter pour les dialectes alsaciens », dans Karën Fort, Claire Gardent et Yannick Parmentier (dir.), *Actes des 5èmes journées du Groupement de Recherche CNRS « Linguistique Informatique, Formelle et de Terrain »*, Nancy, France, p. 30-38.

BERNHARD Delphine, BRAS Myriam, ERHART Pascale, LIGOZAT Anne-Laure et VERGEZ-COURET Marianne, 2019, « Language Technologies for Regional Languages of France: The RESTAURE Project », *International Conference Language Technologies for All (LT4All): Enabling Linguistic Diversity and Multilingualism Worldwide*, Paris, European Language Resources Association (ELRA), p. 272-275.

BLANCHET Philippe, 2000, *La linguistique de terrain. Méthode et théorie. Une approche éthno-sociolinguistique*, Rennes, Presses universitaires de Rennes.

BLASCHKE Verena, SCHÜTZE Hinrich et PLANK Barbara, 2023, « Does Manipulating Tokenization Aid Cross-

Lingual Transfer? A Study on POS Tagging for Non-Standardized Languages », *Tenth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2023)*, Dubrovnik, p. 40-54.

BOULA DE MAREÜIL Philippe, VERNIER Frédéric et RILLIARD Albert, 2017, « Enregistrements et transcriptions pour un atlas sonore des langues régionales de France », *Géolinguistique*, vol. 17, p. 23-48.

BOYER Henri, 2013, « L'impact de l'unilinguisme sur la normativisation de la langue française » dans Kremnitz Georg (dir.), *Histoire sociale des langues de France*, p. 183-188.

BRAS Myriam et VERGEZ-COURET Marianne, 2016, « BaTelÒc: A text base for the Occitan language », *Language Documentation and Conservation in Europe*, vol. Special Publication nº 9, p. 133-149.

BRUNEL Noëlle, 2022, « Édito », *Traduire*, vol. 247, p. 3-5.

CALVET Louis-Jean, 1999, *Pour une écologie des langues du monde*, Plon.

CONNEAU Alexis, KHANDELWAL Kartikay, GOYAL Naman, CHAUDHARY Vishrav, WENZEK Guillaume, GUZMÁN Francisco, GRAVE Edouard, OTT Myle, ZETTLEMOYER Luke et STOYANOV Veselin, 2020, « Unsupervised Cross-lingual Representation Learning at Scale », *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, p. 8440-8451, [https://doi.org/10.48550/arXiv.1911.02116].

COQUEBERT DE MONTBRET Eugène et LABOUDERIE Jean de (dir.), 1831, *Mélanges sur les langues, dialectes et patois : renfermant, entre autres, une collection de versions de la Parabole de l'enfant prodigue en cent idiomes ou patois différens*, Paris, Almanach du commerce : Delaunay.

CRISTINOI Antonia, 2022, « La traduction dans la documentation des langues », *Traduire*, vol. 247, p. 6-15.

DE MARNEFFE Marie-Catherine, MANNING Christopher D., NIVRE Joakim et ZEMAN Daniel, 2021, « Universal dependencies », dans *Computational linguistics*, vol. 47, nº 2, p. 255-308.

DEVLIN Jacob, CHANG Ming-Wei, LEE Kenton et TOUTANOVA Kristina, 2019, « BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding », *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, (volume 1 : Long and Short Papers), Minneapolis, Minnesota, Association for Computational Linguistics, p. 4171-4186.

DOURDET Jean-Christophe, VERGEZ-COURET Marianne et LAY Marie-Hélène, 2019, « Telpos - Texte électronique en poitevin-saintongeais, enjeux et difficultés », *Colloque « Langues minoritaires » : quels acteurs pour quel avenir ?*, Strasbourg, [https://hal.science/hal-02892750].

ESCH Daan van, LUCASSEN Tamar, RUDER Sebastian, CASWELL Isaac et RIVERA Clara, 2022, « Writing system and speaker metadata for 2,800+ language varieties », *Proceedings of the*

language resources and evaluation conference, Marseille, European Language Resources Association, p. 5035-5046.

FAVRE Léopold, 1879, *Parabole de l'enfant prodigue en divers dialectes ; patois de la France*, avec une introduction sur la formation des dialectes et patois de la France, par L. Favre, Niort, L. Favre.

GIAGKOU Maria, PIPERIDIS Stelios, LABROPOULOU Penny, DELIGIANNIS Miltos, KOLOVOU Athanasia et VOUKOUTIS Leon, 2022, « Collaborative Metadata Aggregation and Curation in Support of Digital Language Equality Monitoring », *Proceedings of the Workshop Towards Digital Language Equality within the 13th Language Resources and Evaluation Conference*, Marseille, France, European Language Resources Association, p. 27-35.

GROBOL Loïc, REGNAULT Mathilde, ORTIZ SUAREZ Pedro, SAGOT Benoît, ROMARY Laurent et CRABBÉ Benoit, 2022, « BERTrade: Using Contextual Embeddings to Parse Old French », *Proceedings of the language resources and evaluation conference*, Marseille, France, European Language Resources Association, p. 1104-1113.

GUESPIN Louis et MARCELLESI Jean-Baptiste, 1986, « Pour la glottopolitique » *Langages n°* 83, *Glottopolitique, sous la direction de Jean-Baptiste Marcellesi*, p. 5-34.

HAMMARSTRÖM Harald, FORKEL Robert, HASPELMATH Martin et BANK Sebastian, 2023, « Glottolog 4.8 », [https://glottolog.org/glottolog/language].

HOLGADO Cristina Garcia et VERGEZ-COURET Marianne, 2024, « Empowering Low-Resource Regional Languages with Lexicons : A Comparative Study of NLP Tools for Morphosyntactic Analysis », *The 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, Turin, Association for Computational Linguistics, p. 5747-5756.

IMANI Ayyoob, LIN Peiqin, KARGARAN Amir Hossein, SEVERINI Silvia, SABET Masoud Jalili, KASSNER Nora, MA Chunlan, SCHMID Helmut, MARTINS André F. T., YVON François et SCHÜTZE Hinrich, 2023, « Glot500: Scaling Multilingual Corpora and Language Models to 500 Languages », *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics* (volume 1 : Long Papers), Toronto, Canada, Association for Computational Linguistics, p. 1082-1117.

KARGARAN Amir Hossein, IMANI Ayyoob, YVON François et SCHÜTZE Hinrich, 2023, « GlotLID: Language Identification for Low-Resource Languages », *Findings of the Association for Computational Linguistics: EMNLP 2023*, Singapore, Association for Computational Linguistics, p. 6155-6218.

KARGARAN Amir Hossein, YVON François et SCHÜTZE Hinrich, 2024, « GlotScript: A Resource and Tool for Low Resource Writing System Identification », *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation*, Torino, p. 7774-7784.

KEVERS Laurent et RETALI-MEDORI Stella, 2020, « Towards a Corsican Basic Language Resource Kit », dans Calzolari Nicoletta, Béchet Frédéric, Blache Philippe, Choukri Khalid, Cieri Christopher, Declerck Thierry, Goggi Sara, Isahara Hitoshi, Maegaard Bente, Mariani Joseph, Mazo Hélène, Moreno Asuncion, Odijk Jan et Piperidis Stelios (dir.), *Proceedings of the Twelfth Language Resources and Evaluation Conference*, Marseille, European Language Resources Association, p. 2726-2735.

KING Kendall A, 2001, *Language revitalization processes and prospects: Quichua in the Ecuadorian Andes*, vol. 24, Multilingual Matters.

KREUTZER Julia, CASWELL Isaac, WANG Lisa, WAHAB Ahsan, VAN ESCH Daan, ULZII-ORSHIKH Nasanbayar, TAPO Allahsera, SUBRAMANI Nishant, SOKOLOV Artem, SIKASOTE Claytone, SETYAWAN Monang, SARIN Supheakmungkol, SAMB Sokhar, SAGOT Benoît, RIVERA Clara, RIOS Annette, PAPADIMITRIOU Isabel, OSEI Salomey, SUAREZ Pedro Ortiz, ORIFE Iroro, OGUEJI Kelechi, RUBUNGO Andre Niyongabo, NGUYEN Toan Q., MÜLLER Mathias, MÜLLER André, MUHAMMAD Shamsuddeen Hassan, MUHAMMAD Nanda, MNYAKENI Ayanda, MIRZAKHALOV Jamshidbek, MATANGIRA Tapiwanashe, LEONG Colin, LAWSON Nze, KUDUGUNTA Sneha, JERNITE Yacine, JENNY Mathias, FIRAT Orhan, DOSSOU Bonaventure F. P., DLAMINI Sakhile, DE SILVA Nisansa, BALLI Sakine Çabuk, BIDERMAN Stella, BATTISTI Alessia, BARUWA Ahmed, BAPNA Ankur, BALJEKAR Pallavi, AZIME Israel Abebe, AWOKOYA Ayodele, ATAMAN Duygu, AHIA Orevaoghene, AHIA Oghenefego, AGRAWAL Sweta et ADEYEMI Mofetoluwa, 2022, « Quality at a Glance: An Audit of Web-Crawled Multilingual Datasets », dans *Transactions of the Association for Computational Linguistics*, vol. 10, p. 50-72.

MARKL Nina, HALL-LEW Lauren et LAI Catherine, 2024, « Language Technologies as If People Mattered: Centering Communities in Language Technology Development », dans Calzolari Nicoletta, Kan Min-Yen, Hoste Veronique, Lenci Alessandro, Sakti Sakriani et Xue Nianwen (dir.), *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, Torino, Italia, ELRA; ICCL, p. 10085-10099.

MARTINET André, 1969, *Le français sans fard*, Paris, Presses Universitaires de France.

MENCÉ-CASTER Corinne, 2023, « De la traduction médiévale à la traduction de la Renaissance : quelle(s) conceptualisation(s) de la langue et quel(s) enjeu(x) ? », dans *e-Spania*, vol. 46, [https://doi.org/10.4000/e-spania.48526].

MICHELI Vincent, D'HOFFSCHMIDT Martin et FLEURET François, 2020, « On the importance of pre-training data volume for compact language models », *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, p. 7853-7858.

MILLOUR Alice, KEVERS Laurent, BRASILE Lorenza et GHIA Alberto, 2024, « Agettivu, aggitivu o aghjettivu?

POS Tagging Corsican Dialects », *LREC-COLING 2024*, Torino, [https://hal.science/hal-04534608].

MOUNIN George (dir.), 1963, *Les problèmes théoriques de la traduction*, Paris, Gallimard.

NDJITAT TATCHOU Cyrille, 2016, « Écotraduction : une politique de traduction au service de la glottodiversité », *Le linguiste*, vol. 62, n° 4, p. 25-27.

NIKOLAEV Dmitry, ARVIV Ofir, KARIDI Taelin, KENNETH Neta, MITNIK Veronika, SAEBOE Lilja Maria et ABEND Omri, 2020, « Fine-Grained Analysis of Cross-Linguistic Syntactic Divergences », dans Jurafsky Dan, Chai Joyce, Schluter Natalie et Tetreault Joel (dir.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, p. 1159-1176, [https://aclanthology.org/2020.acl-main.109].

RITCHIE Sandy, ESCH Daan van, OKONKWO Uche, VASHISHTH Shikhar et DRUMMOND Emily, 2024, « LinguaMeta: Unified metadata for thousands of languages », dans *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation*, Torino, Italia, ELRA; ICCL, p. 10530-10538.

ROGERS Anna et LUCCIONI Sasha, 2024, « Position: Key Claims in LLM Research Have a Long Tail of Footnotes », dans *Proceedings of the 41st International Conference on Machine Learning, PMLR*, p. 42647-42665.

RUIZ FABO Pablo, WERNER Carole, BERNHARD Delphine, ERHART Pascale et HUCK Dominique, 2021, juin, « MeThAL : Ressources numériques pour une relecture du théâtre en alsacien », [https://doi.org/10.5281/zenodo.4908212].

SCHØSLER Lene, 2020, « Le rapport entre continuité référentielle et expression du sujet envisagé dans une perspective diasystématique », dans Dupuy Estèle, Millogo Victor et Lay Marie-Hélène (dir.), *La continuité référentielle ou le choix des mots*, Frankrig, Presses universitaires de Rennes, p. 123-145.

STOSIC Dejan, MARJANOVIĆ Saša, BERNHARD Delphine, BACH Xavier, BRAS Myriam, KEVERS Laurent, RETALI MEDORI Stella, VERGEZ-COURET Marianne et WERNER Carole, 2024, « The ParCoLab Parallel Corpus and its Extension to Four Regional Languages of France », dans Calzolari Nicoletta, Kan Min-Yen, Hoste Veronique, Lenci Alessandro, Sakti Sakriani et Xue Nianwen (dir.), *LREC-COLING 2024*, Vol. Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC–COLING 2024), Torino, ELRA and ICCL, p. 16014-16023.

TAMBURELLI Marco et TOSCO Mauro, 2021, *Contested languages : The Hidden Multinlingualism in Europe*, Amsterdam/Philadelphie, John Benjamins.

VASWANI Ashish, SHAZEER Noam, PARMAR Niki, USZKOREIT Jakob, JONES Llion, GOMEZ Aidan N, KAISER Łukasz et POLOSUKHIN Illia, 2017, « Attention is All you Need », dans

Advances in Neural Information Processing Systems, vol. 30, Curran Associates, Inc., p. 2-11.

VERGEZ-COURET Marianne, BERNHARD Delphine, NAUGE Michael, BRAS Myriam, RUIZ Pablo et WERNER Carole, 2024, « Managing Fine-grained Metadata for Text Bases in Extremely Low Resource Languages: the Cases of Two Regional Languages of France », dans Melero Maite, Sakti Sakriani et Soria Claudia (dir.), *SIGUL 2024*, Vol. Proceedings of the 3rd Annual Meeting of the Special Interest Group on Under–resourced Languages @ LREC–COLING 2024, Torino, p. 212-221.

VRIES Wietse de, WIELING Martijn et NISSIM Malvina, 2022, « Make the Best of Cross-lingual Transfer: Evidence from POS Tagging with over 100 Languages », *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics* (volume 1: Long Papers), Dublin, Ireland, Association for Computational Linguistics, p. 7676-7685.

WOODBURY Anthony C., 2007, « On thick translation in linguistic documentation », *Language Documentation and Description*, vol. 4, p. 120-135.

NOTES

1 Nous employons ce terme avec le sens que lui donne Blanchet (2000: 131) : « La minoration est qualitative, elle joue sur le statut. La minorisation est quantitative, elle joue sur les pratiques. L'addition de ces deux processus liés conduit le groupe ethno-socioculturel minoré et minorisé à la situation de groupe (de langue) minoritaire ».

2 [https://divital.gitpages.huma-num.fr/fr/], consulté le 27 novembre 2024.

3 [https://universaldependencies.org/], consulté le 27 novembre 2024.

4 Un modèle est une représentation mathématique d'un phénomène particulier et est construit par apprentissage à partir de données d'entraînement.

5 Par exemple, on trouve l'indicatif nor pour la macro-langue « norvégien » qui recouvre deux langues individuelles : le nynorsk nno et le bokmål nob, cf. [https://iso639-3.sil.org/code/nor], consulté le 27 novembre 2024.

6 [https://github.com/cisnlp/GlotLID], consulté le 27 novembre 2024.

7 Nous avons engagé des démarches auprès de l'organisme en charge de la gestion d'ISO 639-3 pour demander la création d'un indicatif pour le poitevin-saintongeais.

8 [https://meta.wikimedia.org/wiki/Language_proposal_policy], consulté le 27 novembre 2024.

9 https://globalrecordings.net/fr/rolv], consulté le 27 novembre 2024.

10 [https://revirada.eu/], consulté le 27 novembre 2024.

11 [https://translate.google.com], consulté le 27 novembre 2024.

12 Il peut être utile de signaler ici que l'occitan n'est pas un cas unique. Au contraire, toutes les langues, même les plus diffusées peuvent souffrir de cette normalisation forcée, avec l'anglais américain au détriment des autres variétés d'anglais, le français dit standard au détriment des autres variétés de français, l'espagnol castillan au détriment des variétés latines, le portugais du Portugal au détriment du portugais parlé au Brésil…

13 « Ce sont les Français eux-mêmes (sous l'influence de leurs grammairiens) qui ont été élevés dans le respect du statu quo normatif, dans la crainte de forger de nouveaux mots, de faire fonctionner la productivité du système » (Martinet, 1969, cité dans Boyer, 2013 : 186).

RÉSUMÉS

Français

Cet article propose une réflexion sur les défis de la documentation des langues minorisées dans l'espace numérique à partir des travaux réalisés dans le cadre du projet DIVITAL. Les premiers travaux du projet ont concerné la collecte de corpus et leur documentation par des métadonnées à grain fin. Ces travaux ont mis en évidence deux défis majeurs : (i) l'identification des langues et de leurs variantes, dans le cadre des normes de codification des noms de langues, et (ii) la création de nouvelles ressources en lien avec les pratiques actuelles de ces langues.

English

This article looks at the challenges of documenting minority languages in the digital environment, based on work carried out as part of the DIVITAL project. The project's initial work involved collecting corpora and documenting them using fine-grained metadata. This work has highlighted two major challenges: (i) the identification of languages and their variants, within the framework of standards for the codification of language names, and (ii) the creation of new resources linked to the current practices of these languages.

Deutsch
Dieser Artikel stellt Überlegungen zu den Herausforderungen der Dokumentation von Minderheitensprachen im digitalen Raum an, ausgehend von den Arbeiten, die im Rahmen des DIVITAL-Projekts durchgeführt wurden. Die ersten Arbeiten des Projekts betrafen die Sammlung von Korpora und ihre Dokumentation durch feinkörnige Metadaten. Diese Arbeiten haben zwei große Herausforderungen aufgezeigt: (i) die Identifizierung der Sprachen und ihrer Varianten im Rahmen der Normen für die Kodierung von Sprachnamen und (ii) die Schaffung neuer Ressourcen in Verbindung mit der aktuellen Praxis dieser Sprachen.

INDEX

Mots-clés
langue minorisée, traitement automatique des langues, corpus, norme, traduction

Keywords
minority language, natural language processing, corpus, norm, translation

Schlagwortindex
Minderheitensprache, maschinelle Sprachverarbeitung, Korpus, Norm, Übersetzung

AUTEURS

Delphine Bernhard
LiLPa UR 1339, université de Strasbourg.
Maîtresse de conférences en informatique à la faculté des langues de l'université de Strasbourg. Ses travaux de recherche portent sur le traitement automatique des langues, la fouille de textes et les ressources linguistiques. Elle se consacre en particulier aux langues moins bien dotées en ressources et aux défis liés au développement de méthodes et de ressources appropriées pour ces langues.
IDREF : https://www.idref.fr/112578063
ORCID : http://orcid.org/0000-0001-7857-5873
HAL : https://cv.archives-ouvertes.fr/delphine-bernhard

Marianne Vergez-Couret
FoReLLIS UR 15076, université de Poitiers.
Maîtresse de conférences en sciences du langage à l'université de Poitiers. Elle inscrit ses travaux de recherche dans le champ de la sémantique du discours et participe au développement de ressources et outils pour l'analyse linguistique outillée de l'occitan et du poitevin-saintongeais.

IDREF : https://www.idref.fr/153254564
ORCID : http://orcid.org/0000-0002-0483-0525
HAL : https://cv.archives-ouvertes.fr/marianne-vergez-couret
ISNI : http://www.isni.org/0000000356884370

Estèle Dupuy
FoReLLIS UR 15076, université de Poitiers.
Maîtresse de conférences en sciences du langage à l'université de Poitiers. Diachronicienne, elle inscrit ses travaux de recherche dans le champ de la sémantique du discours et des relations syntactico-sémantiques visant à la continuité référentielle, de la lexicologie appliquée, la représentation de l'oral et l'oral représenté et de la psycholinguistique. Elle participe également au développement de ressources pour l'analyse linguistique outillée.
IDREF : https://www.idref.fr/061100021
ORCID : http://orcid.org/0000-0003-0924-6579
HAL : https://cv.archives-ouvertes.fr/estele-dupuy