Cahiers du plurilinguisme européen
ISSN : 2970-989X

12 | 2020
Langues minoritaires : quels acteurs pour quel avenir ?

L'avenir numérique des langues minoritaires : bilan du projet RESTAURE pour l'alsacien, l'occitan et le picard

Delphine Bernhard, Myriam Bras, Anne-Laure Ligozat, Aleksandra Miletic, Jean Sibille, Amalia Todirascu et Marianne Vergez-Couret

https://www.ouvroir.fr/cpe/index.php?id=1253

DOI : 10.57086/cpe.1253

Référence électronique
Delphine Bernhard, Myriam Bras, Anne-Laure Ligozat, Aleksandra Miletic, Jean Sibille, Amalia Todirascu et Marianne Vergez-Couret, « L'avenir numérique des langues minoritaires : bilan du projet RESTAURE pour l'alsacien, l'occitan et le picard », *Cahiers du plurilinguisme européen* [En ligne], 12 | 2020, mis en ligne le 01 décembre 2020, consulté le 28 octobre 2026. URL : https://www.ouvroir.fr/cpe/index.php?id=1253

Droits d'auteur
Licence Creative Commons – Attribution - Utilisation non commerciale - Partage dans les mêmes conditions 4.0 (CC-BY-NC-SA)

L'avenir numérique des langues minoritaires : bilan du projet RESTAURE pour l'alsacien, l'occitan et le picard

Delphine Bernhard, Myriam Bras, Anne-Laure Ligozat, Aleksandra Miletic, Jean Sibille, Amalia Todirascu et Marianne Vergez-Couret

PLAN

1. Présentation du projet RESTAURE
   1.1. Participants et objectifs
   1.2. Langues du projet
2. Défis et solutions mises en œuvre
   2.1. Acquisition et annotation de données
2.2 Variations dialectales et orthographiques
3. Impacts du projet
   3.1. Bilan scientifique
   3.2. Contributions au rayonnement des langues du projet
Conclusion et perspectives

TEXTE

1 Le numérique est un des enjeux de la revitalisation des langues minoritaires. Dans ce domaine, l'écart entre langues « bien dotées » (moins d'une dizaine de langues)[1] et langues « peu dotées » se creuse considérablement (Rehm *et al.*, 2014 ; Soria et Mariani, 2013). Cet écart est documenté également en France, où l'on constate que les langues régionales sont très peu dotées en ressources et outils numériques, en comparaison au français (Leixa, Mapelli et Choukri, 2014). Il est indispensable pour toutes les langues, et *a fortiori* pour les langues dites « minoritaires », de se faire une place dans l'ère du numérique pour renforcer leur visibilité, faciliter leur utilisation, accompagner leur enseignement. Des recommandations ont d'ailleurs été récemment émises à ce sujet, dans le cadre du *Digital Language Diversity Project* (Ceberio Berger *et al.*, 2018). Or, les langues disposant de peu de ressources ont en commun que leur informatisation a une faible rentabilité qui ne compense pas des coûts de développement

importants. Le défi que constitue la constitution de ressources et d'outils électroniques pour ces langues est donc considérable.

2 Le projet RESTAURE [2] (RESsources informatisées et Traitement AUtomatique pour les langues REgionales), financé par l'ANR (2015-2018) [3] a eu pour objectif de fournir des ressources numériques (en particulier à travers la constitution de corpus et lexiques) et d'outils de traitement automatique des langues (TAL) pour trois langues dites « régionales » de France : l'alsacien, l'occitan et le picard.

3 Nous allons dans un premier temps présenter ce projet, à travers ses objectifs et les langues traitées. Puis seront détaillés certains défis méthodologiques qui ont été soulevés : données et outils numériques rares et éparses, descriptions des langues incomplètes, variations dialectales et graphiques. Ces défis ne sont pas spécifiques aux langues du projet et sont représentatifs des difficultés lors de travaux portant sur les langues dites « peu dotées ». Nous discuterons ensuite des solutions qui ont été proposées, sur la base de recommandations visant à améliorer la vitalité numérique [4] des langues minoritaires et peu dotées (Soria, Mariani et Zoli, 2013 ; Ceberio Berger *et al.*, 2018) : coopération, utilisation de standards, réutilisabilité des ressources et outils. Enfin, nous conclurons en détaillant les leçons tirées du projet RESTAURE.

1. Présentation du projet RESTAURE

1.1. Participants et objectifs

4 Le projet RESTAURE a réuni des chercheurs de quatre unités de recherche situées à Strasbourg (Université de Strasbourg — LiLPa UR 1339), Toulouse (Université Toulouse Jean-Jaurès — CLLE-ERSS), Amiens (Université de Picardie Jules Verne — Habiter le monde) et Orsay (LIMSI), spécialistes à la fois des trois langues du projet (alsacien, occitan et picard) et du traitement automatique des langues. Il avait trois objectifs principaux :

- développer des ressources (corpus et lexiques) ;
- développer des outils pour l'acquisition et l'analyse de corpus écrits ;

- diffuser ces ressources et outils auprès des chercheurs et des non spécialistes.

5 Ces objectifs répondaient à des besoins en ressources et outils numériques identifiés lors du dépôt du projet et prenant en compte l'existant. La situation des différentes langues avant le démarrage du projet est résumée dans le tableau 1. Ce tableau rend compte de nombreux déficits et d'une situation hétérogène en terme d'expériences et de besoins. Elle reflète également l'état des lieux établi en 2014 dans l'*Inventaire des ressources linguistiques des langues de France* (Leixa, Mapelli et Choukri, 2014) qui indique un volume faible de ressources linguistiques pour l'alsacien (8/10)[5] et les langues d'oïl (7/10), dont le picard[6] fait partie, et moyen pour l'occitan[7] (5/10), qui se situe au même niveau que le breton et en deçà du basque (4/10) ou du catalan (4/10). Le français obtient le meilleur classement parmi les langues de France avec un score de 3/10.

Tableau 1 : Ressources et outils avant le début du projet

Ressource/ outil	**alsacien**	**occitan**	**picard**
Corpus brut	∅	*BaTelÒc* base expérimentale (Bras et Thomas, 2008)	PICARTEXT
Corpus annoté	∅	∅	∅
Lexique et dictionnaires	(Bernhard, 2014)	*Dico d'òc* (Congrès Permanent de la Lenga Occitana) en cours de construction	∅
Tokéniseur	∅	∅	∅
Étiqueteur morphosyntaxique	(Bernhard et Ligozat, 2013a ; Bernhard et Ligozat, 2013b)	(Vergez-Couret, 2013 ; Vergez-Couret et Urieli, 2014)	∅
Analyseur syntaxique	∅	∅	∅
Niveau de classement établi par (Leixa, Mapelli et Choukri, 2014)	8	5	7

6 Il est à noter que les notions de corpus, lexique ou dictionnaire utilisées dans le cadre du projet RESTAURE désignent des ressources directement utilisables pour des travaux en traitement automatique

des langues[8]. Ainsi, des textes ou des lexiques qui seraient disponibles dans des formats numériques non structurés ou semi-structurés (pages web, PDF, document traitement de texte) ne constituent pas à proprement parler des ressources directement exploitables sans travaux préparatoires : extraction et balisage des éléments d'intérêt, suppression des informations inutiles, transformation vers un format standard comme TEI (TEI Consortium, 2020).

7 De plus, certaines ressources existantes, comme le *Dico d'òc* contiennent des données structurées directement exploitables pour des travaux en traitement automatique des langues mais ne sont pas nécessairement disponibles en totalité pour la recherche en raison du droit de propriété intellectuelle.

1.2. Langues du projet

8 Le tableau 2 ci-dessous dresse un bref état comparatif des trois langues du projet, selon diverses caractéristiques : famille linguistique, situation sociolinguistique, production écrite et standardisation à l'écrit. Il montre les différences entre les langues du projet, mais aussi leurs points communs.

Tableau 2 : Comparatif des trois langues du projet RESTAURE

Caractéristiques	**alsacien**	**occitan**	**picard**
Famille linguistique	langue germanique / haut-allemand	langue romane / gallo-roman méridional	langue romane / gallo-roman septentrional
Situation sociolinguistique	Dialectes utilisés principalement dans des contextes informels. Pas d'enseignement à l'école publique, mais dans quelques écoles associatives.	Utilisation de la langue dans toutes ses variantes dialectales dans des contextes formels (enseignement, recherche, presse, médias) et informels. Enseignement public et associatif (initiation bilingue, immersion)	Dialectes utilisés principalement dans des contextes informels. Pas d'enseignement à l'école publique.

Production écrite	Utilisé à l'écrit depuis au moins la seconde moitié du XVII^{e} siècle, dans deux genres principaux : pièces de théâtre et poésie. D'autres genres sont aussi représentés : prose poétique, chansons, comptines, contes, traductions et adaptation d'œuvres dans d'autres langues. Les textes en prose sont assez rares.	Production écrite littéraire abondante et continue depuis les troubadours au Moyen-Âge. Tous les genres sont représentés : poésie, prose, théâtre, chansons, contes, comptines, presse, textes techniques, incluant traductions.	Utilisé à l'écrit dans tous les genres littéraires dont les plus représentés : théâtre, fable, roman, récits brefs, poésie, chanson.
Standardisation typo-orthographique	Propositions récentes de conventions orthographiques (par exemple ORTHAL (Zeidler et Crévenat-Werner, 2008)), mais leur utilisation n'est pas généralisée. Nombreuses graphies individuelles	Co-existance de deux standards principaux : graphie mistralienne et graphie classique. Nombreuses graphies individuelles.	Pas de standardisation.

2. Défis et solutions mises en œuvre

9 Les défis rencontrés lors du projet ont directement trait aux langues traitées. Il s'agit, comme nous l'avons montré précédemment, de langues peu dotées, ce qui entraîne des difficultés liées à l'acquisition et à l'annotation de données. Par ailleurs, ces langues ne sont pas uniformes et normées comme peut l'être le français standard à l'écrit, ce qui induit des variations dialectales et orthographiques.

10 Les solutions mises en œuvre répondent quant à elles à trois principes essentiels : la collaboration entre chercheurs, la réutilisation et le recyclage d'outils existants et l'utilisation de standards. Ces principes suivent les recommandations données par Soria *et al.* (2013) dans un article en faveur du développement des technologies des

langues pour les langues dites régionales et/ou minoritaires. Dans ce même article, Soria *et al.* préconisent de documenter les ressources et outils et de les partager, avec des licences permettant leur réutilisation par d'autres chercheurs. Ces diverses recommandations vont dans le sens des principes FAIR (*Findable Accessible Interoperable Reusable*) pour repérer les ressources numériques, les rendre accessibles et interopérables et faciliter leur réutilisation (Wilkinson *et al.*, 2016).

2.1. Acquisition et annotation de données

11 Le premier défi d'importance concerne les données et le « goulot d'étranglement » que constitue le passage des données brutes aux données enrichies par des annotations linguistiques ; ce dernier problème se rencontre d'ailleurs également pour les langues mieux dotées, mais dans une moindre mesure.

12 La figure 1 ci-dessous représente ce goulot d'étranglement et les différents défis auxquels il faut faire face.

Figure 1 : Passage des données brutes aux données annotées



13 **1.** La collecte des données brutes (corpus de textes) est rendue difficile par la faible quantité de ressources disponibles. À cela s'ajoute la difficulté à trouver des données dont les droits d'utilisation permettent une diffusion libre des corpus collectés pour des besoins de recherche, en particulier pour les textes les plus récents (Kevers et Retali-Medori, 2019). En conséquence, les données collectées sont

de taille relativement faible, en comparaison aux corpus généralement exploités en TAL [9].

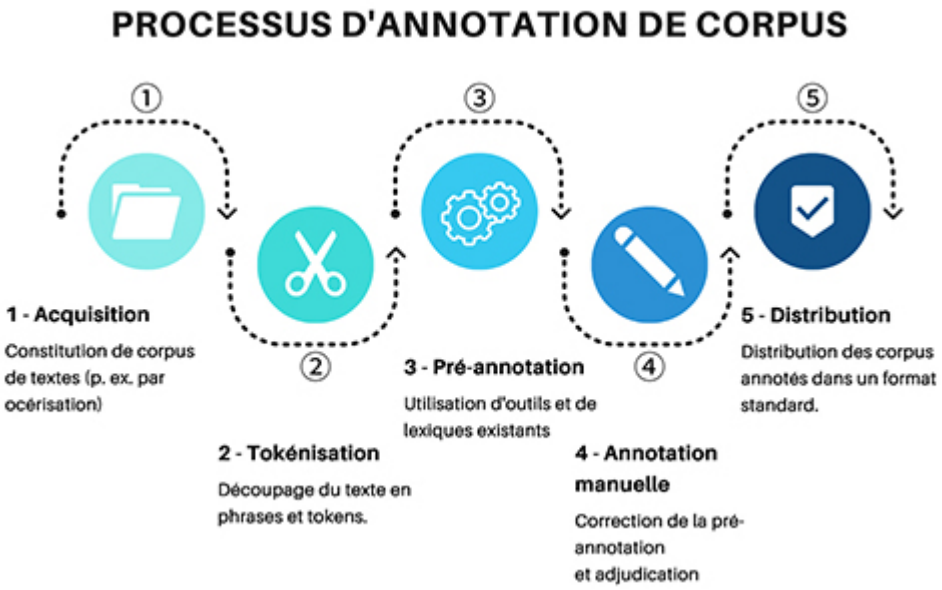
14 **2.** Des descriptions linguistiques précises et complètes sont nécessaires pour enrichir les corpus d'annotations (partie du discours, propriétés morphosyntaxiques) et produire des guides d'annotation. Il n'est pas toujours facile de trouver des grammaires de bonne qualité pour les langues régionales de France et, si elles existent, elles peuvent être dépassées ou incomplètes.

15 **3.** Le recrutement de personnes qui sont à la fois des locuteurs ayant une bonne maîtrise de la langue et des annotateurs expérimentés n'est pas une chose aisée.

16 **4.** Enfin, comme pour toute tâche d'annotation de corpus, il est nécessaire d'appliquer des procédures d'assurance qualité et de rédiger une documentation, en vue de la diffusion des ressources annotées. La vérification de la qualité des données annotées passe par une phase d'adjudication. Cette phase est également chronophage.

17 Les participants au projet avaient différents domaines de spécialité (sociolinguistique, linguistique descriptive, dialectologie, traitement automatique des langues, humanités numériques) et différents niveaux d'expérience dans la production de ressources et la manipulation d'outils de TAL. Or, le développement d'outils de TAL est coûteux et nécessite des compétences spécifiques. Les technologies évoluent de surcroît très rapidement et demandent donc des améliorations ou mises à jour constantes sous peine de voir les outils devenir rapidement obsolètes. Le processus d'annotation représenté par la figure 2 a été l'occasion pour les chercheurs de coopérer et s'entraider en fonction de leurs compétences. Un autre point important a été le recyclage d'outils et de ressources existantes, ainsi que l'encouragent Soria *et al.*, 2013.

Figure 2 : Processus d'annotation de corpus



18 Par exemple, le travail sur l'étape 1 d'acquisition de corpus par reconnaissance optique de caractères (OCR) a donné lieu à un travail commun entre Strasbourg et Toulouse, afin d'évaluer et d'entraîner deux outils différents, Tesseract (Smith, 2007) et Jochre (Urieli et Vergez-Couret, 2013), pour l'alsacien et l'occitan (Vergez-Couret *et al.*, 2017). Le développement d'un tokéniseur pour le picard a été réalisé conjointement par des chercheurs de Strasbourg et d'Amiens (Bernhard *et al.*, 2017). Le travail d'annotation de corpus et de constitution de lexiques pour le picard effectué à Amiens a été largement soutenu par les chercheurs du LIMSI à Orsay. L'annotation des corpus alsaciens et occitans s'est faite à l'aide d'un outil appelé AnaLog (Lay et Pincemin, 2010), avec une pré-annotation réalisée à l'aide d'outils étiquetage morphosyntaxique existants : Talismane (Urieli et Vergez-Couret, 2013), TreeTagger (Schmid, 1994), Apertium (Armentano I Oller, 2008).

19 La collaboration entre chercheurs nécessite toutefois des efforts de coordination et se heurte aussi à des problèmes liés au cadre même d'un projet coopératif : enchaînement des recrutements, temps disponible pour les enseignants-chercheurs, problème de la distance, dialogue entre disciplines très différentes. En ce sens, la figure 2 donne une représentation idéale du processus d'annotation qu'il peut

être nécessaire d'ajuster en fonction des langues et de ce qu'il est possible de faire pour les chercheurs impliqués.

20 Cela étant, nous pouvons affirmer rétrospectivement que de nombreuses tâches n'auraient pu être accomplies, ou alors sous une forme moins sophistiquée, sans la collaboration entre les différentes équipes. Cette coopération a permis de compenser, dans une certaine mesure, le manque de ressources humaines et de spécialistes du traitement automatique pour les langues régionales étudiées. Le travail parallèle sur plusieurs langues a permis de gagner en efficacité en bénéficiant des expériences réalisées sur d'autres langues. Les problèmes qui se sont posés dans une langue ont conduit à une vigilance accrue sur ces sujets dans les autres langues.

2.2 Variations dialectales et orthographiques

21 L'alsacien, l'occitan et le picard ne sont pas des langues homogènes ni totalement standardisées à l'écrit. Différentes variétés ou dialectes peuvent être identifiés dans chaque région. Les conventions orthographiques pour l'écrit sont soit assez récentes, soit peu utilisées, soit conçues de manière à rendre compte des particularités dialectales. Ces variations ont un impact négatif sur le développement d'outils de TAL. D'une part, l'absence ou le non-respect de normes orthotypographiques impacte la première étape de la chaîne de traitement (*tokénisation*), qui consiste à identifier les unités de base que sont les mots, signes de ponctuation, symboles ou autres. L'utilisation irrégulière ou inattendue des signes de ponctuation ou des espaces complique cette phase (Bernhard *et al.*, 2017). Par exemple, en picard, l'apostrophe peut marquer l'élision d'une voyelle au milieu d'un mot (ramintuv'lant/souvenant) ou à la frontière (m'/me l'/le). Dans le premier cas, il n'est pas séparateur, alors qu'il l'est dans le deuxième cas :

> In m' ramintuv'lant l' timps passé,…
> En me souvenant le temps passé,…

22 Par ailleurs, les outils de TAL sont très sensibles au problème des mots inconnus ou « hors vocabulaire » (OOV - *Out Of Vocabulary*).

Les mots hors vocabulaire sont des mots qui ne sont pas connus d'un système car il est tout simplement impossible de lister tous les mots possibles d'une langue (formes rares, néologismes, entités nommées, formes erronées, etc.). Les mots hors vocabulaire sont problématiques pour le traitement automatique de toutes les langues, et encore plus dans les langues peu dotées et présentant des variantes orthographiques (voir par exemple [Snoeren, Adda-Decker et Adda, 2010] pour le luxembourgeois). Les stratégies visant à réduire le nombre de mots inconnus cherchent généralement à normaliser les variantes orthographiques vers une forme connue : forme contemporaine dans le cas des états anciens d'une langue (Bollmann, 2019), forme standard pour la communication médiée par ordinateur (Mosquera, Lloret et Moreda, 2012) ou pour des variantes dialectales (Frey, Glaznieks et Stemle, 2015). Il est également possible de simplement détecter les variantes, sans pour autant chercher à les normaliser (Dasigi et Diab, 2011 ; Barteld, Biemann et Zinsmeister, 2019), ce qui évite d'avoir à choisir une norme ou de trouver des solutions pour les formes qui ne trouvent pas facilement une correspondance dans cette norme (formes disparues ou spécifiques à la variante étudiée). Une solution intermédiaire, lors de l'annotation manuelle d'un corpus, consiste à indiquer une glose dans une autre langue (Jarrar *et al.*, 2016), ce qui permet d'identifier facilement les variantes ayant la même traduction, tout en offrant une désambiguïsation sémantique. Nous avons privilégié cette dernière solution (voir section 3.2 pour des exemples).

3. Impacts du projet

23 Il existe diverses manières de mesurer l'impact du projet, en fonction de différents points de vue. Il est ainsi possible de faire un bilan scientifique « objectif » des ressources produites, en incluant leur mode de diffusion et leur potentiel de réutilisation dans le cadre de nouveaux travaux de recherche. Un autre bilan, plus difficile à réaliser, concerne les éventuelles contributions au développement des langues dites minoritaires concernées, que ce soit dans le milieu de la recherche universitaire ou auprès des locuteurs de ces langues.

3.1. Bilan scientifique

24 Le tableau 3 ci-dessous dresse un récapitulatif des travaux réalisés par les équipes du projet RESTAURE, en incluant également les travaux cités dans le tableau 1.

Tableau 3 : Récapitulatif des ressources et outils produits dans le cadre du projet RESTAURE

Ressource / outil	alsacien	occitan	picard
Corpus brut	(Bernard *et al*, 2018a)	BaTelÒc base opérationnelle (Bras et Vergez-Couret, 2016)	PICARTEXT
Corpus annoté	(Bernard *et al*, 2018a)	(Bras *et al.*, 2018)	(Martin, Rey et Reynés, 2018)
Lexiques et dictionnaires	(Bernhard, 2014 ; Bernhard *et al.*, 2018b ; Steiblé et Bernhard, 2018)	LoFlòc (Vergez-Couret, 2016 ; Bras *et al.*, 2017)	∅
Tokéniseur	(Bernard, 2018)	(Vergez-Couret, 2019)	(Todirascu, 2018 ; Ligozat, 2018)
Étiqueteur morphosyntaxique	(Bernhard et Ligozat, 2013a ; Bernhard et Ligozat, 2013b)	(Vergez-Couret, 2013 ; Vergez-Couret et Urieli, 2014 ; Vergez-Couret et Urieli, 2015)	∅
	(Magistry *et al.*, 2018 ; Magistry *et al.*, 2019)		
Analyseur syntaxique	∅	∅	∅

25 En particulier, nous avons choisi de partager les corpus annotés produits par le projet RESTAURE dans le format CONLL-U[10], défini dans le projet *Universal Dependencies* (Nivre *et al.*, 2016). Ce projet définit par ailleurs 17 parties du discours considérées comme universelles[11]. L'utilisation de standards de la communauté scientifique est importante, car elle permet de faciliter la réutilisation des ressources grâce à l'utilisation d'un modèle commun.

26 Nous avons utilisé ces catégories dans la diffusion de nos corpus annotés, après avoir projeté les étiquettes utilisées pour l'annotation initiale (VO) vers les parties du discours *Universal Dependencies* (UD) (Miletic *et al.*, 2019). Ces corpus donnent également la traduction

en français pour chaque mot ainsi que le lemme, en plus de la forme trouvée dans le texte. Comme nous l'avons expliqué dans la section 2.2, l'ajout de la glose en français est utile car elle permet d'identifier les variantes à partir de leur traduction. Elle rend aussi possible la constitution de lexiques bilingues à partir des corpus annotés.

27 Les tableaux 4 à 6 ci-dessous donnent des exemples pour l'alsacien, l'occitan et le picard.

Tableau 4 : Exemple d'annotation pour l'alsacien

tokens	VO	UD	lemme	glose
Mìtem	APPRART	ADP	mìt	avec
		DET	de	le
Sabayon	NOUN	NOUN	Sabayon	sabayon
ìwwerzìehje	VERB	VERB	ìwwerzìehje	napper
ùn	CONJ	CCONJ	ùn	et
mìt	ADP	ADP	mìt	avec
de	DET	DET	de	les
g'hobelte	ADJ	ADJ	g'hobelt	effilé
Màndle	NOUN	NOUN	Màndel	amande
bstraie	VERB	VERB	bstraie	saupoudrer
.	PUNCT	PUNCT	.	.

Tableau 5 : Exemple d'annotation pour l'occitan

tokens	VO	UD	lemme	glose
Cossí	Rx	ADV	cossí	comment
aquò	Pd	PRON	aquò	ça
pòt	Vm	VERB	poder	pouvoir
èsser	Vm	VERB	èsser	être
?	F	PUNCT	?	?

Tableau 6 : Exemple d'annotation pour le picard

tokens	VO	UD	lemme	glose

Il	PRONPERS	PRON	Il	Il
est	VERBCONJ	VERB	ète	est
rétampi	ADV	ADV	rétampir	debout
d'puis	ADP	ADP	depuis	depuis
bientout	ADV	ADV	bientôt	bientôt
troés	NUM	NUM	troés	trois
ins	NOUN	NOUN	in	ans
.	PUNCT	PUNCT	.	.

28 L'utilisation de ces standards permettra à l'avenir d'entraîner facilement d'autres outils pour la tokénisation et l'étiquetage morphosyntaxique comme UDPipe (Straka et Straková, 2017) ou Stanza (Qi *et al.*, 2020). En effet, la plupart des outils de TAL actuels sont capables d'apprendre à partir des données. Les méthodes ont évolué, passant d'approches essentiellement fondées sur des règles à des techniques d'apprentissage automatique, qui sont en principe applicables à une grande variété de langues. La condition principale est que des données soient disponibles pour les réutiliser. Nous avons donc choisi de nous concentrer sur la collecte et l'annotation des données, plutôt que sur le développement d'outils. Comme souligné précédemment, les outils peuvent ensuite être réutilisés ou recyclés.

29 Pour la diffusion des ressources produites, nous avons veillé à respecter les principes FAIR (cf. section 2), afin que le travail puisse bénéficier à d'autres chercheurs :

- Les ressources sont associées à un identifiant pérenne (DOI), sont décrites par des métadonnées et déposées dans un entrepôt de données (Zenodo) ;
- Des liens explicites sont effectués entre les articles publiés (décrits sur HAL) et les ressources et outils déposés sur Zenodo ;
- Les ressources et outils sont librement téléchargeables [12] et réutilisables, suivant une licence Creative Commons CC-BY-SA.

3.2. Contributions au rayonnement des langues du projet

30 Une autre manière de mesurer l'impact du projet est de vérifier s'il a eu un impact positif sur le rayonnement des langues minoritaires

concernées. Deux types de publics sont ici concernés : les chercheurs et les locuteurs des langues.

31 Du point de vue de la recherche, la mise à disposition des ressources du projet devrait permettre de favoriser les travaux d'autres chercheurs. Ainsi, le corpus annoté pour l'alsacien a été utilisé dans (Millour *et al.*, 2020) et la base textuelle *BaTelÒc* est utilisée fréquemment par les chercheurs qui travaillent sur l'occitan, par exemple (Esher 2018), (Bach, à paraître) ou sur d'autres langues romanes, comme le catalan (Garcia-Sebastià, 2018, Pujol i Campeny, 2020). Les chercheurs du projet continuent à exploiter les corpus annotés, le corpus occitan a ainsi été utilisé pour poursuivre la chaine de traitement avec l'étape de l'annotation syntaxique (Miletic *et al.*, 2019, 2020).

32 Un autre aspect important concerne la reconnaissance de ces travaux par la communauté scientifique. Or, le travail sur des langues peu dotées nécessite souvent de construire des ressources et outils à partir de rien ou presque, ce qui peut conduire à un sentiment d'infériorité par rapport aux langues qui disposent de plus de ressources (ressources humaines, ressources financières et état de l'art plus avancé). La production de ressources linguistiques exige du temps et des moyens, et les deux sont rares pour les langues peu dotées. Ces contraintes extrinsèques sont difficiles à contrôler, mais ne doivent pas compromettre la volonté des chercheurs de continuer à travailler sur ces langues. Pour ce faire, les organismes de financement ainsi que les comités de programme et de lecture doivent reconnaître les défis particuliers que pose le travail sur les langues peu dotées. L'inévitable « retard » de ces langues sur les langues mieux dotées conduit trop souvent à une évaluation scientifique négative faisant état d'un manque d'originalité des travaux. Aider les chercheurs à relever ces défis relève selon nous du simple respect d'un principe d'égalité entre les humains et les langues qu'ils parlent, comme le résume la formule de Marcel Félix Castan, emblématique du Forom des Langues du Monde de Toulouse, « les langues et les cultures du monde sont égales entre elles comme les citoyens d'une même république »[13].

33 Certains des travaux développés, en particulier les bases textuelles et les dictionnaires en ligne sont largement utilisés par les apprenants

et les locuteurs experts que sont les traducteurs et les enseignants des langues concernées. Nous en avons des preuves avec l'utilisation croissante de l'application *Dico d'Òc* par les apprenants de l'occitan (enseignement secondaire, universitaires, formation pour adultes) et par l'utilisation croissante également de *Bateloc* par les enseignants d'occitan et les traducteurs.

34 Enfin, l'impact sur les locuteurs ordinaires est plus indirect. En effet, la plupart des ressources et outils constitués dans le cadre du projet ne sont pas directement destinés au grand public. Il n'en reste pas moins que les ressources collectées peuvent permettre de développer des outils utiles à tous. Ainsi, les corpus collectés pour l'alsacien et l'occitan ont été utilisés pour deux claviers prédictifs pour smartphones [14]. Par ailleurs, l'existence de travaux scientifiques sur ces langues conduit à les légitimer et augmenter leur visibilité. Ces travaux suscitent aussi l'attention des médias : le projet RESTAURE a ainsi bénéficié d'une couverture médiatique par différentes chaînes de télévision et stations de radio (France 3 Sud, Alsace 20, France Bleu Elsass, etc.).

Conclusion et perspectives

35 Nous avons, dans cet article, fait le bilan d'un projet visant à améliorer l'assise numérique de trois langues régionales de France. Sans pour autant évacuer les défis inhérents à un travail collaboratif et interdisciplinaire sur des langues très différentes, nous avons montré qu'il est possible d'aboutir à des résultats concrets grâce à la mise en place de principes simples mais efficaces : collaboration entre chercheurs, réutilisation et recyclage d'outils, utilisation de standards et diffusion des ressources. Nous avons également tiré un certain nombre de leçons de ce projet, qui pourront être utiles aux chercheurs souhaitant se lancer dans une entreprise similaire : importance de la coopération, nécessaire prise en compte du statut particulier des langues peu dotées par rapport au numérique, exploitation de l'existant, efforts axés sur les données plutôt que sur les outils. Pour reprendre à notre compte l'adage bien connu (« A dwarf standing on the shoulders of a giant may see farther than a giant himself », Robert Burton, utilisé par Soria, Mariani et Zoli, 2013 dans le titre de leur l'article), nous pouvons résumer ces diverses leçons de la manière suivante :

« Un nain debout sur les épaules d'autres nains peut voir presque aussi loin qu'un géant. »

36 Nous avons conscience que le chemin à parcourir est encore long pour assurer l'avenir numérique de l'alsacien, de l'occitan et du picard, mais nous espérons avoir pu contribuer à améliorer la situation pour ces langues.

BIBLIOGRAPHIE

ARMENTANO I OLLER Carme, 2008, « Traduction automatique occitan-catalan et occitan-espagnol : difficultés affrontées et résultats atteints », communication orale au *Neuvième Congrès International de l'Association Internationale d'Études Occitanes*, Aix-la-Chapelle, 24-31 août 2008.

BACH Xavier, 2020, « Morfosemantica dels augmentatius en occitan del Lengadòc », dans COUROUAU Jean-François, FABIÉ David (éds), *Fidelitats e dissidéncias. Actes del XIIn Congrès de l'Associacion internacionala d'estudis occitans. Actes du XII^e Congrès de l'Association internationales d'études occitanes. Albi 10-15/07/2017*, Toulouse, SFAIEO, p. 115-120.

BARTELD Fabian, BIEMANN Chris et ZINSMEISTER Heike, 2019, « Token-based spelling variant detection in Middle Low German texts », dans *Language Resources and Evaluation*, p. 1-30, <https://doi.org/10.1007/s10579-018-09441-5>.

BERNHARD Delphine, 2014, « Adding Dialectal Lexicalisations to Linked Open Data Resources: the Example of Alsatian », dans *Proceedings of the Workshop on Collaboration and Computing for Under Resourced Languages in the Linked Open Data Era* (CCURL 2014), p. 23-29, <https://hal.archives-ouvertes.fr/hal-00966820>.

BERNHARD Delphine, 2018, *Tokeniser for the Alsatian Dialects*. Zenodo, <https://doi.org/10.5281/zenodo.2454993>.

BERNHARD Delphine, ERHART Pascale, HUCK Dominique et STEIBLÉ Lucie, 2018a, *Annotated Corpus for the Alsatian Dialects*, <https://doi.org/10.5281/zenodo.2536041>.

BERNHARD Delphine et LIGOZAT Anne-Laure, 2013a, « Hassle-free POS-Tagging for the Alsatian Dialects », dans ZAMPIERI Marcos et DIWERSY Sascha (dir.) *Non-Standard Data Sources in Corpus Based-Research*, Coll. « ZSM Studien », Shaker, p. 85-92, <http://hal.archives-ouvertes.fr/hal-00860790>.

BERNHARD Delphine et LIGOZAT Anne-Laure, 2013b, « Es esch fàscht wie Ditsch, oder net? Étiquetage morphosyntaxique de l'alsacien en passant par l'allemand », dans *Actes de TALARE 2013 : Traitement Automatique des Langues Régionales de France et*

d'Europe, p. 209-220, <http://hal.archives-ouvertes.fr/hal-00838355>.

BERNHARD Delphine, MAGISTRY Pierre, LIGOZAT Anne-Laure et ROSSET Sophie, 2018b, « Resources and Methods for the Automatic Recognition of Place Names in Alsatian », dans *Corpus-Based Research in the Humanities*, p. 35-44, <https://hal.archives-ouvertes.fr/hal-01702656>.

BERNHARD Delphine, TODIRASCU Amalia, MARTIN Fanny, ERHART Pascale, STEIBLÉ Lucie, HUCK Dominique et REY Christophe, 2017, « Problèmes de tokénisation pour deux langues régionales de France, l'alsacien et le picard », dans *DiLiTAL 2017*, p. 14-23, <https://hal.archives-ouvertes.fr/hal-01539160>.

BOLLMANN Marcel, 2019, « A Large-Scale Comparison of Historical Text Normalization Systems », dans *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)*, Association for Computational Linguistics, p. 3885-3898, <https://doi.org/10.18653/v1/N19-1389>.

BRAS Myriam et VERGEZ-COURET Marianne, 2016, « BaTelÒc: A text base for the Occitan language », dans *Language Documentation and Conservation in Europe*, p. 133-149, <https://scholarspace.manoa.hawaii.edu/bitstream/10125/24675/1/bras_vergez-couret_2016.pdf>.

BRAS Myriam, ESHER Louise, SIBILLE Jean et VERGEZ-COURET Marianne, 2018, *Annotated Corpus for Occitan*, <https://doi.org/10.5281/zenodo.1182949>.

BRAS Myriam et THOMAS Jean, 2011, « Batelòc : cap a una basa informatisada de tèxtes occitans », dans RIEGER A., SUMIEN D. (éds), *L'Occitanie invitée de l'Euregio. Liège 1981 – Aix-la-Chapelle 2008 : Bilan et perspectives/Occitània convidada d'Euregio. Lièja 1981 – Aquisgran 2008: Bilanç e amiras/ Okzitanien zu Gast in der Euregio. Lüttich 1981 – Aachen 2008 : Bilanz und Perspektiven. Actes du Neuvième Congrès International de l'Association Internationale d'Études Occitanes, Aix-la-Chapelle, 24-31 août 2008*, Aachen, Shaker, p. 661-670, <https://hal.archives-ouvertes.fr/hal-00986409>.

BRAS Myriam, VERGEZ-COURET Marianne, HATHOUT Nabil, SIBILLE Jean, SÉGUIER Aure et DAZÉAS Benaset, 2020, « Loflòc : Lexic obèrt flechit occitan », in COUROUAU Jean-François, FABIÉ David (éds), *Fidelitats e dissidéncias. Actes del XIIn Congrès de l'Associacion internacionala d'estudis occitans. Actes du XII^e Congrès de l'Association internationales d'études occitanes. Albi 10-15/07/2017*, Toulouse, SFAIEO, p. 141-15.

CEBERIO BERGER Klara, GURRUTXAGA HERAIZ Antton, BARONI Paola, DAVYTH Hicks, KRUSE Eleonore, QUOCHI Valeria, RUSSO Irene, SALONEN Tuomo, SARHIMAA Anneli et SORIA Claudia, 2018, *Digital Language Survival Kit. The DLDP Recommendations to Improve Digital Vitality*, <http://www.dldp.eu/sites/default/files/documents/DLDP_Digital-Language-Survival-Kit.pdf>.

DASIGI Pradeep et DIAB Mona, 2011, « CODACT: Towards Identifying Orthographic Variants in Dialectal Arabic », dans *Proceedings of the 5th International Joint Conference on Natural Language Processing*, p. 318-326, <http://aclweb.org/anthology-new/I/I11/I11-1036.pdf>.

ESHER Louise, 2021, « Grammaticalisation et emploi des expressions un pauc, un bocin,… "un peu" en occitan », dans BANEGAS-SAORIN Mercedes & SIBILLE Jean, *Entre francisation et démarcation : Usages hérités et usages renaissantistes des langues régionales de France/Carnets d'atelier de sociolinguistique*, n° 13, Paris, L'Harmattan, p. 61-78.

FREY Jennifer-Carmen, GLAZNIEKS Aivars et STEMLE Egon W., 2015, « The DiDi Corpus of South Tyrolean CMC Data », dans *Proceedings of the 2nd Workshop of the Natural Language Processing for Computer-Mediated Communication/Social Media*, p. 1-6.

GARCIA SEBASTIÀ Josep Vicent, 2018, *Les construccions de temps transcorregut en el català de l'edat moderna i contemporània : acostament segons la lingüística de corpus i la gramàtica cognitiva*, Thèse de doctorat, Université d'Alicante.

JARRAR Mustafa, HABASH Nizar, ALRIMAWI Faeq, AKRA Diyam et ZALMOUT Nasser, 2016, « Curras: an annotated corpus for the Palestinian Arabic dialect », dans *Language Resources and Evaluation*, p. 1-31, <http://link.springer.com/article/10.1007/s10579-016-9370-7>.

KEVERS Laurent et RETALI-MEDORI Stella, 2019, « Copyright in the context of tooling up Corsican and other less-resourced languages », dans *Proceedings of the International Conference on Language Technologies for All (LT4All)*, <https://lt4all.org/media/papers/O8/144.pdf>.

LAY Marie-Hélène et PINCEMIN Bénédicte, 2010, « Pour une exploration humaniste des textes : AnaLog », dans *Proceedings of 10th International Conference Journées d'Analyse statistique des Données Textuelles*, <http://www.ledonline.it/ledonline/JADT-2010/allegati/JADT-2010-1045-1056_106-Lay.pdf>.

LEIXA Jérémy, MAPELLI Valérie et CHOUKRI Khalid, 2014, *Inventaire des ressources linguistiques des langues de France*. ELDA-DGLFLF-2013A, Paris, ELDA/DGLFLF.

LIGOZAT Anne-Laure, 2018, *OpenNLP tokenization model for Picard*, <https://doi.org/10.5281/zenodo.1484520>.

MAGISTRY Pierre, LIGOZAT Anne-Laure et ROSSET Sophie, 2018, « Étiquetage en parties du discours de langues peu dotées par spécialisation des plongements lexicaux », dans *Actes de la conférence sur le Traitement Automatique des Langues Naturelles*, <https://hal.archives-ouvertes.fr/hal-01793092>.

MAGISTRY Pierre, LIGOZAT Anne-Laure et ROSSET Sophie, 2019, « Exploiting languages proximity for part-of-speech tagging of three French regional languages », dans *Language Resources and Evaluation*, p. 1-26.

MARTIN Fanny, REY Christophe et REYNÉS Philippe, 2018, *Annotated Corpus for Picard*, <https://doi.org/10.5281/zenodo.1485988>.

MCGILLIVRAY Barbara, POIBEAU Thierry et RUIZ FABO Pablo, 2020, « Digital Humanities and Natural Language Processing: Je t'aime… Moi non plus », dans *Digital Humanities Quarterly*, vol. 14, <http://www.digitalhumanities.org/dhq/vol/14/2/000454/000454.html>.

MILETIC Aleksandra, BERNHARD Delphine, BRAS Myriam, LIGOZAT Anne-Laure et VERGEZ-COURET Marianne, 2019, « Converting POS-tag and Lemma Annotations into the Universal Dependencies Format: A Case Study on Alsatian and Occitan », dans *Actes de la Conférence sur le Traitement Automatique des Langues Naturelles*, ATALA, p. 427-435, <https://hal.archives-ouvertes.fr/hal-02123743>.

MILETIC Aleksandra, BRAS Myriam, ESHER, Louise, SIBILLE Jean, VERGEZ-COURET Marianne, 2019, « Building a treebank for Occitan: what use for Romance UD Corpora? », dans Gerdes K., Kahane S. (eds.) Proceedings of the International Conference on Dependency Linguistics, SyntaxFest – Depling 2019, Paris, France, <https://hal.archives-ouvertes.fr/hal-02380554>.

MILETIC Aleksandra, BRAS Myriam, VERGEZ-COURET Marianne, ESHER, Louise, POUJADE Clamença, SIBILLE Jean, 2020, « Building a Universal Dependencies Treebank for Occitan », dans *Proceedings of the 12th Conference on Language Resources and Evaluation* (LREC 2020), p. 2932-2939, <https://hal.archives-ouvertes.fr/hal-02892715>.

MILLOUR Alice, FORT Karën et MAGISTRY Pierre, 2020, « Répliquer et étendre pour l'alsacien "Étiquetage en parties du discours de langues peu dotées par spécialisation des plongements lexicaux" », dans *Actes du 2e atelier Éthique et TRaitemeNt Automatique des Langues (ETeRNAL)*, p. 29-37, <https://hal.archives-ouvertes.fr/hal-02750224>.

MOSQUERA Alejandro, LLORET Elena et MOREDA Paloma, 2012, « Towards Facilitating the Accessibility of Web 2.0 Texts through Text Normalisation », dans *Proceedings of the Workshop on Natural Language Processing for Improving Textual Accessibility* (NLP4ITA), <http://www.taln.upf.edu/nlp4ita/pdfs/mosquera-nlp4ita2012.pdf>.

NIVRE Joakim, MARNEFFE Marie-Catherine de, GINTER Filip, GOLDBERG Yoav, HAJIC Jan, MANNING Christopher D., MCDONALD Ryan, PETROV Slav, PYYSALO Sampo, SILVEIRA Natalia, TSARFATY Reut et ZEMAN Daniel, 2016, « Universal Dependencies v1: A Multilingual Treebank Collection », dans *Proceedings of the Tenth International Conference on Language Resources and Evaluation* (LREC 2016), European Language Resources Association (ELRA), <http://www.lrec-conf.org/proceedings/lrec2016/pdf/348_Paper.pdf>.

PUJOL I CAMPENY Afra, « à paraître » (lieu d'édition encore inconnu, document uniquement disponible sur ResearchGate pour le moment), « La expresión de la polaridad positiva enfática en las lenguas romances: anteposición no focal y marcadores

focales positivos en catalán y occitano (https://www.researchgate.net/profile/Afra_Pujol_I_Campeny/publication/340297327_La_expresion_de_la_polaridad_positiva_enfatica_en_las_lenguas_romances_anteposicion_no_focal_y_marcadores_focales_positivos_en_catalan_y_occitano/links/5f16beb445851515ef3be9f1/La-expresion-de-la-polaridad-positiva-enfatica-en-las-lenguas-romances-anteposicion-no-focal-y-marcadores-focales-positivos-en-catalan-y-occitano) », *Studia Linguistica Romanica*, <https://www.researchgate.net/publication/340297327>.

QI Peng, ZHANG Yuhao, ZHANG Yuhui, BOLTON Jason et MANNING Christopher D., 2020, « Stanza: A Python Natural Language Processing Toolkit for Many Human Languages », dans *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, p. 101-108, <https://www.aclweb.org/anthology/2020.acl-demos.14.pdf>.

REHM Georg, USZKOREIT Hans, DAGAN Ido, GOETCHERIAN Vartkes, DOGAN Mehmet Ugur MERMER Coskun, VÁRADI Tamás, KIRCHMEIER-ANDERSEN Sabine, STICKEL Gerhard, PRYS JONES Meirion, OETER Stefan et GRAMSTAD Sigve, 2014, « An update and extension of the META-NET Study "Europe's Languages in the digital age" », dans *Proceedings of the Workshop on Collaboration and Computing for Under-Resourced Languages in the Linked Open Data Era*, p. 30-37.

SCHMID Helmut, 1994, « Probabilistic Part-of-Speech Tagging Using Decision Trees », dans *Proceedings of the International Conference on New Methods in Language Processing*, p. 44-49, <http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.28.1139>.

SMITH Ray, 2007, « An overview of the Tesseract OCR engine », dans *Proceedings of the 9th International Conference on Document Analysis and Recognition*, IEEE, p. 629-633, <http://www.computer.org/csdl/proceedings/icdar/2007/2822/02/28220629-abs.html>.

SNOEREN, Natalie D., ADDA-DECKER, Martine et ADDA, Gilles, 2010, « The Study of Writing Variants in an Under-resourced Language: Some Evidence from Mobile N-Deletion in Luxembourgish », dans *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, European Language Resources Association (ELRA), p. 601-605, <http://www.lrec-conf.org/proceedings/lrec2010/pdf/258_Paper.pdf>.

SORIA Claudia et MARIANI Joseph, 2013, « Searching LTs for minority languages », dans *Actes de TALaRE Traitement Automatique des Langues Régionales de France et d'Europe*, p. 235-247, <https://doi.org/10.13140/2.1.1798.7203>.

SORIA Claudia, MARIANI Joseph et ZOLI Carlo, 2013, « Dwarfs sitting on the giants' shoulders – how LTs for regional and minority languages can benefit from piggybacking major languages », dans *Proceedings of* XVII *FEL Conference*, Foundation for Endangered Languages, p. 73-79.

STEIBLÉ Lucie et BERNHARD, Delphine, 2018, « Pronunciation Dictionaries for the Alsatian Dialects to Analyze Spelling

and Phonetic Variation », dans *11th edition of the Language Resources and Evaluation Conference*, 7-12 May 2018, <https://hal.archives-ouvertes.fr/hal-01704814>.

STRAKA Milan et STRAKOVÁ Jana, 2017, « Tokenizing, POS Tagging, Lemmatizing and Parsing UD 2.0 with UDPipe », dans *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, Association for Computational Linguistics, p. 88-99, <http://www.aclweb.org/anthology/K/K17/K17-3009.pdf>.

TEI CONSORTIUM, 2020, *TEI P5: Guidelines for Electronic Text Encoding and Interchange*, <https://doi.org/10.5281/zenodo.3667251>.

TODIRASCU Amalia, 2018, *Tokeniser for Picard*. Zenodo, <https://doi.org/10.5281/zenodo.1493642>.

URIELI Assaf et VERGEZ-COURET Marianne, 2013, « Jochre, océrisation par apprentissage automatique : étude comparée sur le yiddish et l'occitan », dans *Actes de TALARE 2013 : Traitement Automatique des Langues Régionales de France et d'Europe*, p. 221-234, <https://hal-univ-tlse2.archives-ouvertes.fr/hal-00979665>.

VERGEZ-COURET Marianne, 2013, « Tagging Occitan using French and Castillan Tree Tagger », dans *Proceedings of the 6th Language & Technology Conference: Less Resourced Languages, new technologies, new challenges and opportunities*, p. 78-82, <https://hal.archives-ouvertes.fr/hal-00986426>.

VERGEZ-COURET Marianne, 2016, *Description du lexique Loflòc*. CLLE-ERSS, <https://hal.archives-ouvertes.fr/hal-01338774>.

VERGEZ-COURET Marianne, 2019, *Tokenization for Occitan (Gascon and Lengadocian)*. Zenodo, <https://doi.org/10.5281/zenodo.2533873>.

VERGEZ-COURET Marianne, BERNHARD Delphine, URIELI Assaf, BRAS Myriam, ERHART Pascale et HUCK Dominique, 2017, « Océrisation de textes pour les langues régionales. Regards croisés sur l'occitan et l'alsacien », dans *10e colloque international ISKO France*, ISTE Editions Ltd, p. 250-269, <https://hal.archives-ouvertes.fr/hal-01252241>.

VERGEZ-COURET Marianne et URIELI Assaf, 2014, « Pos-tagging different varieties of Occitan with single-dialect resources », dans *COLING 2014*, p. 21, <http://www.aclweb.org/anthology/W/W14/W14-53.pdf>.

VERGEZ-COURET Marianne et URIELI, Assaf, 2015, « Analyse morphosyntaxique de l'occitan languedocien : l'amitié entre un petit languedocien et un gros catalan », dans *TALARE 2015*, <https://hal.archives-ouvertes.fr/hal-01214566>.

WILKINSON Mark D., DUMONTIER Michel, AALBERSBERG IJsbrand Jan, APPLETON Gabrielle, AXTON Myles, BAAK Arie, BLOMBERG Niklas, BOITEN Jan-Willem, SANTOS Luiz Bonino da Silva, BOURNE Philip E., BOUWMAN Jildau, BROOKES Anthony J., CLARK Tim, CROSAS Mercè, DILLO Ingrid, DUMON Olivier, EDMUNDS Scott, EVELO Chris T., FINKERS Richard,

GONZALEZ-BELTRAN Alejandra, GRAY Alasdair J. G., GROTH Paul, GOBLE Carole, GRETHE, Jeffrey S., HERINGA, Jaap, HOEN, Peter A. C. 't, HOOFT, Rob, KUHN Tobias, KOK Ruben, KOK Joost, LUSHER Scott J., MARTONE Maryann E., MONS Albert, PACKER Abel L., PERSSON Bengt, ROCCA-SERRA Philippe, ROOS, Marco, SCHAIK, Rene van, SANSONE Susanna-Assunta, SCHULTES Erik, SENGSTAG Thierry, SLATER Ted, STRAWN George, SWERTZ, Morris A., THOMPSON, Mark, LEI, Johan van der, MULLIGEN Erik van, VELTEROP Jan, WAAGMEESTER Andra, WITTENBURG Peter, WOLSTENCROFT Katherine, ZHAO Jun et MONS Barend, 2016, « The FAIR Guiding Principles for scientific data management and stewardship », dans *Scientific Data*, vol. 3, nº 1, p. 1-9, <https://doi.org/10.1038/sdata.2016.18>.

ZEIDLER Edgar et CRÉVENAT-WERNER Danielle, 2008, *Orthographe alsacienne : bien écrire l'alsacien de Wissembourg à Ferrette*, Colmar, France, J. Do Bentzinger.

NOTES

1 Les langues « bien dotées » sont des langues pour lesquelles des ressources linguistiques de base sont disponibles : corpus de textes écrits, corpus de parole, corpus parallèles, ressources lexicales et grammaires (Rehm *et al.*, 2014).

2 <http://restaure.unistra.fr/>. Les travaux décrits dans cet article ont bénéficié du soutien de l'ANR (projet RESTAURE – référence ANR-14-CE24-0003). Nous remercions toutes les personnes ayant participé au projet et toutes celles qui nous ont aidés dans l'acquisition des données.

3 Référence ANR-14-CE24-0003.

4 Ceberio Berger *et al.* (2018) proposent trois indicateurs de la vitalité numérique. La capacité numérique (*digital capacity*) fait référence aux outils et infrastructures nécessaires pour communiquer numériquement : système d'écriture, accès à Internet, habileté numérique, claviers prédictifs. La présence et l'utilisation numérique (*digital presence and use*) rend compte de l'utilisation effective de la langue pour créer des contenus numériques (mails, réseaux sociaux, sites web, etc.). Enfin, la performance numérique (*digital performance*) correspond à des indicateurs tels que des réseaux sociaux ou logiciels localisés ou encore l'existence d'outils de traduction automatique.

5 Un classement de 1/10 indique une excellente base de ressources linguistiques. À l'inverse, un classement de 10/10 indique une base faible

ou inexistante.

6 <https://www.u-picardie.fr/LESCLaP/PICARTEXT/Public/>

7 <http://redac.univ-tlse2.fr/bateloc/>

8 Nous avons donc une définition plus restrictive que celle utilisée dans le rapport de Leixa *et al.*, 2014. Par ailleurs, nous ne nous intéressons qu'aux travaux sur l'écrit, tandis que Leixa *et al.* incluent également des ressources et outils pour l'oral.

9 Cela rejoint les observations faites par McGillivray *et al.* (2020) au sujet des différences entre corpus utilisés dans le domaine des Humanités Numériques et corpus pour le TAL.

10 <https://universaldependencies.org/format.html>

11 <https://universaldependencies.org/u/pos/index.html>

12 <https://zenodo.org/communities/restaure?page=1&size=20>

13 <http://www.arnaud-bernard.net/index.php/forom_des_langues.html>

14 AnySoftKeyboard, développé par la société basque Elhuyar, et disponible sur Android ; Microsoft SwiftKey, disponible sur Android et iOS.

RÉSUMÉS

Français

Le numérique est un des enjeux de la revitalisation des langues minoritaires. Or, l'informatisation de ces langues a une faible rentabilité qui ne compense pas des coûts de développement importants. Le défi que constitue la constitution de ressources et d'outils numériques pour ces langues est donc considérable. Le projet RESTAURE (Ressources informatisées et Traitement Automatique pour les langues régionales), financé par l'ANR (2015-2018) a eu pour objectif de fournir des ressources numériques, en particulier à travers la constitution de corpus et lexiques, et d'outils de traitement automatique des langues pour trois langues régionales de France : l'alsacien, l'occitan et le picard. Dans cet article, nous présentons le bilan du projet RESTAURE, et en particulier les défis méthodologiques qui ont été soulevés — données et outils numériques rares et éparses, variations dialectales et graphiques, descriptions des langues incomplètes — ainsi que les solutions qui ont été proposées — coopération, utilisation de standards. Nous tirons également les leçons principales de ce projet qui, nous l'espérons, pourront être utiles à d'autres chercheurs et à d'autres langues.

English
Digital technology is one of the challenges in the revitalization of minority languages. However, the computerization of these languages has a low profitability that does not compensate for significant development costs. The challenge of creating digital resources and tools for these languages is therefore considerable. The RESTAURE project (Ressources informatisées et Traitement Automatique pour les langues régionales), funded by the project-based funding agency for research in France ANR (2015-2018), aimed to provide digital resources, in particular through the constitution of corpora and lexicons, and automatic language processing tools for three regional languages of France: Alsatian, Occitan and Picard. In this article, we present the results of the RESTAURE project, and in particular the methodological challenges that were raised – rare and scattered digital data and tools, dialectal and graphic variations, incomplete language descriptions – as well as the solutions that were proposed – cooperation, use of standards. We also draw the main lessons from this project, which we hope will be useful to other researchers and to other languages.

Deutsch
Die Einführung digitaler Technologien ist eine der Herausforderungen bei der Revitalisierung von Minderheitensprachen. Die Computerisierung dieser Sprachen hat jedoch eine geringe Rentabilität, die die erheblichen Entwicklungskosten nicht ausgleicht. Die Herausforderung, digitale Ressourcen und Werkzeuge für diese Sprachen zu schaffen, ist daher beträchtlich. Das Projekt RESTAURE (Ressources informatisées et Traitement Automatique pour les langues régionales), das von der französischen projektbasierten Förderstelle für Forschung ANR (2015-2018) finanziert worden ist, hatte zum Ziel, digitale Ressourcen bereitzustellen, insbesondere durch die Erstellung von Korpora und Lexika sowie automatische Sprachverarbeitungswerkzeuge für drei Regionalsprachen Frankreichs: Elsässisch, Okzitanisch und Picard. In diesem Artikel stellen wir die Ergebnisse des RESTAURE-Projekts vor, insbesondere die methodischen Herausforderungen, die aufgeworfen wurden - seltene und verstreute digitale Daten, dialektale und grafische Variationen, unvollständige Sprachbeschreibungen - sowie die Lösungsvorschläge - Zusammenarbeit, Verwendung von Standards. Wir ziehen auch die wichtigsten Lehren aus diesem Projekt, von denen wir hoffen, dass sie für andere Forscher und für andere Sprachen von Nutzen sein werden.

INDEX

Mots-clés
alsacien, langue peu dotée, langue régionale de France, occitan, picard, traitement automatique de la langues

Keywords
alsatian, low-resource language, natural language processing, occitan, picard, regional language of France

Schlagwortindex
automatische Sprachverarbeitung, Elsässisch, Okzitanisch, Picard, Regionalsprachen Frankreichs, gering ausgestattete Sprache

AUTEURS

Delphine Bernhard
Maîtresse de conférences en informatique à l'université de Strasbourg. Ses travaux de recherche se situent dans le domaine du traitement automatique des langues et concernent en particulier la constitution de ressources et d'outils pour les langues peu dotées. Elle s'intéresse également aux problématiques de l'aide à la lecture et de la simplification de textes.
IDREF : https://www.idref.fr/112578063
ORCID : http://orcid.org/0000-0001-7857-5873
HAL : https://cv.archives-ouvertes.fr/delphine-bernhard

Myriam Bras
Professeure de linguistique au département de sciences du langage du l'université Toulouse Jean Jaurès et chercheuse dans l'équipe de linguistique du laboratoire CLLE. Ses domaines de recherche vont de la sémantique du temps au discours en français et en occitan en passant par le développement de ressources pour la linguistique occitane comme la base textuelle BaTelòc et différents corpus annotés et outils de traitement automatique de l'occitan.
IDREF : https://www.idref.fr/068709862
ISNI : http://www.isni.org/0000000038746562

Anne-Laure Ligozat
Maîtresse de conférences en informatique à l'ENSIIE et effectue sa recherche au LIMSI. Son domaine de recherche est le traitement automatique des langues, et elle s'intéresse en particulier à l'extraction d'information et à la simplification de textes.
IDREF : https://www.idref.fr/112991440
ORCID : http://orcid.org/0000-0002-2188-3426
HAL : https://cv.archives-ouvertes.fr/anne-laure-ligozat
ISNI : http://www.isni.org/000000014089721X

Aleksandra Miletic
Chercheuse post-doctorale dans l’équipe de linguistique du laboratoire CLLE. Son travail se situe dans les domaines du traitement automatique des langues et de la syntaxe descriptive. Elle s’intéresse particulièrement à la création des corpus dotés d’annotations linguistiques et à leur utilisation dans la description et quantification de phénomènes syntaxiques.
IDREF : https://www.idref.fr/244418268

Jean Sibille
Chargé de recherche HDR au CNRS (Laboratoire CLLE, université Toulouse Jean-Jaurès), en linguistique occitane et romane. Il a soutenu une thèse de linguistique sur La passion de saint André, un drame religieux du XVIe siècle en dialecte occitan de Briançon. Il travaille actuellement sur différentes questions de syntaxe, de morphologie et de dialectologie occitanes.
IDREF : https://www.idref.fr/032468865

Amalia Todirascu
Professeure en linguistique de corpus et linguistique outillée à la faculté de lettres de l’université de Strasbourg et membre de l’UR 1339 LiLPa de la même université. Ses travaux de recherche s’inscrivent dans le domaine du traitement automatique des langues et de la linguistique de corpus. Elle travaille en particulier sur les problèmes d’annotation manuelle et automatique (de la coréférence), la simplification automatique et le développement des ressources pour la didactique des langues.
IDREF : https://www.idref.fr/130431796
ORCID : http://orcid.org/0000-0002-3092-3549
HAL : https://cv.archives-ouvertes.fr/amalia-todirascu

Marianne Vergez-Couret
Maîtresse de conférences en linguistique au département de sciences du langage et au laboratoire FoReLLIS de l’université de Poitiers. Elle effectue ses travaux de recherche d’une part dans le domaine de la sémantique du discours et d’autre part dans le domaine du traitement automatique des langues, en particulier pour le développement de ressources et d’outils pour la langue occitane.
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