Myriadisation de ressources linguistiques pour le traitement automatique de langues non standardisées

Alice Millour

SORBONNE UNIVERSITÉ

ÉCOLE DOCTORALE V

Laboratoire de recherche Sens Texte Informatique Histoire (STIH)

T H È S E

pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ SORBONNE UNIVERSITÉ

Discipline : Mathématiques, informatique et applications aux sciences de l'Homme

Présentée et soutenue par :

Alice MILLOUR

le 14 décembre 2020

Myriadisation de ressources linguistiques pour le traitement automatique de langues non standardisées

Sous la direction de :

Mme Karën FORT – Maîtresse de conférences, Sorbonne Université / STIH
M. Claude MONTACIÉ – Professeur, Sorbonne Université / STIH

Membres du jury :

M. Laurent BESACIER – Professeur, Laboratoire d'Informatique de Grenoble / GETALP – Rapporteur
Mme Iris ESHKOL TARAVELLA – Professeure, Université Paris Nanterre / MoDyCo – Examinatrice
Mme Karën FORT – Maîtresse de conférences, Sorbonne Université / STIH – Co-directrice
M. Claude MONTACIÉ – Professeur, Sorbonne Université / STIH – Directeur
Mme Delyth PRYS – Professeure, Bangor University / Language Technologies Unit – Examinatrice
M. Benoît SAGOT– Directeur de recherche, Inria / ALMAnaCH – Rapporteur

Remerciements

Mes premiers remerciements vont à Karën Fort. Merci d'avoir consacré tant de cœur et d'énergie à t'assurer que j'étais prête pour ce qui m'attendait. Merci, entre nombreuses et autres choses, de ne pas avoir laissé ta confiance s'ébranler, et de m'avoir fait une place parmi tes pairs. C'est précieux, et ça m'a portée.

Je remercie également Claude Montacié, qui a été le premier à m'orienter vers un parcours de recherche et dont la présence a été décisive aux moments clé de cette thèse. Merci à Barthélémy Jobert, alors président de l'université Paris-Sorbonne, de m'avoir accordé le financement qui m'a permis de réaliser cette thèse.

Je remercie mes rapporteurs, Laurent Besacier et Benoît Sagot pour leurs précieux retours et commentaires. Un grand merci également à Iris Eshkol Taravella et à Delyth Prys d'avoir accepté de prendre part au jury de cette thèse.

Je remercie ici toutes les personnes qui ont fait partie de cette aventure et qui m'ont nourrie de leurs réflexions, de leurs critiques et de leurs encouragements. Par ordre d'apparition, merci à Marie-Laure Pflanz pour ses conseils et ses livres, et à Bernard Victorri de m'avoir aidée à me préparer à l'audition ayant conduit à l'obtention de ma bourse de thèse. Merci à Franck Neveu et à Maxime Amblard de m'avoir donné la chance de présenter mon travail dans divers contextes, à Françoise Guérin pour son écoute et ses retours. Merci à Rodrigo Agerri pour son accueil à Donostia et ses conseils, et merci à Jean-Yves Antoine qui n'a eu de cesse de m'encourager et de me conseiller à chaque occasion où nous nous sommes croisés.

Ce travail a été l'occasion de collaborations, rendues notamment possibles par la DGLFLF et par Enet Collect que je remercie ici. Un grand merci à André Thibault, Delphine Bernhard, Bruno Guillaume et Nicolas Lefebvre, membres du projet PLURAL ; à Pierre Magistry pour sa patience et sa disponibilité ; et à Marianne Grace Araneta, Ivana Lazić Konjik, Liam Murray, Neasa Ní Chiaráin, Yann-Alan Pilatte et Annalisa Raffone d'avoir accepté de nous suivre, Karën et moi, dans le projet de Katana et d'en avoir nourri le développement avec leur énergie, leurs bonnes idées et leur bonne humeur.

Je remercie Gwladys Feler et Harmonie Begue de nous avoir accordé leur confiance. Je remercie également ici Fabiola Henri pour sa disponibilité.

Merci à tous les participants sans qui tout ceci n'aurait eu aucun sens. Un merci particulier à Mireille Libmann et à Jean-Noël Schàng Kempf, pour leur enthousiasme et leur disponibilité.

Ce travail doit beaucoup à mon entourage, souvent confiant et toujours bienveillant.

En particulier, je remercie la gaie équipe de Serpente : Gaël Lejeune, Morgan Kitzmann, Margot Déage, Hugo Jeanningros, Hugo Touzet, Victor Coutolleau, Jean-Baptiste Tanguy pour leur présence, en 206, au bar et ailleurs. Merci à Yoann Dupont d'avoir bien voulu consacrer le temps nécessaire à la résolution de mes soucis LaTeX quand ça ne m'amusait plus assez. Enfin, merci à Shehrazad Lakaf pour sa présence rassurante à des moments déterminants de mon parcours à la Sorbonne.

Merci à Émilie Paillous de m'avoir ouvert en douceur les portes du développement Web et merci à Laura Maréchal d'avoir été la gardienne de mes pages pendant la rédaction.

Je remercie qui de droit de m'avoir permis de croiser les routes de Xavier Pothron, de Léa Lévy, de Pauline Roeser, de Théo Trouillon, et de Franck de Goër.

Merci à Grégoire Rialan et à Nina Cooper pour leur amitié sans faille.

Merci à Ariane d'être inclassable et ma sœur.

Merci à mes frères, Julien et Pierre, et à mes parents, Claire et Christian. Merci d'applaudir mes réussites et de consoler mes échecs avec l'intelligence et la tendresse qui vous caractérisent.

Andy, tu m'as accompagnée à chaque étape de cette aventure. Te lo agradezco.

Table des matières

Table des figures

Liste des tableaux

Introduction

L'explosion des usages numériques représente à la fois une menace et une opportunité pour la diversité linguistique. Le traitement automatique des langues (TAL) joue naturellement un rôle dans l'accompagnement des communautés linguistiques envers l'utilisation de leurs langues sur Internet et dans le monde numérique en général. En particulier, assurer la présence d'une langue dans l'univers technologique ne peut s'envisager sans le développement d'outils variés correspondant à des pratiques numériques sans cesse renouvelées, qu'il s'agisse par exemple de claviers de saisie, de moteurs de recherche à la hauteur des attentes actuelles ou encore de moteurs de reconnaissance vocale.

Si la diversité linguistique du monde numérique n'est pas représentative de la diversité de ses usagers c'est notamment que les ressources textuelles linguistiques requises par de tels développements sont coûteuses à plusieurs égards. Dans cette thèse, nous explorons les possibilités offertes par la production participative, ou *myriadisation*, pour permettre le développement de ressources linguistiques numériques pérennes pour toute langue susceptible d'en bénéficier.

En particulier, nous montrons que dans le cas de langues non standardisées, c'est-à-dire dont l'écriture ne suit pas de norme consensuelle, seule la mise en place d'un dialogue (au sens large) avec les locuteurs permet d'envisager le développement d'outils correspondant à la pratique réelle de leurs utilisateurs finaux.

Dans la suite de cette introduction, nous présentons d'abord les enjeux nous ayant poussée à mener cette recherche. Puis nous présentons le cadre dans lequel elle a été menée : dans un premier temps nous décrivons le contexte formel de la thèse et nous explicitons comment celui-ci a conditionné nos choix quant aux communautés linguistiques avec lesquelles nous avons expérimenté la myriadisation. Dans un second temps, nous présentons les contraintes méthodologiques que nous nous sommes imposée. La troisième partie de cette introduction est consacrée à l'organisation du corps du présent document.

Rôles du TAL et des ressources langagières

La recherche en traitement automatique des langues a émergé pour combler un besoin, celui d'un support technique à deux activités effectuées manuellement jusqu'alors : d'une part l'étude fondamentale des langues et des mécanismes cognitifs associés, et d'autre part divers traitements pouvant être opérés sur celles-ci. L'ambition initiale ayant motivé les premiers travaux de traitement automatique du langage, dont les balbutiements datent de 1954, fut d'automatiser la traduction d'une langue source vers une langue cible. Les recherches poursuivant cet objectif ont conduit au développement de nombreuses « technologies du langage » qui accompagnent les besoins concomitants des nouveaux usages du numérique. Aujourd'hui, les pratiques évoluant ont érigé l'existence d'outils numériques au rang de nécessité pour un nombre croissant de communautés linguistiques connectées. L'impact du traitement automatique des langues a donc progressé : de discipline permettant d'automatiser des activités humaines, elle est aujourd'hui devenue un champ de recherche dont les systèmes qui en découlent peuvent favoriser (ou non) l'expression de communautés linguistiques particulières.

La rapidité d'écriture permise par les claviers de saisie, associée à la quasi-instantanéité des transmissions, ont conduit à l'apparition d'une nouvelle catégorie de matériau linguistique : la conversation médiée par ordinateur (en anglais, *computer-mediated conversation*), en particulier la conversation *écrite*. La part prise par le numérique dans notre communication a démocratisé de

nouvelles pratiques venues bousculer le genre textuel en y introduisant des marques de l'oralité, en le contraignant à des formats divers, depuis le « langage texto » aux *tweets*, ou en passant par l'utilisation de signes propres à l'écriture numérique comme les émoticônes ou les mots-dièse.

Dans le cas des langues standardisées, la conversation médiée par ordinateur fait ainsi émerger de nouvelles « variantes » linguistiques pouvant notamment s'écarter des normes typographiques et orthographiques prescrites par les conventions d'écriture en vigueur en contexte formel. Néanmoins, les langues majoritaires ne sont pas les seules à être concernées par l'évolution des usages. En effet, nombre de langues, notamment des langues dont la tradition restait jusqu'alors principalement orale et n'ayant pas fait l'objet d'une standardisation graphique, voient leurs pratiques scripturales se démocratiser. C'est à ces nouvelles pratiques, à ce qu'elles permettent et à ce qu'elles appellent en termes d'effort de traitement linguistique que nous nous sommes intéressée dans le cadre de cette thèse.

La question de l'utilisation du TAL pour accompagner une pratique linguistique dans le monde numérique revient à poser deux questions principales. La première est celle de **la nature et de la disponibilité des ressources linguistiques existantes** pour celle-ci : à l'exception des systèmes reposant sur l'existence de règles qui ont prédominé jusqu'à l'apparition des méthodes par apprentissage, tous les outils de traitement en TAL reposent aujourd'hui sur l'existence de ressources textuelles. Notons dès à présent que seule l'existence de ressources pérennes permet d'envisager le développement d'usages durables.

La seconde est l'**adaptabilité des méthodes** ayant été imaginées pour des langues standardisées à des pratiques linguistiques qui ne le sont pas : comment, par exemple, poser la question de la *représentativité* des ressources dans un cadre où, l'orthographe n'étant pas fixée, des dizaines de graphies concurrentes peuvent coexister ? Sous quelles conditions les étapes de conception, d'implémentation, et d'évaluation de méthodes ayant fait leur preuve sur des langues standardisées peuvent-elle être ainsi transposées ?

Les outils par apprentissage sont très largement conçus comme agnostiques vis-à-vis des langues. Cela signifie que l'existence de ressources langagières suffisantes est *a priori* le seul obstacle à l'instanciation d'un outil fonctionnel pour une nouvelle langue. La construction de ressources langagières nécessaires telles que les lexiques ou corpus annotés étant une tâche peu automatisable, et requérant l'implication parfois prolongée d'intervenants aux compétences diverses, cet obstacle est de taille.

Nombre de langues, dont les locuteurs sont pourtant aussi des internautes, ne peuvent en effet pas prétendre à des financements suffisants, ni compter sur la disponibilité d'experts pour assurer le développement des ressources langagières. En revanche, leur présence sur Internet signifie qu'il est possible de rentrer en contact avec les communautés linguistiques concernées. En leur proposant un modèle permettant de les mettre à contribution, il devient donc possible d'exploiter leurs connaissances linguistiques afin de palier le manque de moyens traditionnels.

La production participative *via* Internet est en effet une solution ayant fait ses preuves, notamment pour l'anglais (Poesio *et al.*, 2013) ou le français (Guillaume *et al.*, 2016). Une des clés principales du succès de telles entreprises est de parvenir à motiver un nombre suffisant de participants pour assurer une quantité et une qualité de données satisfaisantes.

À nouveau, la transposition d'une telle pratique à une langue ne présentant pas les mêmes avantages en termes de nombre de locuteurs, de rémunération possible ou d'uniformité des pratiques linguistiques n'est pas immédiate. Nous avons néanmoins formulé les deux hypothèses

suivantes :

1. Il n'y a pas de raison que le succès d'une entreprise participative (en termes de qualité des ressources produites) dépende de la langue à laquelle elle est appliquée.
2. Concernant la quantité de locuteurs à mobiliser, la motivation de ceux-ci quant à l'urgence de disposer de ressources et d'outils adaptés suffit à compenser un nombre de locuteurs moindre.

Afin de tester ces hypothèses, nous avons mis en place plusieurs expériences de myriadisation de ressources linguistiques. Après avoir expérimenté avec l'annotation participative en parties du discours pour l'alsacien et le créole guadeloupéen sur des corpus existants, nous avons mis en place une plateforme de collecte de corpus bruts, d'annotations et de variantes graphiques, instanciée quant à elle pour l'alsacien et le créole mauricien. Ces ressources ont été évaluées et utilisées par la suite pour le développement d'outils d'annotation automatique en parties du discours pour ces langues.

Cadre de la thèse

L'encadrement de cette thèse a été réalisé par K. Fort et C. Montacié au sein du laboratoire Sens Texte Informatique Histoire [1] (STIH) de Sorbonne Université.

Contexte formel et choix linguistiques

Les enjeux que nous nous sommes fixés nous ont poussée à mener un travail pluridisciplinaire et à dépasser le cadre du traitement automatique des langues au sens strict.

D'abord, en impliquant les locuteurs dans la création collaborative de ressources linguistiques, nous nous rapprochons de la linguistique de terrain. Néanmoins, notre objectif n'est pas celui de la documentation linguistique mais bien de la construction de ressources pour le TAL dans un contexte de ressources initiales minimal. Nous ne prétendons pas suivre une méthodologie permettant d'atteindre la qualité linguistique attendue par la linguistique de terrain. En revanche, nous proposons une méthodologie adaptable à n'importe quelle langue à moindre coût et permettant la production de ressources de qualité raisonnable.

Ensuite, nous nous sommes heurtée à des problématiques propres aux sciences de l'information et de la communication, notamment lors de la conception et de la promotion de notre interface de collecte de ressources.

Enfin, nous avons menée une enquête socio-linguistique concernant le positionnement des locuteurs vis-à-vis de l'utilisation de leur(s) langue(s) sur Internet.

Nous l'avons dit, la méthodologie que nous avons développée est conçue pour être indépendante de la langue à laquelle elle est appliquée. Le choix des langues sur lesquelles nous avons travaillé, et notamment celui des langues de France dans un premier temps, a été conditionné par la nature des financements dont cette thèse a fait l'objet : le financement de thèse obtenu a Sorbonne Université, et le financement obtenu dans le cadre de l'appel *Langues et Numérique 2018* proposé

1. Voir : http://stih-sorbonne-universite.fr/.

par la délégation générale à la langue française et aux langues de France[2] (DGLFLF).

En effet, le présent travail a été réalisé dans le laboratoire STIH de Sorbonne Université, et au moment où cette thèse a débuté, celui n'abritait que des recherches sur le français. Notre projet de thèse ayant une prétention multilingue, nous avons eu la possibilité de proposer une recherche sur les langues régionales de France. Il est à noter que la DGLFLF définit les langues de France comme étant « [...] les langues régionales ou minoritaires parlées par des citoyens français sur le territoire de la République depuis assez longtemps pour faire partie du patrimoine culturel national, et ne sont langue officielle d'aucun État ». La France étant le lieu d'une diversité culturelle dont l'étendue ne se limite pas au territoire métropolitain, ces langues sont classées en trois sous-catégories : les langues régionales, les langues non-territoriales et les langues des Outre-mer. Par conséquent, des langues aux caractéristiques aussi diverses que le picard, le berbère et le tahitien sont estampillées « Langues de France » par le Ministère de la Culture. Étant donné la nature participative de notre projet, seules les langues disposant d'une communauté de locuteurs connectée étaient candidates.

Sachant ces contraintes, la disponibilité en 2016 de Delphine Bernhard et de Lucie Steiblé, chercheuses au LiLPa de Strasbourg travaillant à la construction de ressources pour l'alsacien a orienté notre choix vers l'alsacien dans un premier temps. L'instanciation de notre méthodologie pour une nouvelle langue requérant la collaboration avec un(e) locuteur(trice) de la langue, les adaptations ultérieures aux créoles guadeloupéen puis mauricien ont été rendues possibles par le co-encadrement avec Karën Fort des mémoires de Master 1 de Gwladys Feler en 2017 et d'Harmonie Begue en 2019.

Enfin, travailler sur ces trois langues, parlées principalement par des communautés linguistiques bilingues avec le français, nous a permis de le choisir comme langue neutre des interfaces développées et ainsi de garantir que nos plateformes ne soient pas le lieu de la promotion d'une variante dialectale ou d'une convention graphique spécifique pour aucune des langues concernées par nos travaux.

Contraintes méthodologiques

Afin de tester les hypothèses présentées ci-dessus nous avons développé une méthodologie pour la production participative bénévole de ressources linguistiques pérennes conçue comme réplicable à toute langue susceptible d'en bénéficier.

Cet objectif nous a amenée à travailler sous un certain nombre de contraintes qu'il est possible, bien qu'ils ne soient pas strictement distincts, de regrouper dans trois ensembles principaux : le premier concerne la pérennité des ressources produites en utilisant notre méthodologie (C1). Le second concerne la réplicabilité de celle-ci (C2), et le troisième sa nature participative et bénévole (C3).

C1 : Pérennité des ressources produites

Un des objectifs de notre méthodologie est de (faire) produire des ressources pérennes, ou *durables*, c'est-à-dire pouvant être redistribuées et réutilisées à l'avenir. Cela nous oblige notamment à redistribuer sous des licences claires tout ce qui est produit par à travers la méthodologie que

2. Voir : https://www.culture.gouv.fr/Sites-thematiques/Langue-francaise-et-langues-de-France.

nous développons. De ce fait, nous nous sommes interdit d'utiliser des ressources qui n'étaient pas libres de droit, qu'elles soient brutes ou annotées. Cela exclut par exemple l'utilisation d'un vaste ensemble de contenus textuels facilement accessibles mais en réalité indisponibles à l'usage, comme les œuvres littéraires sous licence ou les contenus produits sur certains réseaux sociaux comme Facebook.

C2 : Réplicabilité de la méthodologie

La deuxième contrainte que nous nous sommes imposée consiste à proposer une méthodologie qui soit réplicable à toute langue candidate, c'est-à-dire à toute langue qui dispose d'une communauté de locuteurs connectée. L'absence de toute autre contrainte constitue, en creux, une contrainte forte. D'une part, cela nous a poussée à nous confronter au cas des langues présentes sur Internet mais dont les variations dialectales et graphiques ne sont pas lissées par une orthographe commune aux différentes variantes. D'autre part cela nous a menée à faire des choix d'implémentation les plus agnostiques de la langue considérée possible.

Nous avons notamment :

- tâché de limiter au maximum la dépendance à des caractéristiques propres à la langue considérée, par exemple sa parenté avec une autre langue, ou la disponibilité préalable de telle ou telle ressource,
- fait des choix linguistiques le plus agnostiques possibles, notamment le choix du jeu d'étiquettes proposé par le projet UD (Petrov *et al.*, 2012) pour la tâche d'annotation en parties du discours, qui est reconnu par la communauté et dont la complexité est stable selon les langues,
- proposé une méthodologie qui ne nécessite pas un engagement permanent de spécialistes de la langue considérée, ce qui nous conduit naturellement à la contrainte C3.

C3 : Myriadisation bénévole

La production de ressources linguistiques par myriadisation auprès de locuteurs qui ne sont pas des professionnels de la langue nous a contrainte à des choix de conception permettant de les guider vers la réalisation de ces tâches inhabituelles. Certains de ces choix ont été de nature à alourdir les développements informatiques des différentes plateformes de myriadisation.

D'abord, nous avons été menée à nous questionner sur les conditions assurant la faisabilité des tâches proposées au regard de la compétence des annotateurs. Cela nous a conduit à former les participants d'une part et à réduire la complexité des tâches proposées d'autre part. Par exemple, la complexité d'une tâche de catégorisation n'est pas la même selon qu'on propose aux participants de choisir entre toutes les catégories possibles, entre un nombre réduit d'options probables, ou qu'on leur demande de valider ou d'invalider une option spécifique.

Par ailleurs, à la différence d'un cadre où les annotateurs ont été choisis, nous n'avons pas de connaissance *a priori* de la compétence des participants. Cela nous a menée à mettre en place des procédures de contrôle pour assurer la qualité des données, comme une évaluation dynamique des participants et une agrégation des données conditionnelle à la confiance accordée à chaque annotateur.

Enfin, le recrutement des participants et le maintien de leur motivation passe par un travail sur le *design* de l'environnement de myriadisation qui se doit d'être plaisant et intuitif.

À la frontière entre C2 et C3 :

Dans le contexte d'un travail sur des langues ne pouvant être rattachées à un standard consensuel, nous devons proposer aux locuteurs impliqués des textes avec lesquels ils sont à l'aise, dans le sens où ils sont conformes à leurs pratiques de leur langue. Il est apparu au cours des expériences d'annotation que nous avons menées que les locuteurs pouvaient en effet être gênés voire découragés lorsque les textes à annoter n'étaient pas proposés dans « leur variante », qu'elle soit dialectale ou graphique.

Structure du document

Nous consacrons le premier chapitre de notre état de l'art (partie I) à une présentation des diversités linguistiques, notamment celles qui peuvent être observées dans les technologies de l'information et de la communication et dans le domaine du TAL. Le second chapitre concerne les méthodes de production des corpus bruts et des corpus annotés nécessaires au développement d'un grand nombre d'outils de traitement automatique. Le coût de production de ressources de qualité nous amène au troisième chapitre, où nous présentons les opportunités offertes par la myriadisation à cet égard.

Nous décrivons dans une seconde partie (partie II) nos expériences de myriadisation de ressources linguistiques pour des langues non standardisées. Après avoir présenté les communautés linguistiques avec lesquelles nous avons travaillé, nous décrivons les trois tâches de myriadisation instanciées : la production d'annotation en parties du discours, de ressources textuelles brutes, et de lexiques de variantes graphiques.

Enfin, les ressources recueillies sont présentées une troisième partie (partie III). Nous en proposons une évaluation intrinsèque puis extrinsèque à travers la description de nos expériences d'annotation automatique. Nous présentons en particulier une expérience permettant d'exploiter les variantes graphiques myriadisées. La méthodologie décrite permet d'améliorer les performances d'un outil d'annotation tout en conduisant à la génération automatique de paires de variantes graphiques.

Première partie

État de l'art

Introduction

Afin d'aborder la problématique du traitement automatisé de langues non standardisées, nous proposons un état de l'art qui vise à ancrer notre travail dans les enjeux propres aux trois champs de recherche que sont : (i) l'étude de la diversité linguistique, (ii) la production et l'exploitation de ressources linguistiques et (iii) la myriadisation.

Plus précisément, notre objectif est de montrer les relations existant entre ces trois champs et le domaine du TAL en répondant aux questions suivantes : « quel sont les liens entretenus par la diversité linguistique et le TAL ? », « comment le caractère nécessaire des ressources linguistiques s'inscrit-il dans une ambition plurilingue du TAL ? », et enfin « quelles sont les formes prises par la myriadisation et comment celles-ci peuvent-elles être mises à profit pour accompagner cette ambition plurilingue ? ».

D'abord, il nous paraît important de situer notre recherche dans le contexte linguistique actuel. Les enjeux liés à la diversité linguistique que nous décrivons ont contribué à motiver notre recherche. La promotion, ou non, de la diversité linguistique est un sujet politique pouvant être débattu, mais nous nous attachons ici à proposer un aperçu le plus objectif possible de la diversité effectivement observable. Une partie du TAL étant consacrée à l'outillage numérique des langues, il convient de comprendre quelles sont les communautés et pratiques linguistiques effectivement concernées par le développement de ressources et d'outils de traitement.

Observer la diversité linguistique dans différents environnements permet de mettre au jour un déséquilibre entre les pratiques des locuteurs dans le monde réel et leurs pratiques numériques. Celui-ci n'est pas surprenant, et les raisons qui poussent une communauté linguistique à préférer une langue plutôt qu'une autre dans ses usages numériques sont variées. On peut toutefois s'interroger sur le rôle joué par la recherche en TAL, comprise ici comme discipline permettant le développement d'outils propres à informatiser les langues, dans le maintien de ce déséquilibre.

L'informatisation d'une nouvelle langue est une problématique complexe, les pratiques et les caractéristiques linguistiques étant nombreuses et confrontant les chercheurs à des difficultés variées. Quelle que soit la langue concernée, le développement d'outils de traitement ne peut s'envisager sans l'existence de ressources linguistiques prises au sens large, les techniques développées reposant sur l'utilisation de ressources de types variés, brutes ou annotées, monolingues ou multilingues, manufacturées par des experts ou extraites automatiquement, et pouvant être combinées. Si les stratégies visant à atténuer la dépendance des systèmes à des ressources coûteuses sont nombreuses, la disponibilité de ressources linguistiques minimales reste nécessaire, tout au moins pour pouvoir évaluer les performances des systèmes développés.

Or, l'obtention ou la construction de telles ressources représente une première difficulté. Celle-ci est d'autant plus accrue lorsque l'on constate que les usages numériques ont permis la naissance

de pratiques linguistiques nouvelles, voire inédites à l'écrit pour certaines langues. L'intégration de ces contenus aux processus de TAL appelle des traitements spécifiques et représente un enjeu nouveau pour le domaine.

Pour pallier le manque de ressources dont pâtissent également les langues « majoritaires », terme par lequel nous désignons les langues les plus présentes sur Internet aujourd'hui, la myriadisation est une pratique courante : elle est invoquée comme une méthodologie permettant de tirer parti de l'intelligence collective des locuteurs pour leur faire produire des ressources linguistiques utiles pour le TAL.

L'état de l'art que nous proposons est ainsi constitué de trois chapitres. Le premier (chapitre 1) est consacré à l'observation de la diversité linguistique dans trois sphères : parmi les locuteurs, dans les technologies de l'information et de la communication, et dans le domaine du TAL. Ce chapitre, au cours duquel nous présentons la diversité des pratiques linguistiques observées, nous amène à la question des ressources nécessaires pour les intégrer aux processus de TAL.

Le second chapitre (chapitre 2) leur est consacré : nous y présentons le caractère non trivial de la construction de ressources linguistiques pour le TAL, en nous focalisant sur les ressources textuelles que constituent les corpus bruts et annotés. Nous y proposons notamment une présentation critique des sources de corpus textuels usuelles, tant d'un point de vue de leur pérennité que de leur représentativité. Ces considérations nous permettent d'aborder d'emblée les difficultés posées par les langues non standardisées.

Enfin, le troisième volet de notre état de l'art (chapitre 3) est un tour d'horizon des différents travaux faisant usage de la myriadisation. Celui-ci nous permet d'illustrer que la myriadisation peut être conçue d'une part comme un palliatif au manque d'« experts linguistiques », et d'autre part comme le moyen dans certains cas d'accéder à une connaissance linguistique nécessaire et détenue par les seuls locuteurs.

Chapitre 1

Diversité(s) linguistique(s)

Sommaire

Le choix de travailler au développement de ressources et d'outils pour certaines langues plutôt que d'autres n'est pas anodin. Nous consacrons donc ce premier chapitre à ancrer notre recherche au regard des enjeux liés à la diversité linguistique dans lesquels elle s'inscrit.

Nous envisageons dans ce chapitre trois sphères d'observation de la diversité linguistique : la première, qu'on qualifie ici de « réelle », est celle des locuteurs, la seconde décrit la diversité dans les technologies de l'information et de la communication (TIC), que nous approximons ici à la diversité linguistique des *internautes*. Enfin, nous envisageons la diversité linguistique dans le domaine du traitement automatique des langues en particulier à travers le terme de « langue peu dotée ».

Nous ne proposons pas ici d'argumentaire détaillé en faveur de la protection de la diversité linguistique. Nous recommandons sur ce sujet la lecture de *Halte à la mort des langues* que Hagège (2000) conclut en évoquant les opportunités offertes par Internet, qu'il décrit comme « un support qui donne une nouvelle voix, même dans le mirage du virtuel, à des idiomes qu'on risquait de ne plus entendre » où « se multiplient les échanges au moyen [...] de ces langues » (yiddish, langues régionales de France etc.) ». Nous recommandons également l'ouvrage *Net.lang : Towards the Multilingual Cyberspace* (Maaya, 2012), produit par le réseau MAAYA [3], qui propose un ensemble de communications en faveur de la diversité culturelle et du multilinguisme dans les technologies de l'information et de la communication.

3. Voir : http://www.maaya.org/, juin 2020.

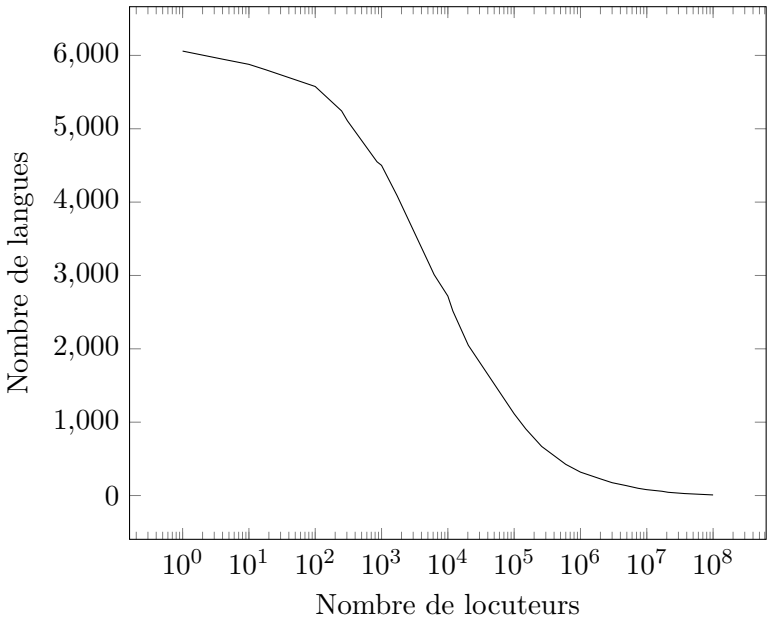


FIGURE 1.1 – Répartition des langues par nombre de locuteurs [6].

Le contenu de ce chapitre a fait l'objet d'une présentation invitée à Nancy dans le cadre du Séminaire Cognition & Langage organisé par M. Amblard et M. Rebuschi (Millour, 2019b).

1.1 Diversité linguistique « réelle »

Dans cette section, nous utilisons les chiffres diffusés par ETHNOLOGUE, LANGUAGES OF THE WORLD (Eberhard *et al.*, 2019), souvent abrégé ETHNOLOGUE [4] et ceux disponibles sur Wikipédia [5]. S'ils peuvent être contestés et présentent nécessairement un caractère approximatif (voir, notamment, (Paolillo et Das, 2006) pour une étude critique détaillée des méthodes de décompte des langues et de leurs locuteurs), ces chiffres permettent de dresser les contours de la diversité linguistique.

On estime à environ 7 000 le nombre de langues maternelles parlées quotidiennement. Cette diversité apparente est à mettre en regard avec le nombre de locuteurs pour chacune de ces langues. La figure 1.1 montre la répartition du nombre de langues parlées par un nombre de locuteurs donné. En effet, seules 2 à 3 % d'entre elles sont parlées par plus d'un million de locuteurs, 1 % par plus de 10 millions de locuteurs et moins d'une vingtaine par plus de 100 millions de locuteurs.

L'*Atlas UNESCO des langues en danger dans le monde* (Moseley, 2010) fait en outre état de 2 346 langues dont le niveau de vitalité est classé de « vulnérable » à « en situation critique », cet état étant établi en fonction du degré de transmission de la langue d'une génération à l'autre.

4. ETHNOLOGUE est une publication en ligne qui inventorie les langues du monde. Ce catalogue est développé par SIL INTERNATIONAL, une organisation non gouvernementale confessionnelle, voir : https://www.sil.org/about.

5. Voir : https://en.wikipedia.org/wiki/List_of_languages_by_total_number_of_speakers, juin 2020.

6. Ce graphique a été reconstitué à partir des chiffres de l'édition d'Ethnologue de 1999, donnés dans le cours `LING001: Introduction to Linguistics` de M. Libermann.

D'après Leclerc (2020),

> « on peut dire qu'une langue est menacée dans sa survie dès qu'elle n'est plus en état d'expansion, dès qu'elle perd de ses fonctions de communication dans la vie sociale ou qu'elle n'est plus pratiquée quotidiennement pour les besoins usuels de la vie, dès qu'elle n'est plus rentable au plan économique, ou dès qu'il n'y a plus suffisamment de locuteurs pour en assurer la diffusion ».

Or, un nombre croissant de communications et d'accès à des services quotidiens passe aujourd'hui par le monde numérique. Par conséquent, les technologies de l'information et de la communication (TIC) jouent un rôle dans le processus d'expansion ou d'amenuisement de l'usage des langues, et l'absence d'une langue donnée du monde numérique contribue directement à la rendre vulnérable.

Notons que l'absence de scripturalisation d'une langue n'est pas un critère pour la considérer comme « en danger ». Inversement, l'existence d'un système graphique pour une langue donnée n'est pas une garantie de sa survie à long terme (Karan, 2014).

1.2 Diversité linguistique dans les technologies de l'information et de la communication (TIC)

Les « technologies de l'information et de la communication » (TIC) sont définies comme suit par l'Institut de statistique de l'UNESCO [7] :

> « Ensemble d'outils et de ressources technologiques permettant de transmettre, enregistrer, créer, partager ou échanger des informations, notamment les ordinateurs, l'internet (sites Web, blogs et messagerie électronique), les technologies et appareils de diffusion en direct (radio, télévision et diffusion sur l'internet) et en différé (podcast, lecteurs audio et vidéo et supports d'enregistrement) et la téléphonie (fixe ou mobile, satellite, visioconférence, etc.). »

Dans la suite de cette section, nous approximons la diversité linguistique « dans les TIC » à la diversité linguistique « sur Internet », qui constitue un observatoire commode des pratiques linguistiques numériques, en particulier à travers les nombreux rapports produits par W3Techs [8].

1.2.1 La représentativité linguistique des TIC

Si la diversité linguistique globale s'amenuise, la diversité linguistique en termes de nombre de langues présentes sur Internet tend pour sa part à s'accroître (Oustinoff, 2012). En effet, la démocratisation des infrastructures nécessaires et l'explosion des communications numériques conduisent un nombre croissant de communautés linguistiques à définir de nouveaux espaces d'expression sur Internet (voir par exemple les cas décrits par Rivron (2012) et Soria *et al.* (2018)). Toutes les langues ne sont pas concernées par ces développements, mais il y a urgence à « informatiser » un nombre croissant de langues, pour reprendre le terme proposé par Berment (2004), c'est-à-dire à les doter d'un ensemble de services (par exemple, des outils de saisie,

7. Voir : http://uis.unesco.org/fr/glossary, juin 2020.
8. Voir : https://w3techs.com/, juin 2020.

de sélection et de correction de textes, de traduction etc.) et de ressources à destination des locuteurs tels que des dictionnaires.

D'après les chiffres fournis par Internet World Stats [9], la couverture Internet concerne 54,5 % des êtres humains soit 4,2 milliards d'internautes. Les 10 langues les plus représentées sur Internet (l'anglais, le mandarin, l'espagnol, l'arabe, le portugais, l'indonésien, le français, le japonais, le russe et l'allemand) y concentreraient 77,2 % des internautes. Notons que le taux de croissance du nombre d'internautes entre 2000 et 2018 a été de +1 091,9 % pour ces dix langues et de +935,8 % pour les autres langues.

Il n'existe pas de décompte précis du nombre de langues présentes aujourd'hui sur Internet. Wikipédia fait état de 299 projets actifs. D'après W3Techs, 200 langues sont représentées dans le « top 10 millions de sites internet », 160 d'entre eux comptant pour moins de 0,1 % de ces sites.

Les corpus ayant été extraits du Web donnent un aperçu intéressant de la diversité linguistique qui y règne : le projet Oscar (Ortiz Suárez *et al.*, 2019) distribue par exemple des corpus issus d'un filtrage du projet Common Crawl [10] pour 166 langues. Le projet An Crúbadán - Corpus Building for Minority Languages permet quant à lui de télécharger des corpus issus du Web pour 2 228 langues.

Le paysage linguistique d'Internet n'est donc pas uniforme. Néanmoins, Internet est loin de représenter la diversité linguistique de ses internautes, comme illustré par la figure 1.2 qui met en regard la proportion des langues des contenus présents sur Internet et la proportion de locuteurs natifs de ces langues [11].

1.2.2 L'oralité à l'écrit sur Internet

Internet est également devenu le berceau de pratiques scripturales nouvelles, notamment pour des communautés linguistiques dont la pratique était jusqu'alors principalement orale et qui se sont emparé de ces espaces d'expression *écrite* (van Esch *et al.*, 2019). Le passage à l'écrit se fait dans certains cas indépendamment des conventions scripturales existantes, voire sans qu'aucune convention n'ait été définie. En effet, bien que l'UNESCO préconise l'élaboration d'un certain nombre de ressources linguistiques (dont une orthographe et un système d'écriture, une grammaire écrite, un dictionnaire et une transcription phonétique) comme condition préalable à la présence pérenne d'une langue dans le cyber-espace (Diki-Kidiri, 2007), l'absence de norme orthographique n'y empêche pas leur usage à l'écrit.

Lorsque aucune orthographe n'a été définie pour une langue, ou lorsqu'une convention existe mais qu'elle n'est pas communément acceptée par une communauté linguistique, l'orthographe est instable et peut varier d'un locuteur à l'autre. Cela conduit à un phénomène de *polymorphisme graphique* [12] : la standardisation orthographique, en tant que définition d'une norme, agit comme

9. Voir : https://www.internetworldstats.com/, juin 2020.
10. Voir : https://commoncrawl.org/, juin 2020.
11. Ce graphique a été réalisé à partir des chiffres publiés par W3Tech (voir : https://w3techs.com/technologies/history_overview/content_language) concernant les langues des contenus et par Wikipédia concernant les nombres de locuteurs natifs.
12. Cette notion provient de la linguistique et est invoquée pour décrire les multiples graphies coexistant dans des états de langues antérieurs à la standardisation du français.

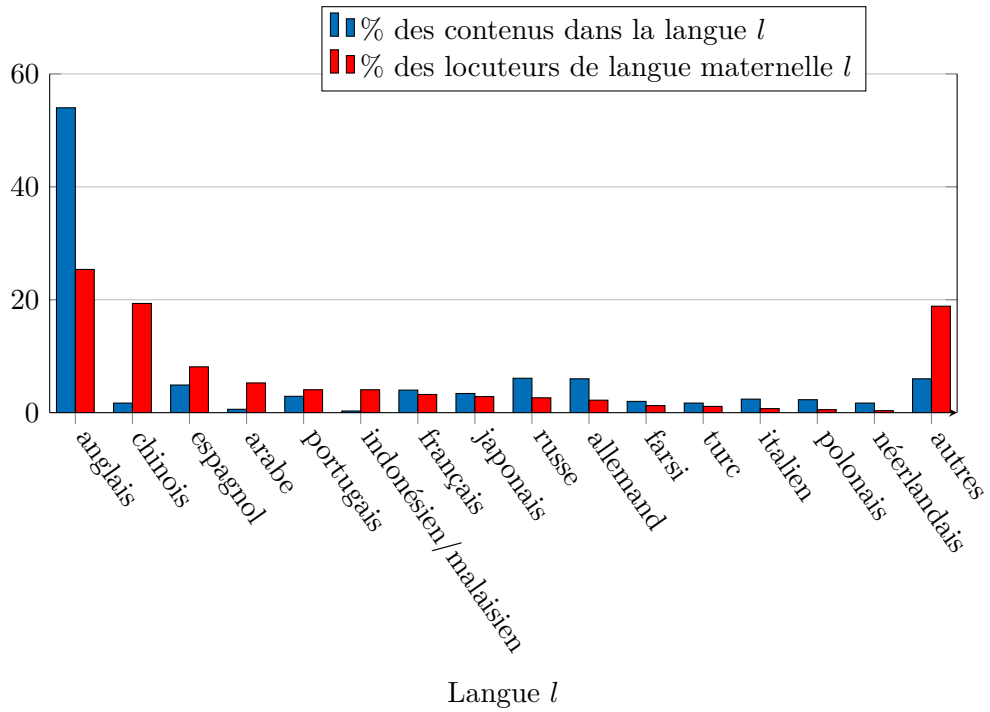


FIGURE 1.2 – Proportion des langues des contenus présents sur Internet par rapport à la proportion de locuteurs natifs de ces langues[11].

un processus d'unification de variantes linguistiques potentielles pouvant coexister à l'oral en leur attribuant une unique représentation scripturale (voir section 4.3.1).

À l'inverse, l'absence d'un tel processus de standardisation autorise la transcription différenciée de variantes pouvant être associées à une langue. La transcription de celle-ci dépend alors de la variante d'expression du locuteur, et de son bagage linguistique (langues parlées et écrites, habitudes scripturales, convention orthographique suivie, connaissance ou non de différents alphabets etc.).

En cela, Internet, en particulier dans sa nature conversationnelle, est devenu un terrain fertile pour l'expression et l'observation de la variation linguistique.

On trouve en effet dans la littérature récente de nombreux travaux traitant de l'observation sur Internet de tels phénomènes de polymorphisme graphique, où coexistent des variantes graphiques au sein d'une communauté linguistique donnée. Les contextes linguistiques correspondants sont très variés notamment en termes de nombre de locuteurs, de statut de la langue ou de vivacité. Notons par exemple les cas :

- des communautés Zapotec et Chatino a Oaxaca, Mexique, détaillé par Lillehaugen (2016) dans le cadre du programme *Voces del Valle*. Lors de ce programme, les locuteurs ont été encouragés a écrire des *tweets* dans leurs langues. Certaines conventions scripturales leur ont été proposées mais signalées comme non obligatoires. L'observation faite par l'auteur est la suivante : « Le résultat fut que les choix orthographiques effectués par la plupart des locuteurs n'étaient pas stables–mais ils écrivaient » [13] ;

13. « *The result was that, for the most part, the writers were non-systematic in their spelling decisions—but*

- de certaines communautés tibétaines hors de Chine, qui développent une « forme écrite basée sur la langue orale » [14], et ce indépendamment de l'orthographe préexistante propre au tibétain littéraire classique (Tournadre, 2014) ;
- du groupe ethnique éton, au Cameroun, au sujet duquel Rivron (2012) observe qu'Internet est le support d'une « extension de l'usage d'une langue dite « maternelle » hors des contextes et registres habituels, associé au développement de la pratique de son écriture. ». Il ajoute : « Il semble y avoir moins de gêne à écrire l'éton dans l'« entre-soi » des groupes Facebook, l'intercompréhension étant postulée malgré l'absence de graphie usuelle codifiée et enseignée » ;
- de locuteurs de dialectes javanais qui « ont leur propre manière d'écrire les mots qu'ils utilisent en fonction de la prononciation qu'ils comprennent » [15], indépendamment de l'orthographe officielle. Chaque dialecte développant une graphie qui lui est propre, un dialecte qui était « à l'origine seulement identifiable à travers les discours parlés (prononciation), est maintenant facilement reconnaissable par son orthographe sur les réseaux sociaux » [16] (Fauzi et Puspitorini, 2018) ;
- de locuteurs de langues régionales telles que l'alsacien, continuum de dialectes alémaniques pour lesquels on observe une grande diversité d'orthographes concurrentes, et ce malgré l'existence d'un système orthographique flexible, l'ORTHAL (Crévenat-Werner et Zeidler, 2008) ;

Un autre cas de polymorphisme graphique peut être observé lorsque l'alphabet usuel n'est pas connu des internautes ou peu aisé à utiliser dans les TIC. Ces cas sont notamment rapportés concernant :

- les communautés transcrivant l'arabe en alphabet latin : comme observé par Tobaili *et al.* (2019), l'arabizi permet des correspondances multiples entre les alphabets arabes et les caractères alphanumériques latins, ce qui révèle les variations dialectales habituellement dissimulées par l'écriture traditionnelle de l'arabe ;
- les communautés qui utilisent l'alphabet latin pour transcrire certains dialectes indiens sans qu'aucune règle de translittération systématique n'ait été définie (Shekhar *et al.*, 2018).

Nous nous référerons à ces langues aux écritures protéiformes comme des langues *multi-variantes*. La variation observée est le résultat de deux mécanismes pouvant se cumuler : la variation « dialectale » et la variation « scripturale ».

D'un point de vue du traitement automatique, ces productions linguistiques représentent un défi. En effet, elles nous obligent à nous confronter d'emblée aux difficultés posées par la variation. En particulier lorsque ces productions multi-variantes représentent la majorité des contenus écrits pour une langue donnée, ou qu'elles divergent d'une forme standard dominante de manière non déterministe ou du moins mal connue.

they were writing. » (Lillehaugen, 2016)

14. « *written form based on the spoken language* » (Tournadre, 2014).

15. « [...]*have their own way of writing down the words they use according to the pronunciation they understand* » (Fauzi et Puspitorini, 2018).

16. « *[...] originally only recognizable through its oral narratives (pronunciation) is now easily recognizable through the spelling used in social media* » (Fauzi et Puspitorini, 2018).

1.3 Diversité linguistique dans le domaine du TAL

1.3.1 Définir les langues « peu dotées »

Munro (2015) et Mielke (2016) montrent, à travers leurs études des langues concernées par les articles publiés à la conférence de l'*Association for Computational Linguistics* ACL, un paysage du traitement automatique des langues relativement uniforme, concentré en particulier sur le traitement du chinois mandarin et de l'anglais.

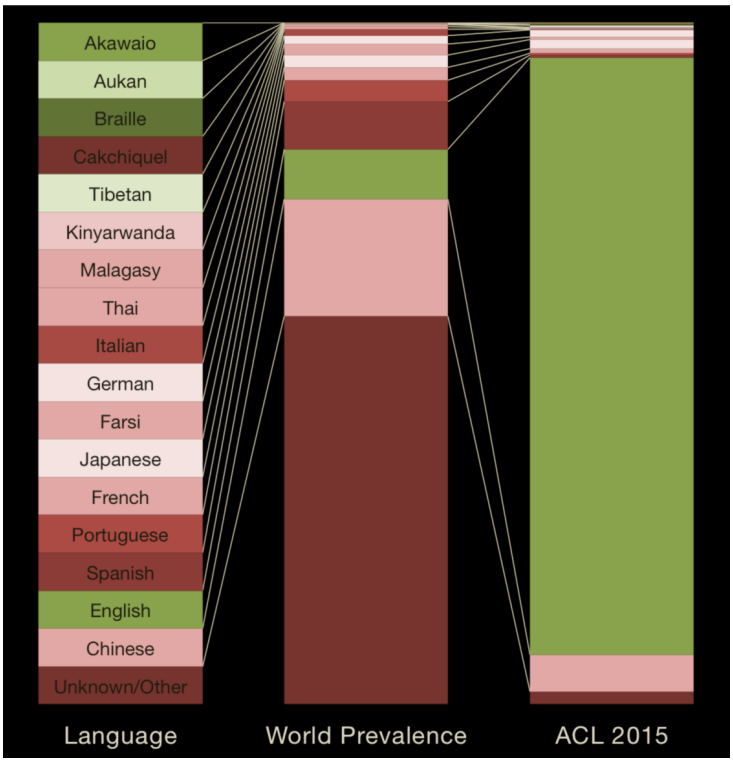


FIGURE 1.3 – Comparaison entre la distribution des locuteurs et le nombre de travaux pour les langues concernées par des travaux publiés à la conférence ACL en 2015 (Munro, 2015).

Le nombre de travaux concernant une langue donnée n'est pas proportionnel au nombre de ses locuteurs (voir la figure 1.3), ce qui n'est pas étonnant au regard de la non représentativité des TIC évoquée ci-dessus.

Il existe cependant un champ du TAL consacré aux langues qui, comparativement à d'autres, disposent de moins de ressources et outils favorisant leur intégration dans le monde numérique. Ces langues, qui se définissent donc en négatif des langues sur lesquelles un nombre grandissant de traitements peut être effectués, sont habituellement qualifiée de « peu dotées » (Berment, 2004).

Cette appellation recouvre nombre de réalités linguistiques différentes, notamment :

- des langues ayant, ou non, le statut de langue officielle (voir par exemple, l'islandais, langue officielle, comparée à l'igbo, parlé par 20 millions de de personnes au Nigeria, où la langue officielle est l'anglais),

- des langues parlées par un nombre plus ou moins important de locuteurs (voir par exemple l'inuktikut, environ 34 000 locuteurs, comparé au lao, 4 à 5 millions de locuteurs),
- des langues présentant, ou non, une parenté avec une langue mieux dotée (voir par exemple l'occitan languedocien au regard du catalan, comparés à l'arménien ou à l'albanais, qui ne sont proches d'aucune langue bien dotée),
- des langues dans des contextes géo-linguistiques variés (voir par exemple certaines langues berbères parlées dans plusieurs pays (Maroc, Algérie, Mali, Niger) au regard de l'inuktikut, géographiquement isolé),
- des langues cohabitant une autre langue, par exemple la langue « nationale » ou langue-toit (*Dachsprache*), sur un territoire donné (voir par exemple le cas des langues indonésiennes au regard du *bahasa indonesia* (seule enseignée et bénéficiant d'une littérature écrite) ou le cas des langues régionales en France).

Si l'appellation de « langue peu dotée » semble faire consensus en français, elle correspond à une multitude d'appellations en langue anglaise, ce qui traduit à la fois la diversité des réalités décrites et l'absence d'une définition consensuelle [17].

En anglais, Berment (2004) propose l'appellation « *little equipped* », mettant ainsi l'accent sur la disponibilité des outils et des usagers. Le terme le plus courant en anglais semble tout de même être celui d'« *under-resourced language* », c'est en tout cas celui qui est utilisé dans les ateliers principaux recueillant des travaux sur ces langues, notamment les six éditions de SLTU (*Spoken Language Technologies for Under-resourced languages*), initié en 2008 pour les travaux de traitement automatique de la parole, et les trois éditions de CCURL (*Collaboration and Computing for Under-Resourced Languages*), dont la première a eu lieu en 2014.

Il nous semble que le qualificatif « peu dotée » (ou, en anglais, « *less-resourced* ») sous entend que la différence entre une langue peu dotée et une langue bien dotée serait uniquement une question de quantité de données.

Berment (2004) affirme également que :

> « L'informatisation des langues peu dotées n'est pas tant une difficulté sur le plan informatique qu'une question de moyens humains et financiers pour permettre à ces populations de se munir des moyens adaptés à leurs écritures et à leurs langues. »

Il apparaît toutefois que les productions linguistiques « non canoniques » appellent à de nouvelles stratégies, et d'après Plank (2016), « la solution n'est pas évidente : tous les facteurs [de variation] ne peuvent pas être contrôlés, et la meilleure façon de dépasser la pratique actuelle, qui consiste à entraîner [des outils] sur des données homogènes provenant d'un seul domaine dans une seule langue, n'a pas encore été identifiée. » [18].

Dans le cas de langues multi-variantes, présentant une variation inter- et intra-locuteurs, il n'est en effet pas réaliste d'envisager disposer un jour des ressources suffisantes pour pouvoir appliquer les méthodes développées pour des productions langagières plus canoniques, et ce quels que soient les moyens déployés.

17. On trouve notamment en anglais « *low-resourced languages* », « *under resourced languages* », « *resource free language* », « *resource disadvantaged languages* », « *non-resourced language* », « *low-density language* », « *lower-density languages* », « *resource-poor language* », « *resource-scarse languages* » , « *less-resourced languages* », « *little equipped* », « *less-represented* », « *scarce resource* », « *poor resource scenario* ».

18. *« The solution is not obvious : we cannot control for all factors, and it is not clear how to best go beyond the current practice of training on homogeneous data from a single domain and language ».* (Plank, 2016)

C'est ce constat qui mène les chercheurs travaillant au traitement de langues peu dotées et non standardisées à proposer des approches limitant la dépendance aux données. Parmi celles-ci, la « normalisation » des contenus non canoniques permettant de se rapprocher de productions linguistiques plus facile à traiter, ou les techniques d'« adaptation au domaine » à comprendre au sens large, c'est-à-dire où le domaine source et le domaine cible peuvent être deux langues différentes. Nous présentons un état de l'art des stratégies mises en place concernant l'annotation en morphosyntaxe dans la section 2.3.2.2.

1.3.2 Observer la diversité linguistique du domaine

Une première manière d'analyser la diversité linguistique en TAL est d'observer les langues présentes dans des communications acceptées dans les conférences du domaine. Munro (2015) et Mielke (2016) présentent, à travers leurs études des langues concernées par les articles publiés à la conférence de l'*Association for Computational Linguistics* ACL, le domaine TAL sous un jour relativement uniforme, concentré en particulier sur le traitement du chinois mandarin et de l'anglais. De son côté, l'étude menée par Benjamin (2018) sur le programme d'ICLDC 2017 (*International Conference on Language Documentation & Conservation*) [19] montre qu'une minorité de publications récentes concernant les langues peu dotées s'attelle effectivement au développement de ressources technologiques pour le TAL. Il montre qu'une partie encore moindre présente de réelles innovations technologiques, alors même que les travaux portant sur de nouvelles langues présentent l'intérêt de confronter les chercheurs à des problématiques linguistiques nouvelles. En cause, notamment, le « vortex monolingue » (« *monolingual vortex* »), défini par Branco (2018) comme étant le cercle vicieux induit par les politiques de financement et les opportunités professionnelles moindres découlant des recherches sur les langues peu dotées.

Les travaux sur les langues peu dotées pouvant faire l'objet d'ateliers ou de conférences spécifiques, il est également intéressant de consulter les plateformes d'archivage de ressources linguistiques.

On observe sur celles-ci une augmentation de la diversité linguistique au sein du domaine, notamment en constatant la diversité croissante de l'*Open Archives Language Community* (OLAC) [20], qui propose un moteur de recherche pour les ressources linguistiques englobant une soixantaine de catalogues dont celui de l'*European Language Resources Association* (ELRA), le *Catalogue of Language Resources* [21] qui contient 1 364 ressources pour un total de 76 langues.

Enfin, la figure 1.4 montre l'évolution du nombre de langues et de corpus arborés présents dans les différentes versions du projet *Universal Dependencies* (UD) [22] : en 5 ans, 59 nouvelles langues ont fait leur apparition dans le projet.

En définitive, si la majorité des travaux concerne les langues qui sont déjà les mieux dotées en terme de ressources et d'outils, la tendance est à l'amélioration de la diversité linguistique.

19. Voir : http://icldc5.icldc-hawaii.org/, juin 2020.
20. Voir : http://search.language-archives.org/index.html, juin 2020.
21. Voir : http://catalog.elra.info/index.php, juin 2020.
22. Voir : https://universaldependencies.org/, juin 2020.

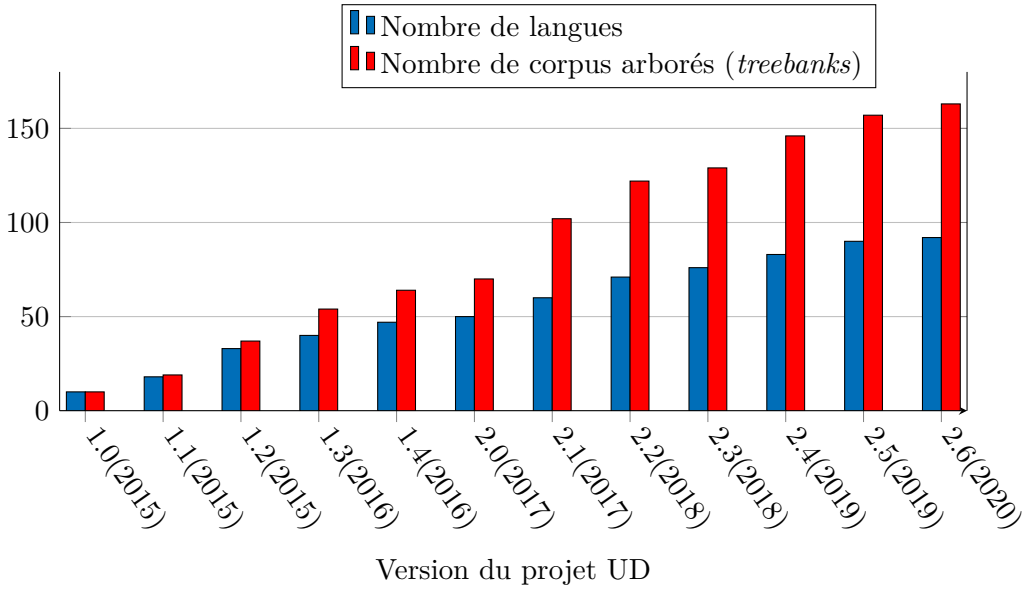


FIGURE 1.4 – Évolution de la diversité linguistique dans le projet UD (*Universal Dependencies*).

1.4 Conclusion

La diversité linguistique réelle, si on la conçoit comme le nombre de langues parlées par des locuteurs natifs, est en déclin. Les TICs représentent toutefois un espace croissant d'expression pour un ensemble de communautés linguistiques plus ou moins vulnérables et le cadre d'expression que constitue le numérique permet à de nouvelles pratiques d'émerger.

Comme nous l'avons vu à travers différents exemples, la présence en ligne de certaines langues peut être constatée indépendamment de l'existence d'éléments formels tels qu'un alphabet ou une orthographe stable, ou de la disponibilité de quelconques outils de TAL. Il nous semble pour autant raisonnable de formuler l'hypothèse que l'existence d'outils adaptés de TAL permet de favoriser l'utilisation numérique d'une langue. En particulier, on peut supposer que les deux mécanismes illustrés par la figure 1.5 ont lieu simultanément.

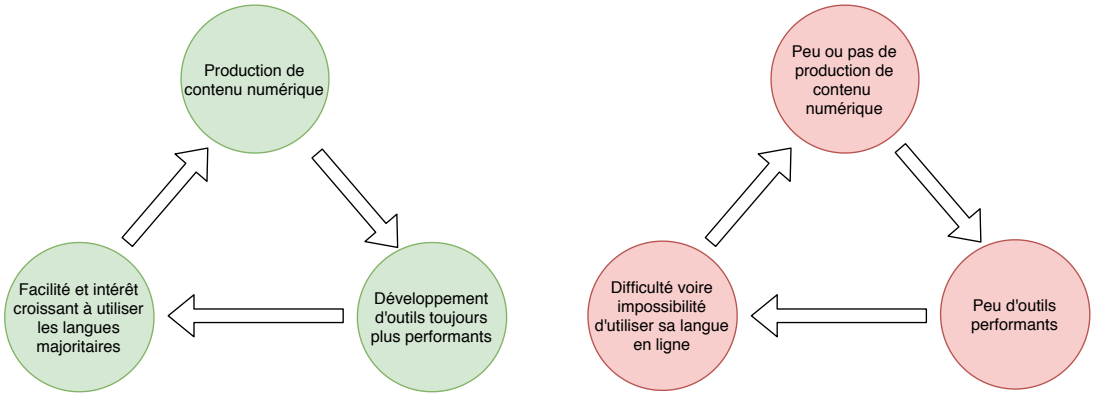


FIGURE 1.5 – Les cercles vertueux et vicieux de la présence linguistique en ligne.

D'une part, un cercle que nous qualifions de vertueux conduit l'existence d'outils à favoriser

la présence en ligne d'une langue donnée, et à accroître son intérêt pour les locuteurs, qui produisent en ligne du contenu numérique lui-même utile pour entraîner de nouveaux outils plus performants. À l'inverse, un cercle que nous qualifions de vicieux peut être observé si l'absence d'outil conduit les locuteurs à préférer une autre langue pour leur expression en ligne. Dans un tel cas, aucune ressource numérique n'est produite, empêchant ainsi le développement d'outils qui pourraient permettre d'enrayer ce mécanisme.

Les observations que nous avons proposées montrent que les recherches en TAL, si elles tendent vers davantage de diversité linguistique, ne sont pas encore en phase avec celle qu'on peut observer dans les TICs : la diversité des pratiques en ligne augmente plus vite que la couverture des outils de TAL. Dans le chapitre suivant, et afin de comprendre la difficulté du TAL à suivre les évolutions linguistiques, nous présentons en quoi les ressources linguistiques tiennent un rôle de pierre angulaire pour le développement d'outils de TAL.

Chapitre 2

Ressources linguistiques

Sommaire

Nous présentons dans ce chapitre les travaux concernant la production de ressources linguistique utiles au développement et à l'évaluation d'outils d'annotation en parties du discours. Nous nous concentrons sur cette tâche en tant que brique de base indispensable à de nombreuses applications de TAL faisant encore défaut à de nombreuses langues.

Les ressources pour le TAL, qu'elles soient brutes ou annotées, sont des ressources coûteuses à plusieurs égards et il convient de savoir tirer le meilleur profit possible des efforts employés pour les développer. C'est pourquoi nous consacrons la section 2.1 à une introduction des conditions de « pérennité » d'une ressource linguistique.

Dans nombre de cas, le traitement automatique des langues peut être envisagé comme un traitement automatique de *corpus* aux caractéristiques variées et plus ou moins contrôlées. Outre leur statut d'objet du traitement, les corpus sont usuellement utilisés pour entraîner, ajuster, évaluer, comparer nombre d'outils développés en TAL, et tiennent par conséquent une place centrale au sein de la discipline.

La création de ressources langagières de qualité est notoirement coûteuse. L'une des rares publications concernant l'annotation d'une ressource langagière et en présentant un coût approximatif concerne le Prague Dependency Treebank (Böhmová *et al.*, 2001). La constitution de ce corpus aurait coûté 600 000 dollars. À ce problème de coût peut s'ajouter celui de la difficulté de

trouver des experts linguistes en nombre suffisant pour certaines langues. Plus récemment, Martínez Alonso *et al.* (2016) et Seddah *et al.* (2020) proposent une analyse des coûts de différentes campagnes d'annotation menées sur des corpus variés du français et de dialectes nord africains transcrits en arabizi respectivement. Ces coûts importants, dont Martínez Alonso *et al.* (2016) proposent une estimation moyenne de l'ordre de trois euros par phrase pour un ensemble de corpus annotés en syntaxe dans le cas du français, dépend naturellement des ressources préexistantes ainsi que de la complexité de l'annotation produite. Dans le cas des corpus annotés du français dont deux sont formés de contenus générés par des utilisateurs, très peu standardisés et fortement bruités, Martínez Alonso *et al.* (2016) soulignent par ailleurs l'impact sur les coûts de la très haute compétence des annotateurs impliqués. Dans le contexte de l'annotation de l'arabizi nord-africain, Seddah *et al.* (2020) insiste sur le coût additionnel engendré par le contexte linguistique concerné. Sont notamment évoquées à ce sujet l'absence initiale de guide satisfaisant, la nécessité de mettre à jour celui-ci au cours de la campagne, ainsi que la nécessité d'entraîner régulièrement des annotateurs pour qui la nature multi dialectale du corpus rend la tâche d'annotation d'autant plus ardue.

Quelle que soit la tâche d'annotation envisagée, les coûts financier et temporel de l'annotation de ressources sont tels qu'il paraît irréaliste que l'ensemble des productions linguistiques, toutes langues et domaines confondus, soit un jour couvert. De telles ressources sont traditionnellement nécessaires à l'entraînement de modèles d'annotation issus de l'apprentissage statistique et profond. Les techniques d'apprentissage non supervisées se proposent de limiter cette dépendance aux ressources annotées. Elles ouvrent en particulier la possibilité de traiter des productions linguistiques pour lesquelles la construction de ressources annotées requerrait l'intervention d'experts de la langue (ou du domaine [23]) qui ne sont pas disponibles pour se consacrer à une telle tâche. S'affranchir des ressources annotées a néanmoins nécessairement un coût : celui de la quantité de ressources brutes à pourvoir à ces algorithmes.

Nous consacrons ainsi la section 2.2.1 de ce chapitre à une présentation des sources de corpus textuelles pouvant être utilisées en TAL. La dernière section de ce chapitre présente les ressources langagières particulières que constituent les corpus annotés, en leur qualité de produit de l'annotation *manuelle* d'une part, et d'objet d'entraînement pour l'annotation *automatique* d'autre part.

2.1 Construire des ressources linguistiques pérennes

La pérennité est définie par Habert (2010) comme :

> « [...] les conditions techniques mais aussi et surtout sociales et humaines qui permettent de produire du « numérique durable », c'est-à-dire tel qu'il permette un accès maintenu, à long terme, des communautés aux connaissances et aux savoirs qui leur sont précieux. »

Wilkinson *et al.* (2016) ont formalisé un ensemble de bonnes pratiques allant dans le sens de cette pérennisation et qui sont regroupées sous l'acronyme FAIR. Ces pratiques visent en effet à agir sur quatre qualités des données scientifiques au sens large :

- la « trouvabilité » (***F**indability*), soit la possibilité d'être *trouvée*, que ce soit par un opérateur humain ou par une machine, notamment *via* des algorithmes d'indexation. Cette

23. Pour plus de détails sur la notion d'expert, voir (Fort, 2017).

trouvabilité est atteinte grâce à l'association des données à des méta-données descriptives de qualité.

- l'« accessibilité » (***A**ccessibility*), soit la mise en place de conditions (pouvant comprendre une authentification) d'accès aux données.
- l'« interopérabilité » (***I**nteroperability*), soit la diffusion des données et méta-données dans des *formats* eux-mêmes accessibles et en usage.
- la « ré-utilisabilité » (***R**eusability*) des données, dont les trois points ci-dessus sont des conditions nécessaires, qui est avant tout une qualité de la description des données : licence d'usage, provenance, documentation selon les pratiques de la communauté scientifique concernée.

Notons qu'il est impossible par nature pour certaines ressources de parvenir à obéir à un ou plusieurs des principes du « FAIR data ». C'est le cas notamment des données ne pouvant être publiées sous une licence permettant leur réutilisation, parce qu'elles contiennent des ressources propriétaire ou des informations sensibles, par exemple.

Les démarches à entreprendre pour s'assurer que les ressources qu'il produit répondent à ces principes constituent une charge de travail additionnelle pour le chercheur. Dans le domaine du TAL, les mécanismes d'incitation assurant la trouvabilité, l'accessibilité et l'interopérabilité des ressources sont néanmoins nombreux. Par exemple, depuis 2010, la revue *Language Resources and Evaluation*[24] (LRE) ainsi que plusieurs conférences du domaine encouragent leurs participants à publier dans la LRE Map[25] les ressources qu'ils construisent et utilisent au moment de la publication d'un article. Consultée en avril 2020, la base contenait ainsi 2 857 enregistrements de ressources pour 100 langues. Cette base permet par ailleurs de créer des références de ressources à l'instar des références bibliographiques, permettant ainsi de les citer dans des publications scientifiques.

Concernant l'archivage pérenne des ressources en tant que tel, c'est-à-dire la certification et le stockage des données (Habert, 2010), la base ORTOLANG permet quant à elle le dépôt facile en ligne « de ressources provenant de laboratoires de recherche français et portant sur toute langue, ou pour toute source de ressources portant sur les langues de France quelle que soit leur origine. »[26].

La documentation à produire dans le cas de ressources langagières pour le TAL afin d'assurer la « ré-utilisabilité » de celles-ci est formée de diverses informations. Notamment l'origine, la langue et la nature des ressources linguistiques représentées, les traitements qui ont été effectués sur les ressources telles que les opérations de lemmatisation, d'annotation et de traduction, mais aussi les opérations de « pré-traitements ». Il nous semble que la majorité de ce qui est souvent décrit comme des « pré-traitements » devraient être considérés comme des traitements à part entière. C'est notamment le cas lorsque ce terme réfère à une opération de mélange de phrases dans un corpus, ou de tokenisation (Habert *et al.*, 1998), qui ont un impact non nul sur les traitements suivants ou les tâches qu'elles permettent de réaliser en aval (voir par exemple l'impact de la tokenisation pour des tâches de reconnaissance d'entités nommées (Benajiba *et al.*, 2009; Kolluru *et al.*, 2011), de traduction automatique (Domingo *et al.*, 2018), ou de similarité lexicale (Bollegala *et al.*, 2020)).

24. Voir : http://www.elra.info/en/dissemination/jlre-language-resources-and-evaluation-journal/, juin 2020.
25. Voir : http://lremap.elra.info/.
26. Voir : https://www.ortolang.fr/information/policy, juin 2020.

Il nous apparaît en outre que la qualité d'une ressource annotée ne peut être envisagée séparément de la qualité de la ressource brute sur laquelle elle est construite.

La question de la disponibilité des ressources est étroitement liée, dans le domaine du TAL, à celle de la reproductibilité de la recherche (Cohen *et al.*, 2016). Les enjeux de reproductibilité nous paraissent revêtir un caractère d'autant plus crucial que la recherche concernée est « précaire » au sens large, c'est-à-dire en termes de ressources, de financements et de chercheurs disponibles.

Pour ces raisons, nous nous attachons dans la suite de ce manuscrit à porter un regard particulièrement attentif à la disponibilité des ressources que nous présentons.

2.2 Ressources textuelles brutes

L'accès à des ressources brutes est par définition moins coûteux que l'accès aux mêmes ressources annotées. Ceux-ci doivent en revanche être distribués sous licence claire : l'obtention et l'utilisabilité réelle des corpus bruts n'est pas triviale et mérite d'être interrogée.

En particulier, comme mentionné dans la section précédente, il convient de questionner la disponibilité, l'accessibilité légale et la qualité des corpus bruts.

Nous présentons dans la section 2.2.1 les pratiques habituelles régissant le développement de corpus textuels pour le TAL. Puis, dans les sections 2.2.2, 2.2.3 et 2.2.4 nous présentons comment trois sources de corpus issues du Web peuvent être utilisées pour la constitution de corpus et qu'elles représentent dans le cas des langues peu dotées et non standardisées.

Nous nous intéressons en particulier à ces ressources dans leur perspective multilingue et multilocuteurs. Étant donnée les phénomènes de variation qui peuvent être rencontrés dans certain cas, nous nous intéressons la prise en compte de la variation compte dans chacun des cas présentés.

Si nous proposons un regard critique sur ces ressources, il ne s'agit en aucun cas de discréditer leur exploitation. Il nous semble néanmoins légitime de relever les biais qu'elles présentent et les difficultés qui se posent dans le cadre de langues peu dotées et non standardisées.

Ainsi, nous présentons dans un premier temps (section 2.2.2) comment le Web en général peut être utilisé comme une sources de corpus textuels grâce à la collecte automatique de pages Web (*Web crawling*).

Puis, nous proposons une présentation (section 2.2.3) du projet Wikipédia, qui, avec 306 Wikipédias actives distribuées sous licence Creative Commons [27], est une des bases multilingues les plus importantes (sinon la plus importante) qui soit librement accessible. Cette ressource est largement utilisée comme base textuelle et de connaissances en TAL. En outre, il s'agit probablement de l'une des initiatives de myriadisation les plus réussies, impliquant des communautés linguistiques nombreuses et variées. Elle représente en cela une référence en termes de production participative et c'est pourquoi nous en proposons une analyse approfondie. En particulier, nous nous intéressons à l'utilisabilité réelle des pages contenues dans les projets Wikipédia.

Enfin (section 2.2.4), nous commentons brièvement l'utilisation des corpus issus des réseaux sociaux.

27. Voir : https://creativecommons.org/, juin 2020.

Les observations présentées dans ces sections ont en partie été publiées dans un article de l'atelier de LREC 2020 SLTU-CCURL 2020 (Millour et Fort, 2020).

2.2.1 Corpus textuels « pour le TAL »

La construction d'une ressource linguistique, même brute, n'est pas neutre et procède le plus souvent de choix de composition.

De bonnes pratiques pour la construction de corpus en linguistique ont été définies par Wynne (2005). Dans ce guide, on retrouve un ensemble de recommandations proposées par (Sinclair, 2005) concernant les paramètres à prendre en compte au moment de la construction d'un corpus. Ces recommandations incluent entre autres des exigences de taille, de technique d'échantillonnage et de représentativité. La discussion menée lui permet de proposer la définition suivante pour le terme « corpus » :

> « Un corpus est une collection de textes sous forme électronique sélectionnés selon des critères externes dans l'optique de représenter, autant que possible, une langue ou une variété de langue, et utilisée comme source de donnée pour la recherche en linguistique. » [28]

Il y est par ailleurs indiqué que le caractère « représentatif » d'un corpus pour l'étude linguistique est conditionné par l'objectif poursuivi : étudier un objet linguistique, qu'il s'agisse d'un phénomène linguistique particulier, de l'évolution d'un phénomène, etc.

À notre connaissance il n'existe pas de guide similaire pour guider la construction de corpus pour le traitement automatique des langues, et pour cause : l'utilisation de corpus en TAL répond à des besoins très divers, pour lesquels la notion de « représentativité » répond à des impératifs mouvants. Il nous apparaît en effet qu'en TAL, un corpus est considéré comme représentatif d'un phénomène donné s'il permet de traiter ce dernier efficacement. Ainsi, le caractère représentatif d'un corpus est conditionné non seulement par la tâche de traitement concernée mais aussi par la technique employée pour le traiter efficacement.

En linguistique comme en TAL, l'obtention de corpus suit souvent une approche pragmatique ou « opportuniste » (McEnery et Hardie, 2011), les sources de corpus utilisées étant celles qui sont accessibles au sens décrit plus haut à un instant t.

En témoigne l'utilisation répandue des corpus de débats parlementaires (EUROPARL (Koehn, 2005), Hansard Corpus (Alexander et Davies, 2015)), journalistiques (Wall Street Journal Corpus (Paul et Baker, 1992), à partir duquel le Penn Treebank est en partie basé (Marcus *et al.*, 1993). Le caractère spécialisé de ces corpus leur confère naturellement un biais, comme démontré par (Sharoff, 2006), à travers la comparaison de l'anglais présent dans les corpus journalistiques, le corpus BNC, ou sur le Web. La comparaison de ces corpus montre en effet que « les corpus journalistiques diffèrent significativement à la fois d'un corpus représentatif ou du corpus de l'Internet, et ne fournit pas un aperçu de l'utilisation moderne du langage » [29].

28. « *A corpus is a collection of pieces of language text in electronic form, selected according to external criteria to represent, as far as possible, a language or language variety as a source of data for linguistic research.* » (Sinclair, 2005).

29. « *The comparison shows that the news corpora differ significantly from either representative or Internet corpora and cannot provide a window into modern language use in general.* » (Sharoff, 2006).

Notons par ailleurs que plusieurs initiatives basées sur la traduction de contenus dans de multiples langues contribuent à rendre accessible des corpus dans un nombre croissant de langues. À ce sujet, Agić et Vulić (2019) présentent par exemple JW300 un corpus parallèle de phrases alignées pour 300 langues issues des revues *Watchtower* et *Awake !* [30]. L'initiative en ligne TATOEBA fait quant à elle était de plus huit millions de phrases traduites collaborativement sur le Web pour 387 langues (23 188 (±40,093) phrases en moyenne [31]).

Aujourd'hui, les corpus correspondant au « mode électronique » tel que défini par Sinclair et Ball (1996), c'est-à-dire en opposition aux corpus de type « écrits » et « oraux » présentent un intérêt croissant pour le domaine du TAL.

En effet, le Web constitue un nouvel objet linguistique d'intérêt à traiter. En tant qu'espace multilingue, riche en métadonnées, permettant dans certains cas la géolocalisation des auteurs des contenus produits, il est donc devenu une source de corpus riche pour lequel le processus de *collecte* diffère largement des entreprises de *construction manuelle* de corpus existant jusqu'alors. Pour ces raisons, nous nous penchons dans la section suivante sur les approches visant à utiliser le Web comme source de corpus.

2.2.2 Le Web comme corpus

En 2001 l'affirmation enthousiaste de Kilgarriff et Grefenstette (2001) « *The corpus of the new millennium is the web.* » promet une révolution de la linguistique de corpus faisant la part belle aux corpus issus d'Internet. Depuis, les critiques autant que les travaux permettant effectivement de tirer parti d'Internet comme corpus se sont succédé. Gatto (2014) propose une analyse détaillée des évolutions des regards sur l'utilisation du Web en linguistique et une réflexion sur l'impact de l'utilisation du Web sur la notion même de corpus. Elle y commente la transition du Web comme source de corpus vers le statut d'objet linguistique indépendant, ainsi que les discussions concernant la reproductibilité (ou non) des expériences menées sur la base de contenus qui ne sont pas conçus comme pérennes. Elle détaille comment le Web peut être considéré comme un « magasin de corpus » (« *Web as a corpus shop* ») (Baroni et Ueyama, 2006), pour la construction de corpus *ad-hoc*.

La diversité des contenus publiés sur le Web, générée par la diversité des usages et des internautes, permet en effet de considérer la possibilité de collecter des corpus équilibrés (au sens où le BNC peut être considéré comme équilibré) tout en atteignant une couverture beaucoup plus large.

Par ailleurs les contenus produits numériquement et diffusés *via* Internet présentent *a priori* le double avantage d'être (i) faciles à collecter, (ii) exploitables rapidement grâce à leur format numérique adapté au traitement automatisé.

Mais le Web est un contenant hétérogène et certaines productions qu'il contient représentent un matériau linguistique nouveau appelant des méthodes de collecte et de traitements automatiques spécifiques. En cela, « Le Web n'est représentatif de rien d'autre que de lui-même. Mais c'est le cas de tout autre corpus » [32] (Kilgarriff et Grefenstette, 2003).

30. Voir https://www.jw.org/en/library/magazines/.
31. Voir : https://tatoeba.org/fra/stats/sentences_by_language, novembre 2020.
32. « *The web is not representative of anything else. But nor are other corpora, in any well-understood sense.* » (Kilgarriff et Grefenstette, 2003).

En effet, le Web n'est pas seulement une extension numérique de ce qui pré-existait, comme on pourrait concevoir la presse numérique comme une extension de la presse papier, ou les courriels comme une extension des correspondances épistolaires. La production *via* des outils numériques et à destination d'Internet a fait émerger de nouvelles communautés numériques aux pratiques linguistiques propres, variant en termes de genre, de registre, de style.

Par ailleurs, comme développé en section 1.2.2, le Web est également le berceau de pratiques nouvelles pour certaines langues pour lesquelles la pratique scripturale était jusqu'alors peu démocratisée.

Dans le cas de telles pratiques linguistiques, pour lesquelles la pratique scripturale se limite, ou presque, à la production de contenu numérique, le *web for corpus*, c'est-à-dire la conception du web comme source de corpus, se confond avec le *web as corpus*, soit la conception du web comme un corpus *per se*.

Du côté du traitement automatique des langues, l'exploitation du Web comme corpus apparaît naturellement comme une manière de s'attaquer au goulot d'étranglement que constitue le manque de données textuelles (« *data bottleneck* »). La répercussion pour le TAL de la formation de communautés linguistiques en ligne semble être l'apparition d'un nouveau « domaine » de traitement pour les outils : celui des contenus générés par les « utilisateurs » (« *User Generated Content* » (UGC)), caractéristique dont se doublent les « locuteurs » lorsqu'ils s'expriment *via* les outils numériques.

Dans la suite de cette section, nous présentons d'abord les techniques de construction de corpus par collecte automatique de pages Web communément utilisées. Nous utilisons cette présentation pour mettre en avant les insuffisances d'une telle méthode pour construire des corpus dans le cas qui nous occupe dans cette thèse, c'est-à-dire le cas de langues peu dotées et multi-variantes.

2.2.2.1 Collecte automatique de pages Web

La collecte automatique de pages Web (*web crawling*) pour la construction de corpus multilingues consiste à télécharger *via* Internet des textes y ayant été publié en utilisant peu de filtres sur les pages concernées. Les pages HTML sont ensuite « nettoyées », c'est-à-dire que seul le contenu intéressant le chercheur est conservé [33] et automatiquement associées, lorsque c'est possible, à la langue à laquelle elles correspondent.

La diversité linguistique sur le Web est telle que cette méthode a été exploitée pour obtenir des corpus pour un grand nombre de langues, dont des langues peu dotées. Quelle que soit la méthode choisie pour collecter des pages Web, il est nécessaire de procéder à une identification de la langue pour classer les documents.

Par exemple, le projet An Crúbadán (Scannell, 2007), première initiative du genre à grande échelle à notre connaissance, utilise une combinaison de trigrammes, de lexiques générés automatiquement et de listes de mots spécifiques à chaque langue pour identifier des contenus du Web écrits en 2 228 langues. D'autres stratégies ont pu être mises en place comme le « *bootstrap* » de corpus grâce à des requêtes spécifiques (Goldhahn *et al.*, 2012) ou le filtrage préalable des URLs à utiliser, comme décrit par Barbaresi (2013) dans le cas de langues peu dotées. Le corpus OSCAR (Ortiz Suárez *et al.*, 2019) est issu du filtrage et de la classification linguistique

33. En fonction de l'objectif poursuivi, il peut s'agir du texte enrichi ou non de balises HTML signalant la structure du document, par exemple.

dans 166 langues du corpus du Web formé par les documents distribués par le projet Common Crawl [34], facilitant ainsi notablement son exploitation par langue. Les travaux de Schwenk *et al.* (2020) présentent et évaluent une méthodologie permettant d'aligner alignés automatiquement des corpus monolingues extraits du Web (Wenzek *et al.*, 2020) pour y faire émerger des corpus parallèles. La collection de corpus ccmatrix qui en est issue contient environ trois millions de paires de phrases alignées pour un total de 28 langues.

Mentionnons en outre l'initiative OPUS (Tiedemann, 2012), qui a pour vocation de fédérer et de proposer une interface d'accès à un ensemble très varié de corpus alignés provenant de traduction de contenus publiés sur le Web.

L'identification linguistique des corpus implique, dans la majorité des cas, l'existence d'informations statistiques sur la distribution de motifs linguistiques caractéristiques pour chacune des langue à classer. Ce genre de donnée n'est néanmoins pas toujours disponible ni adapté à l'identification de langues multi-variantes. En effet, la finesse de l'attribution d'une langue à un document peut être limitée par les performances des outils nécessaires au traitement. C'est le constat de Barbaresi (2013) concernant les paires de langue indonésienne et malaise, deux langues proches pour lesquelles « il est pertinent de les considérer ensemble car il est parfois difficile de tracer une frontière nette entre leurs variantes linguistiques, et c'est d'autant plus vrai pour des outils d'identification automatique ».

Dans un cas comme celui-ci, la performance de l'outil automatique semble freiner l'exigence linguistique et niveler par le bas l'objectif poursuivi. Il nous semble en réalité que c'est la dépendance de la méthode à une technique d'identification qui pose ici problème.

Dans tous les cas, cela nous amène à penser que si la collecte automatique de pages Web permet incontestablement d'obtenir des ressources de qualité pour de nombreuses langues comme l'allemand, l'anglais, le français, l'italien (Baroni *et al.*, 2009), cette méthode n'est sans doute pas adaptée à la collecte de corpus pour des langues dont les outils d'identification ne sont pas suffisamment performants.

2.2.2.2 Moteurs de recherche

Afin de s'affranchir de la nécessité d'identifier la langue a posteriori, il est possible, comme présenté par Sharoff (2006) d'utiliser les moteurs de recherches pour filtrer en amont les pages collectées. Notamment, il est possible de construire des requêtes de combinaisons de termes choisis dans une liste de mots élaborée pour une langue donnée. Le choix des mots est déterminant dans la nature des textes renvoyés par le moteur de recherche. La méthode proposée est « applicable à toute langue dont la présence sur Internet est plus ou moins significative. » [35]. Comme évoqué par Gurrutxaga *et al.* (2010) cette technique présente effectivement des limites, notamment pour la construction de corpus spécialisés, pour lesquels les termes utiles peuvent être partagés par plusieurs langues, ou dans le cas de langues à morphologie riche telle que le basque pour lesquelles l'utilisation de mots clés n'est pas efficace.

34. Voir : https://commoncrawl.org/, juin 2020.
35. « *The proposed procedure [...] in applicable to any language with more or less significant Internet presence.* » (Sharoff, 2006).

2.2.2.3 Distribution de corpus issus du Web

La distribution de corpus collectés automatiquement sur le Web est contestable d'un point de vue légal, de nombreux pays (dont la France) ne reconnaissant pas le « *fair use* » appliqué dans le monde anglo-saxon. C'est un des « inconvénients » pointés par Baroni et Ueyama (2006) : « [...] si un chercheur prévoit de distribuer un corpus du Web important, constitué de millions de documents, il (ou elle) aura la tâche difficile d'obtenir l'autorisation d'utiliser les documents de tous les titulaires de droits d'auteur. » [36].

Pour échapper aux restrictions, les corpus collectés sont ainsi parfois distribués sans métadonnées ou sous forme de liste de phrases mélangées aléatoirement, comme décrit par (Habernal *et al.*, 2016) dans le cas des corpus COW (Schäfer et Bildhauer, 2013) et Leipzig Corpora (Goldhahn *et al.*, 2012). Ces opérations présentent l'inconvénient majeur d'homogénéiser le corpus collecté, et la perte d'informations qui en découle se répercute à plusieurs niveaux :

- Le mélange des phrases conduit à la construction d'une ressource unique et hétérogène pour chaque entité linguistique identifiée. Dans le cas d'une langue multi-variante, cela conduit au mélange des différentes variantes et empêche ainsi toute étude plus fine de celles-ci. Or, il n'est pas démontré que la manière la plus efficace de traiter automatiquement une langue multi-variante soit d'utiliser un corpus hétérogène. Nos propres expériences sur ce sujet tendent d'ailleurs à soutenir la thèse inverse (voir section 7.1.4).
- le mélange des phrases casse la structure du document, limitant les exploitations ultérieures au contexte de la phrase.

Pour remédier à ces limitations, Lyding *et al.* (2014) et Barbaresi et Würzner (2014) proposent de procéder à l'identification automatique de licence de distribution, notamment de licences Creative Commons [37]. Le problème de la distribution de corpus peut effectivement être abordé en choisissant parmi le contenu du Web les pages qui sont publiées sous des licences claires. C'est notamment le cas des pages publiées dans les projets issus de la Wikimedia Foundation et en particulier de Wikipédia, auquel nous nous intéressons dans la section suivante.

2.2.3 Wikipédia

Wikipédia est une encyclopédie multilingue collaborative en ligne soutenue par la Wikimedia Foundation, une organisation à but non lucratif créée en 2003.

Cette encyclopédie en ligne est largement utilisée comme ressource linguistique au sein de la communauté du TAL, en tant que base d'informations structurées dont dérivent ressources sémantiques et ontologies, mais aussi et avant tout comme source de corpus textuels facilement accessibles. Son accessibilité bénéficie en effet aux langues bien dotées comme à celles qui le sont moins.

Sa popularité et sa structure collaborative en ont fait un des environnements les plus propices à la production de contenus par une multitude de locuteurs, un article pouvant lui-même contenir des contributions de plusieurs auteurs.

36. « *Third, if a researcher plans to distribute a large Web corpus made of million of documents, (s)he will have a very hard time obtaining permission to use the documents from all the copyright holders.* (Baroni et Ueyama, 2006) ».
37. Voir : https://creativecommons.org/, juin 2020.

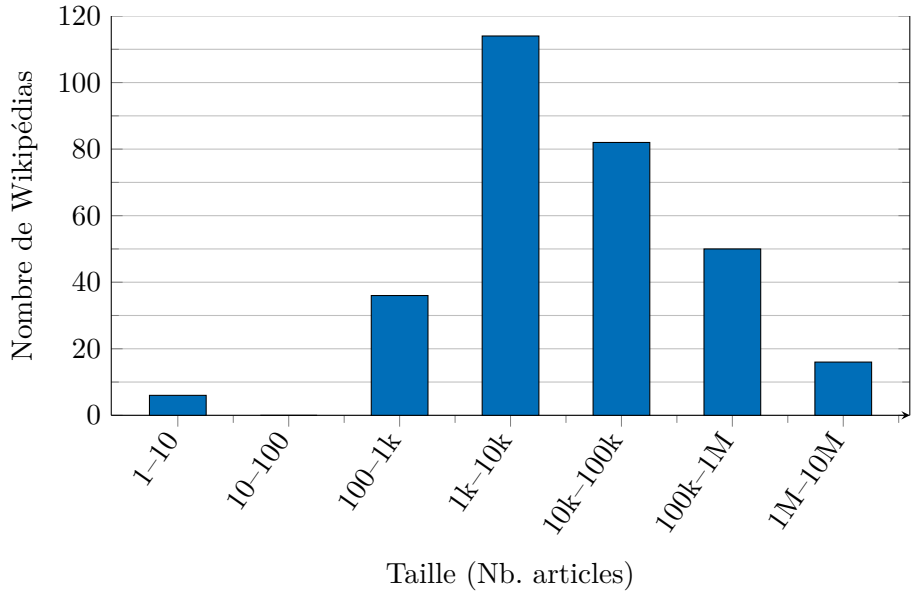


FIGURE 2.1 – Répartition des Wikipédias par taille en nombre d'articles.

En cela, Wikipédia représente une ressource précieuse pour les langues peu dotées pour lesquelles elle représente dans certains cas une partie importante du contenu numérique librement disponible. Son caractère collaboratif couplé à l'absence de standard dans certains cas amènent Wikipédia à abriter des diversités dialectales et scripturales.

Après une brève introduction sur la taille et la qualité des Wikipédias existantes, nous présentons les différentes stratégies mises en place pour intégrer l'absence de standard consensuel pour certaines langues.

2.2.3.1 Taille et qualité des Wikipédias

Les chiffres fournis dans cette section correspondent à la situation en avril 2020.

Il existe des projets Wikipédia pour 309 langues, et 299 d'entre eux sont actifs [38]. Comme illustré sur la figure 2.1, 81 Wikipédias présentent entre 1 000 et 10 000 articles, 100 entre 10 000 et 100 000 articles, 49 entre 100 000 et 1 million d'articles et 17 plus d'un million d'articles, .

Si l'observation de la taille des Wikipédias en termes de nombre d'articles donne un aperçu intéressant de la diversité linguistique du projet, la taille n'est pas le meilleur indicateur pour avoir une idée de la quantité de données de qualité effectivement disponibles dans chaque Wikipédia. Pour cela, l'indicateur de « profondeur » `depth` [39] a été défini par WIKIMEDIA : cet indicateur fournit une estimation de la qualité d'une Wikipédia donnée sur la base du nombre d'articles qu'elle contient, mais aussi du nombre de modifications et de la proportion de pages « non-article » telles que les pages de profil utilisateur, les redirections, etc.

Cet indicateur varie de 0 à 1 063 (Wikipédia Francique ripuaire), sa valeur moyenne est 72, sa valeur médiane 32 et son écart type 67. La Wikipédia anglaise a un score de `depth` de 999. La

38. Voir : https://meta.wikimedia.org/wiki/List_of_Wikipedias.
39. Voir : https://meta.wikimedia.org/wiki/Wikipedia_article_depth.

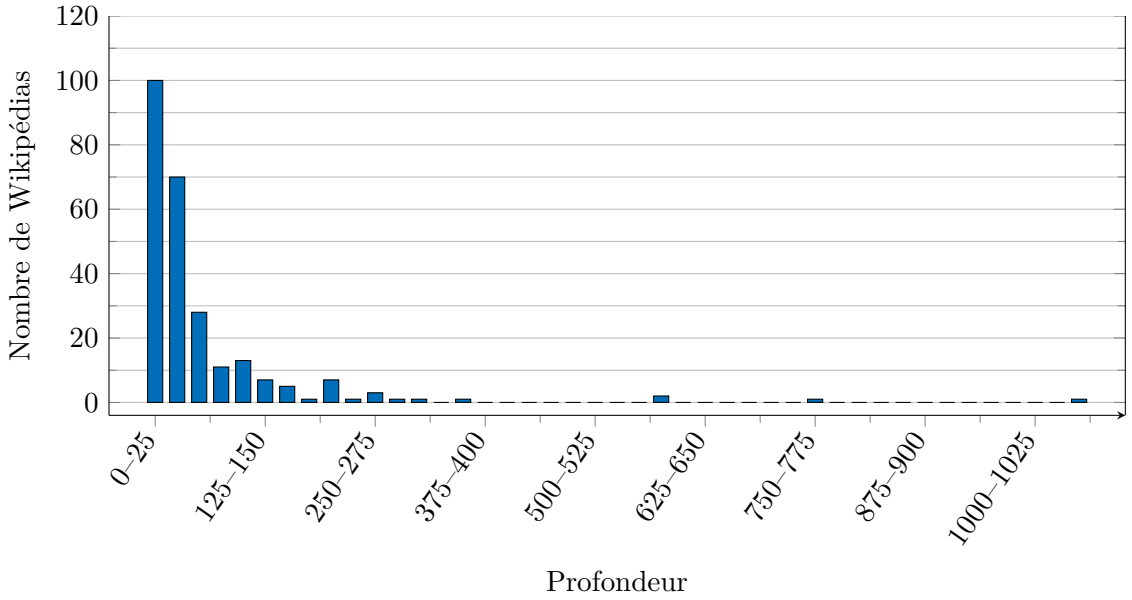


FIGURE 2.2 – Répartition des Wikipédias par profondeur.

figure 2.2 montre la répartition des Wikipédias selon leurs scores de `depth`.

	Taille (Rang (Nb. articles))	**depths**	**Locuteurs natifs (Nb. approx.)**	**Utilisateurs actifs ***
Anglais	1 (6 013 707)	991	379M	137 409
Cebuano	2 (5 378 563)	**2**	15M	148
Suédois	3 (3 738 252)	**7**	10M	2 759
Français	5 (2 208 537)	235	110M	22 023
Winaray	11 (1 263 914)	**4**	2.6M	65
Aragonais	100 (36 706)	63	10 000	76
Vepse	167 (6 369)	39	1 500	23
Hawaïen	195 (3 839)	8	20 000	14
Francique ripuaire	217 (2 860)	1 063	900 000	14

* les « utilisateurs actifs » sont les utilisateurs inscrits qui ont fait au moins une modification dans les 30 derniers jours.

TABLEAU 2.1 – Comparaison de huit projets Wikipédia.

Les statistiques de chaque Wikipédia présentées dans le tableau 2.1 sont fournies par Wikimédia dans l'article `List of Wikipedias`[40]. Le nombre approximatif de locuteurs par langue est l'estimation fournie par ETHNOLOGUE (Eberhard *et al.*, 2019) ou est celui renseigné sur la page Wikipédia de la langue concernée.

On observe dans ce tableau que trois des 10 premières Wikipédias contenant le plus d'articles présentent un score de `depth` inférieur à 10, soit très faible.

40. Voir : https://meta.wikimedia.org/wiki/List_of_Wikipedias.

L'utilisation de la traduction pour créer de nouveaux articles dans une langue mais aussi de *bots* tels que le robot `Lsjbot` [41], qui crée des articles pour les Wikipédias en suédois, en cebuano, en winaray et en néerlandais expliquent ce phénomène (voir notamment (Guldbrandsson, 2013) dans le cas de la Wikipédia suédoise).

Le tableau montre également que les membres de petites communautés linguistiques telles que les communautés aragonaises ou vepsiennes se sont emparées de l'opportunité offerte par Wikipédia pour développer leur présence numérique. L'article `Liste des Wikipédias par nombre de locuteurs par article` [42] permet d'identifier les Wikipédias les plus importantes compte tenu de la taille de leurs communautés linguistiques respectives.

2.2.3.2 Langues et orthographes de Wikipédia

Dans sa `Politique de proposition de langue`, Wikimedia (2020) stipule que chaque Wikipédia doit correspondre à une langue présentant un Code ISO 639 1-3 [43] valide, ou, dans des cas exceptionnels, une balise de langage BCP 47 [44] uniquement.

Il est de plus stipulé :

> « La langue doit être suffisamment unique pour ne pas pouvoir coexister avec une autre sur un wiki plus général. Dans la plupart des cas, cela exclut les dialectes régionaux et les différentes formes écrites de la même langue. » [45]

Cette définition des langues acceptées conduit à des Wikipédias contenant des articles écrits dans des dialectes étroitement liés. C'est le cas, par exemple, de la « *Alemannischi Wikipedia* » [46], dans laquelle chaque article est associé à une catégorie correspondant à sa langue d'écriture (*Schwyzerdütsch* (suisse alémanique), *Badisch* (badois), *Elsassisch* (dialectes alsaciens), *Schwäbisch* (allemand souabe) ou *Vorarlbergisch* (dialecte autrichien parlé au Vorarlberg)).

Il en est de même pour les Wikipédias du limbourgois [47] (qui distingue les articles en limbourgeois du Pays-Bas, *Nederlands Limburg*, et de Belgique, *Belsj Limburg*) et de la Wikipédia du bas saxon des Pays-Bas [48] (qui classe les articles dans 13 catégories dialectales différentes [49]).

Il existe d'autres exemples de Wikipédias multi-dialectales (par exemple la Wikipédia Bihari, qui couvre plus de dix dialectes parlés en Inde et au Népal, ou la Wikipédia occitane couvrant un continuum de dialectes romains parlés dans quatre pays), mais, à notre connaissance, seules les trois Wikipédias sus-mentionnées catégorisent explicitement leurs articles.

Concernant les conventions graphiques, il ne semble pas y avoir de politique globale, et chaque Wikipédia traite le cas de sa langue différemment. Par exemple, certaines pages de la sec-

41. Voir : https://fr.wikipedia.org/wiki/Lsjbot, juin 2020.
42. `List of Wikipedias by speakers per article`, voir https://meta.wikimedia.org/wiki/List_of_Wikipedias_by_speakers_per_article, juin 2020.
43. Voir : https://iso639-3.sil.org/, juin 2020.
44. Voir : https://tools.ietf.org/html/bcp47, juin 2020.
45. « *The language must be sufficiently unique that it could not coexist on a more general wiki. In most cases, this excludes regional dialects and different written forms of the same language.* » (Wikimedia, 2020).
46. Voir : https://als.wikipedia.org/wiki/Wikipedia:Houptsyte, juin 2020.
47. Voir : http://li.wikipedia.org, juin 2020.
48. Nous remercions D. van Esch qui nous a orientée vers les Wikipédias limbourgeoise et du bas saxon des Pays-Bas (communication personnelle du 20 juillet 2020).
49. Voir la section `Artikels up dialekgrup` de la page http://nds-nl.wikipedia.org, juin 2020.

tion alsacienne de la Wikipédia Alémanique sont écrites en suivant les recommandations Orthal (Crévenat-Werner et Zeidler, 2008), d'autres ne les suivent pas. *A contrario*, il est recommandé sur l'incubateur Wikimédia pour le créole mauricien (code ISO 639-3 `mfe`) [50] d'utiliser la graphie standard :

> « Merci d'utiliser l'orthographe standardisée à jour pour le créole mauricien. Certaines pages ont déjà été écrites de manière « non-standardisée » et doivent être remplacées » [51].

La Wikipédia en arabe égyptien, ou masri (code ISO 639-3 `arz`), est la seule Wikipédia qui soit rédigée dans un dialecte de l'arabe. Elle est écrite en grand majorité avec les caractères arabes. Néanmoins, une page fait l'inventaire des articles écrits avec l'alphabet latin, à destination des « personnes qui peuvent parler le masri mais n'écrivent qu'en alphabet latin » [52].

2.2.3.3 Wikipédia comme corpus

La proportion d'articles d'une Wikipédia qui est utile au niveau linguistique est difficile à évaluer. Par exemple, un examen manuel de la section alsacienne de la Wikipédia alémanique a montré que 97 % des 1 860 articles (environ 200 000 mots) étaient des articles d'une phrase décrivant une ville ou un lieu géographique. Une fois ces articles filtrés, la taille de cette Wikipédia est tombée à 50 000 mots.

Les Wikipédias présentent en général un contenu de qualité, en ce sens qu'il est bien formé et peu bruité. La contrepartie de cette qualité est que la contribution à une Wikipédia est une tâche exigeante.

La qualité attendue par la nature encyclopédique du projet Wikipédia, ainsi que l'environnement structuré et d'apparence officielle lui tenant lieu peuvent représenter un obstacle pour les contributeurs potentiels.

Bien que de petites communautés linguistiques se soient emparées de ce projet, il est peu vraisemblable que les articles encyclopédiques soient le type de contenu le plus naturel à produire pour des locuteurs dont la pratique de l'écrit est relativement récente.

De plus, comme commenté par R. Gerbert, coordinateur de Wikimédia France, le développement de Wikipédias pour des langues plus petites se heurte à l'obstacle des sources nécessaires pour soutenir le propos des articles. En effet, pour certaines langues il n'existe pas suffisamment de littérature au sens large.

Il semble en effet que le développement de certaines Wikipédias contrevient au rappel effectué dans la `Politique de proposition de langue`, Wikimedia (2020) :

> « La fondation Wikimédia n'a pas pour but de développer de nouvelles entités linguistiques ; un corpus important doit pré-exister pour chaque langue. » [53]

50. Voir : https://incubator.wikimedia.org/wiki/Wp/mfe/Main_Page, juin 2020.
51. « *Please use the correct up-to-date standardized spelling of the Mauritian Creole language. Some pages have already been written in as "unstandardized" spelling which need to be replaced.* ».
52. « *people who can speak Masry but can only write in the Latin alphabet* », voir `Introduction in English`, sur la page https://arz.wikipedia.org, juin 2020.
53. « *The Wikimedia Foundation does not seek to develop new linguistic entities ; there must be an extensive body of works in that language.* » (Wikimedia, 2020).

Enfin, il est difficile pour les petites Wikipédias de faire face à la croissance des plus grandes, donc d'être compétitives en termes d'intérêt pour leurs utilisateurs dans des contextes bilingues. Cela a par exemple été rapporté dans The Digital Language Diversity Project (2017) au sujet des locuteurs du sarde utilisant en majorité la Wikipédia italienne.

Il existe des initiatives intéressantes pour surmonter cette difficulté en exploitant les articles écrits dans la langue majoritaire. C'est par exemple le cas de l'expérience présentée par Alegria *et al.* (2013), où les Wikipédias espagnoles et basques sont utilisées comme corpus. Les articles en espagnol sont traduits automatiquement puis corrigés par des bénévoles, ce qui permet d'étendre semi-automatiquement la Wikipédia basque tout en améliorant la qualité de l'outil de traduction automatique.

2.2.4 Corpus produits sur les réseaux sociaux

En tant qu'espace d'expression croissant, les réseaux sociaux peuvent être considérés comme une source de corpus précieuse.

Les contenus produits par les utilisateurs de Twitter et Facebook sont probablement plus représentatifs de l'utilisation conversationnelle, mais ils ne représentent pas des sources durables de corpus textuels. Par exemple, Facebook n'autorise pas la libre utilisation des données, et le consentement de tous les participants doit être demandé. Cela a été précisé dans un message sur la liste `CORPORA`[54] par E. Ringger, le 27 octobre 2015. Le cas de Twitter est différent, car les *tweets* sont des textes courts, qui pourraient être considérés comme des citations et donc plus faciles à utiliser. Cependant, cette autorisation d'utilisation ne s'applique pas aux créations artistiques, telles que les haïkus, de sorte que les *tweets* doivent être vérifiés manuellement. Par ailleurs et pour éviter les problèmes de droits d'auteur, les *tweets* sont souvent désignés par un identifiant. Or, ceux-ci peuvent être supprimés ou modifiés par leurs créateurs après une opération de collecte, ce qui génère des écarts entre le contenu publié et le contenu collecté.

Par ailleurs, d'un point de vue technique, et bien que les *tweets* puissent parfois être géolocalisés, il est toujours nécessaire d'identifier la langue de chacun des contenus (Graham *et al.*, 2014; Lui et Baldwin, 2014; Williams et Dagli, 2017). Or l'identification linguistique sur les réseaux sociaux peut être rendue mal aisée par la nature conversationnelle des écrits et les pratiques d'alternance codique (*code-switching*) des utilisateurs multilingues[55] (Lüpke, 2011).

Le projet `Nierika`[56] s'attaque directement à ces questions et mise sur le consentement éclairé des participants et l'utilisation de méta-données informatives. Ce projet, qui a été présenté à LT4ALL en décembre 2019 à Paris, vise en effet à utiliser les réseaux sociaux pour collecter des ressources linguistiques, tout en abordant la question du consentement et du respect de la vie privée. À notre connaissance, aucun résultat concernant le projet n'a été publié jusqu'à présent.

54. Voir : https://mailman.uib.no/listinfo/corpora, juin 2020.
55. « *In particular on social media, where users cultivate a register that is written but very reminiscent of oral communication, multilingual speakers draw in very creative fashion on their entire multilingual repertoire, but in a way that defies easy categorisation of language.* » (Lüpke, 2011).
56. Voir : https://vaniushar.github.io/about.

2.2.5 Conclusions

L'obtention ou la constitution de corpus bruts pour des langues peu dotées utiles pour leur traitement automatique représente un enjeu de taille, et le Web est le lieu de l'expression d'une grande diversité linguistique, il représente une opportunité pour des langues dont la présence scripturale se limite parfois à leur usage numérique. S'ils représentent une source précieuse de données, les corpus « opportunistes » issus de la collecte de données existantes sur le Web présentent néanmoins un certain nombre d'inconvénients, dont notamment :

- une couverture trop faible pour servir de base à des développements linguistiques ultérieurs, en particulier concernant la représentativité linguistique. Il est en particulier peu probable que des corpus dits « opportunistes » soient équilibrés quant aux variantes scripturales pouvant exister pour une langue par exemple.
- une disponibilité relative : la nature et les licences des contenus présents sur le Web requièrent parfois, pour être convertis en corpus à proprement parler, des opérations qui aboutissent à la perte d'informations précieuses, telles que certaines métadonnées ou la structure du document.
- la nécessité de procéder à l'identification de la langue qui n'est pas toujours connue au moment de la collecte. Cela peut requérir l'exploitation de ressources additionnelles, dont l'existence n'est pas garantie.
- des problèmes de qualité : il ne paraît pas souhaitable par exemple qu'un corpus issu d'une traduction soit considéré comme appartenant au corpus d'une langue donnée sans que cela soit signalé.

De manière générale, lorsqu'on essaye d'appliquer des méthodes fonctionnant pour des langues bien dotées à des langues peu dotées et en particulier non standardisées, on se heurte à de nouvelles difficultés qui mettent en lumière les limites des postulats sous-jacents d'une telle démarche, notamment, comme nous l'avons vu, la nécessité de pouvoir distinguer automatiquement les langues entre elles. Par ailleurs, il existe un risque à considérer une langue comme un ensemble homogène : une variante peut se retrouver sur-représentée dans un corpus ce qui peut poser des problèmes de performances, comme nous le décrivons dans la section 6.3.2.

Il parait par conséquent intéressant de disposer des ressources linguistiques auxquelles une diversité de locuteurs ont participé. Nous présentons dans la section 6.3 nos expérimentations à ce sujet.

2.3 Constitution et exploitation de corpus annotés

Dans cette section, nous présentons comment les ressources linguistiques enrichies que constituent les corpus annotés sont d'une part constitués grâce à l'annotation *manuelle*, et d'autre part utilisés pour l'apprentissage d'outils d'annotation *automatique*. Nous considérons principalement la tâche d'annotation morpho-syntaxique, sur laquelle nous nous sommes penchée dans le cadre de ce travail.

Bien que les systèmes récents de traitement automatique tendent vers de moins en moins de supervision (Lample *et al.*, 2017; Grave *et al.*, 2018), les technologies de traitement reposent encore largement sur l'existence de corpus de textes annotés. Par ailleurs, même si le développement de certains outils peut être fait sans requérir de ressources annotées, comme c'est le cas

des méthodes par règles ou des modèles d'apprentissage non supervisé, les ressources annotées sont toujours nécessaires pour *évaluer* ces outils.

2.3.1 L'annotation *manuelle* de corpus

Le contenu de cette section est inspiré du premier chapitre de l'ouvrage Fort (2016).

Par définition, un corpus annoté est l'association d'un ensemble de ressources linguistiques brutes, le *corpus*, et d'un *schéma d'annotation* appliqué sur celles-ci. L'activité d'annotation est elle-même une activité *interprétative* (Leech, 1997; Habert, 2006).

La dualité entre les ressources et le schéma amène à concevoir l'annotation comme un processus agile (Voormann et Gut, 2008) dont les itérations successives permettent d'affiner la connaissance du corpus et de raffiner le schéma en fonction. Dans le cas d'une annotation morpho-syntaxique, cela peut donc se traduire par « une représentation du corpus plus précise » et « un étiqueteur morpho-syntaxique plus sophistiqué ».

2.3.1.1 Utilisation de la pré-annotation

Un des enjeux propres aux tâches linguistiques à accomplir est la réduction de la *complexité* de la tâche d'annotation. Fort *et al.* (2012) la définit selon six axes : la *discrimination* des unités à annoter, la *délimitation* des unités à annoter, l'expressivité du langage d'annotation, la dimension du jeu d'étiquettes, l'ambiguïté et le contexte à prendre en compte.

En proposant une pré-annotation de qualité suffisante, il est possible d'agir efficacement sur la dimension du jeu d'étiquettes. Cela a été démontré sur la tâche d'annotation morpho-syntaxique par Dandapat *et al.* (2009) et Fort et Sagot (2010) qui montrent notamment qu'un corpus de 50 phrases peut suffire à entraîner un outil de pré-annotation permettant de réduire notablement (de l'ordre le la division par trois) le temps d'annotation humain tout en conservant la qualité (notamment l'exactitude des annotations produites et les accords inter-annotateurs) de cette annotation manuelle.

2.3.1.2 Accord inter-annotateur et coefficient Kappa

Pour la construction de corpus annotés de référence, il est d'usage de calculer des accords inter-annotateurs pour faire converger les annotations vers une référence (Fort, 2012).

L'accord entre deux annotateurs peut être calculé grâce au coefficient Kappa de Cohen définit par Cohen (1960). Au-delà de l'accord observé, ce coefficient permet de prendre en compte qu'une partie de l'accord est due au hasard, et que l'utilisateur peut être biaisé. En effet, les utilisateurs peuvent annoter selon leur propre interprétation des consignes. On cherche donc à ne considérer que la proportion d'accord atteinte se trouvant au-dessus du hasard.

Sont ainsi calculés pour deux annotateurs Ann_1 et Ann_2 :

- l'accord observé A_0,

- l'accord attendu $A_e = \sum_q \frac{n_{Ann_1\_q}}{i} . \frac{n_{Ann_2\_q}}{i}$

  avec $\begin{cases} i = \text{nombre d'items} \\ n_{Ann_x\_q} = \text{probabilité que l'annotateur} Ann_x \text{choisisse la catégorie } q \end{cases}$

Le coefficient kappa est par la suite calculé tel que $\kappa = \frac{A_0 - A_e}{1 - A_e}$.

Si Mathet *et al.* (2012) ont cherché à faire correspondre une réalité à ces métriques, établir un seuil au-delà duquel un accord serait considéré comme bon reste hasardeux. Artstein et Poesio (2008) fixent un seuil arbitraire de 0,8 au-delà duquel un accord inter-annotateur peut être considéré comme satisfaisant.

2.3.2 Apprentissage et annotation *automatique*

L'apprentissage automatique est utilisé en TAL pour une grande diversité de tâches (Norvig, 2017). L'application de ces méthodes à une diversité de langues nécessite de disposer de ressources suffisantes. Lorsque les ressources manquent, il est possible de tirer parti de caractéristiques de la langue à traiter ou de l'existence d'autres outils ou ressources.

2.3.2.1 Stratégies pour pallier le manque de ressources

De nombreuses stratégies peuvent être envisagées pour pallier le manque de ressources requises par un apprentissage strictement supervisé. C'est le cas des approches *semi-supervisées* qui reposent sur l'utilisation conjointes de corpus annotés et d'autres ressources ou outils existant.

Certaines d'entre elles tirent parti de corpus parallèles qui permettent la projection d'annotations (Agić *et al.*, 2016), c'est-à-dire un « transfert » de connaissances d'une langue bien dotée vers une langue moins dotée. Zennaki *et al.* (2016) tirent parti de l'existence de corpus bilingues pour entraîner des outils d'annotation non supervisés *via* une architecture de réseaux neuronaux.

Une autre méthode consiste à tirer parti de la proximité étymologique et morphologique pouvant exister entre une langue bien dotée et une qui l'est moins (Hana *et al.*, 2004; Scherrer et Sagot, 2013). Il est ainsi possible d'identifier des couples de cognats entre ces langues proches et de traduire aisément les mots dits « outils » dans la langue bien dotée ce qui améliore la performance des outils existant lorsqu'ils sont appliqués sur la langue peu dotée. Cette approche est également celle de (Bernhard et Ligozat, 2013) pour la création d'outils pour l'alsacien à partir de l'allemand.

D'autres intègrent des outils additionnels, tels que des transducteurs à états finis pour analyser les mots inconnus, comme décrit par Garrette *et al.* (2013), ou de ressources externes pour complémenter un corpus annoté de taille réduite, telles que le `Wiktionnaire` dans le cas de (Li *et al.*, 2012) ou (Täckström *et al.*, 2013) qui tire parti conjointement de l'existence de corpus bilingues alignés et de lexiques frustes [57].

L'apport d'un lexique pour compenser la petite taille d'un corpus d'entraînement a été démontrée par divers travaux : par exemple, Vergez-Couret *et al.* (2014) a montré que `Talismane` (Urieli, 2013) peut être entraîné pour l'occitan et atteindre une précision de 89 % avec un corpus

57. L'exactitude des outils d'annotation obtenus n'a pu être testée que sur des langues bien dotées pour lesquelles il existe des corpus de référence annotés.

d'entraînement de 2 500 mots, grâce notamment à l'intégration d'un lexique de 225 000 entrées. Les expériences de Sagot (2016) au cours desquelles `MElt` a été entraîné pour une trentaine de langues confirment l'intérêt d'utiliser un lexique lorsque les données d'entraînement sont de petite taille. Par exemple, dans le cas de l'estonien, les performances de `MElt` gagnent 4,8 points pour atteindre une exactitude de 94,4 lorsque l'outil est entraîné avec un corpus de 7 687 mots et un lexique complémentaire de 135 095 mots. Dans le cas du roumain, le gain est de 3,3 points et l'exactitude atteint 94,35 avec un corpus de 9 291 mots et un lexique complémentaire de 428 194 entrées.

Une approche de plus en plus employée consiste à tirer parti des progrès des approches neuronales basées sur la disponibilité d'une grande quantité de ressources brutes pour la construction de plongements lexicaux modélisant la distribution des mots dans un corpus. Les travaux de Han et Eisenstein (2019) montrent comment ceux-ci peuvent être utilisés pour adapter à des contextes linguistiques variés des modèles ayant pu être entraînés sur une quantité de corpus suffisante correspondant à un contexte linguistiquement proche et mieux doté. Ils présentent notamment une expérience d'annotation en parties du discours d'anglais historique et d'un corpus de l'anglais issu de Twitter à partir d'un modèle pré-entraîné sur de l'anglais davantage canonique. Notons que le transfert d'annotations peut bénéficier des contenus structurés tels que les projets Wikipédia. Pan *et al.* (2017) en tirent par exemple parti pour transférer des annotations en entités nommées à partir de corpus annotés en anglais et pour 282 langues. Cette approche est étendue par Pfeiffer *et al.* (2020) à deux autres tâches d'annotations et à un ensemble de langues très peu dotées absentes des jeux de données multilingues à disposition.

Nous nous attardons ici sur la méthode proposée par Magistry *et al.* (2018) reposant sur une architecture du type Bi-LSTM et dont nous présentons dans la section 7.2 une expérience de reproduction menée avec un des auteurs de l'article original. Dans ce travail, lorsque les données brutes sont insuffisantes pour couvrir le vocabulaire cible, la construction de plongements lexicaux pour les mots hors vocabulaire s'opère en tirant parti de la graphie de ceux-ci, grâce à des modèles d'apprentissages fonctionnant au grain caractère. C'est le cas du système Mimick (Pinter *et al.*, 2017) que Magistry *et al.* (2018) utilisent par exemple pour entraîner ce qu'ils appellent les « plongements morphosyntaxiques » (*MorphoSyntactic Embeddings* (MSE)) dans une expérience où les plongements sont spécialisés à la tâche d'annotation en parties du discours. En particulier, l'apprentissage des plongements du corpus annoté est poussé à prendre en compte des paramètres utiles à l'analyse morphosyntaxique (morphèmes des mots cibles, parties du discours des mots du contexte, etc.). Magistry *et al.* (2018) ont ainsi montré qu'il est possible d'apporter une solution partielle au manque de données disponibles en utilisant la spécialisation de plongements lexicaux. Ils montrent dans le cas de l'étiquetage automatique du picard, du malgache et de l'alsacien que l'utilisation complémentaire de *MorphoSyntactic Embeddings* permet de dépasser les performances d'architectures de type supervisées classiques. Avec un corpus annoté de 12 600 mots, et un corpus brut de 200 000 mots, les performances d'étiquetage atteignent une exactitude de 0,91.

2.3.2.2 Annotation automatique de production langagière variée

Les langues non standardisées (ou *non canoniques* (Plank, 2016)) sont susceptibles de présenter des variations à tous les niveaux de l'analyse linguistique, de la phonétique à la sémantique.

La question de leur intégration se pose dans quantité de cas dépassant celui des langues peu

dotées et non standardisées, notamment celui des langues anciennes, par exemple le moyen allemand (Barteld, 2017) ou l'ancien français (Sagot, 2019), des contenus produits par des internautes (Krumm *et al.*, 2008; Seddah *et al.*, 2012) dans des cadres aussi variés que les projets `Wikipédia`, les réseaux sociaux, ou les jeux-vidéo en ligne, ou bien encore des communications médiées par ordinateur, ou *Computer Mediated Communication* (CMC) (Melero *et al.*, 2012; Chanier *et al.*, 2014). Des langues bien dotées, à l'instar du chinois mandarin (de Chine continentale, de Hong-Kong et de Taïwan) ou du portugais (brésilien et du Portugal), sont également sujettes à cette variabilité (Tseng *et al.*, 2005; Garcia *et al.*, 2014). Or, à ce jour, les outils développés et évalués pour une langue donnée sont en réalité conçus de manière peu robuste à toute forme de variation (Plank, 2016).

L'approche la plus répandue dans la littérature consiste à normaliser des corpus de manière à homogénéiser l'ensemble en en supprimant la variabilité (Ljubešic *et al.*, 2016; Samardzic *et al.*, 2015) (voir (Cox, 2010), pour une discussion sur la *rentabilité* de la normalisation). Ces techniques s'appliquent généralement lorsqu'il existe une norme majoritaire, qu'on est capable d'identifier les segments s'en écartant, et qu'on connaît les mécanismes de normalisation.

Pour identifier ces mécanismes, il est possible d'utiliser des corpus parallèles lorsqu'ils existent en taille suffisante : les transformations peuvent être déduites des segments alignés comme illustré dans le cas du basque par Etxeberria Uztarroz *et al.* (2014). Notons que la diminution de la variabilité peut également être envisagée *via* l'utilisation de techniques de translittération (Hana *et al.*, 2011; Pingali *et al.*, 2017), ou de dictionnaires de prononciation pour entraîner des modèles de transcription phonétique permettant de réduire la variabilité scripturale (Steible et Bernhard, 2018).

2.4 Conclusion

Quelle que soit la langue considérée, les ressources linguistiques demeurent une denrée coûteuse. Dans les contextes linguistiques où ces ressources manquent, il est difficile de développer des outils qui soient compétitifs avec les systèmes développés pour les langues majoritaires. Le manque de données contraint les chercheurs à imaginer diverses stratégies compensatoires, comme par exemple la combinaison de ressources de tailles réduites mais de natures variées. Le manque de ressources est ainsi d'autant plus criant que le matériau linguistique est non canonique, l'hétérogénéité des pratiques rendant toute ambition de couverture lexicale illusoire.

Quelle que soit la méthode employée pour intégrer la variation observée aux traitements effectués, celle-ci requiert d'avoir accès à une description de la variation présente dans les corpus, à une grande quantité de textes permettant d'inférer les motifs de variation, ou à suffisamment de corpus parallèles pour inférer des règles de normalisation.

Notons par ailleurs qu'en ce qui concerne les langues peu dotées non standardisées, l'un des enjeux du respect de la diversité des variétés existantes est d'éviter de faire de la création de ressources et d'outils de TAL un vecteur non intentionnel de standardisation. C'est en ce sens que nous avons mené nos expériences de collecte de ressources linguistiques, en particulier de lexiques de variantes graphiques alignées.

Chapitre 3

Productions participatives de ressources langagières

Sommaire

La myriadisation (*crowdsourcing*) s'est imposée depuis une dizaine d'années comme l'une des solutions pragmatiques aux freins que constitue le manque de moyens et de linguistes disponibles pour la construction de ressources langagières.

Afin de dresser les contours de notre travail et d'introduire les contraintes qui ont été les nôtres, nous présentons dans un premier temps (section 3.1) plusieurs travaux qui ont proposé d'expliciter les diverses composantes qui permettent de caractériser complètement une entreprise de myriadisation.

Ce tour d'horizon des taxonomies existantes nous permet de présenter les nombreux cadres envisageables pour la mise en place d'une activité de myriadisation.

Nous distinguons ensuite deux activités de production participative concernant les ressources linguistiques. Dans les deux cas, la seule compétence attendue *a priori* des locuteurs est la maîtrise de leur langue. La différence entre ces deux activités tient de la nature et de la variété du public à laquelle elle est adressée.

La première activité consiste à externaliser une tâche linguistique pouvant être considérée comme « objective » (une annotation, une traduction) et repose sur la compétence linguistique de locuteurs indifférenciés. La seconde mise sur la *diversité* des locuteurs participant afin de recueillir des données variées et représentatives de la diversité de leurs pratiques linguistiques.

Nous présentons donc dans les sections 3.2 et 3.3 les projets de myriadisation concernant ces deux types d'activités pour la production de ressources linguistiques.

3.1 Des myriadisations

3.1.1 Les dimensions de la myriadisation

Le terme *crowdsourcing* apparaît pour la première fois dans un article publié dans le magazine *Wire* (Howe, 2006). Jeff Howe en propose la définition suivante :

> « Le *crowdsourcing* est l'activité consistant à faire réaliser, *via* un appel ouvert, à un groupe non défini, généralement important, une tâche habituellement réalisée par un agent défini (généralement un employé). ». [58].

Ce terme « *crowdsourcing* » est formé sur la contraction des mots « *crowd* » (« foule ») et « *outsourced* » (« externalisée »). Aucun terme français ne correspond exactement au mot-valise anglais, nous emploierons de manière indifférenciée pour notre part les termes « production participative » et « myriadisation », proposé par G. Adda et introduit dans Sagot *et al.* (2011).

Il existe plusieurs taxonomies permettant de décrire et distinguer les différentes activités de myriadisation.

Burger-Helmchen et Pénin (2011) ont proposé de séparer les activités de myriadisation en trois catégories en fonction du type d'activité concernée. Ils définissent ainsi de manière disjointe le « *crowdsourcing* d'activités routinières », « de contenu » et « d'activités inventives ». Les premières concernent des activités simples mais trop chronophages pour être réalisées par une personne ou par un petit groupe, les secondes des activités ayant pour objectif la construction d'un contenu exhaustif, les dernières s'adressent à la résolution de « problèmes parfois complexes et/ou créatifs ». Dans le premier cas, les participants sont interchangeables au sens où c'est le caractère massif du cumul des contributions qui présente un intérêt pour le concepteur, et que l'appel ouvert est ici un moyen de passer à l'échelle le nombre de contributeurs, et par extension la quantité de données produites par unité de temps.

Dans les second et troisième cas, chaque participant apporte une contribution particulière qui lui est propre et qui vient *compléter* celles des autres plus qu'elle ne s'y *ajoute*.

Ce découpage nous paraît intéressant dans la mesure où ces différentes classes d'activités procèdent de motivations distinctes de la part de leurs concepteurs et requièrent la mise en place de processus de recrutement des participants, de conception des tâches et d'évaluation des données produites propres à chacun. Néanmoins, les tâches chronophages mais non triviales, pouvant aller jusqu'à requérir une formation ne trouvent pas de place dans la typologie proposée par Burger-Helmchen et Pénin (2011).

58. « *Crowdsourcing is the act of taking a job traditionally performed by a designated agent (usually an employee) and outsourcing it to an undefined, generally large group of people in the form of an open call.* » (Howe, 2006).

Quinn et Bederson (2011) proposent une distinction entre les activités de « *crowdsourcing* », « *social computing* » et « *human computation* », trois activités d'« intelligence collective » (*collective intelligence*). L'activité de « *crowdsourcing* » se distinguerait notamment de celle de *human computation*[59] en cela qu'elle convoquerait l'intelligence collective pour résoudre une tâche habituellement réalisée par un *humain* et non par une *machine*.

Geiger *et al.* (2011) proposent une étude détaillée des taxonomies existantes dont il tire une liste de composantes organisées selon quatre axes : (i) la nécessité ou non de préselectionner les contributeurs et selon quels critères, (ii) l'accessibilité des contributions (pour les contributeurs), (iii) la méthode d'agrégation des contributions, (iv) le type de rétribution.

Fort (2016) propose d'organiser les entreprises de production participatives selon deux dimensions propre au rapport établi avec le contributeur : i) la rétribution (ou l'absence de rétribution) des participants et ii) le caractère transparent (ou non) de la dimension participative. Cette typologie permet de distinguer le « *crowdsourcing* bénévole » (par exemple `Wikipédia`), les plateformes de travail parcellisé *microworking* qui proposent la réalisation de tâches transparentes en échange d'une micro-rémunération (par exemple `Amazon Mechanical Turk`), et les jeux ayant un but, qui dissimulent plus ou moins la tâche à réaliser sous une couche de fonctionnalités ludiques (Lafourcade *et al.*, 2015). Dans ce dernier cas, les récompenses accessibles aux participants *via* les fonctionnalités ludiques peuvent s'accompagner de rétribution sous forme de cadeaux, de bons d'achats etc.

Ces taxonomies donnent un aperçu des nombreuses formes que peut prendre une activité de myriadisation. Elles font émerger trois axes principaux que sont :

1. la tâche, ou le type de tâche, à (faire) réaliser,
2. la nature du public auquel elle est proposée,
3. les conditions dans lesquelles elle est réalisée (par exemple le type de rétribution ou l'accès aux contributions).

Naturellement, ces axes ne sont pas indépendants les uns des autres et chaque choix de conception doit prendre en compte la description de l'activité dans sa globalité. La conception (où le choix) d'une *plateforme* permettant de faire réaliser une *tâche* donnée à une *foule* donnée implique par conséquent de répondre à des questions transverses telles que : Comment concevoir la tâche pour qu'elle soit effectivement réalisable par la foule visée ? Comment la rétribution s'organise-t-elle compte tenu de la conception de la tâche ? Comment les conditions de réalisation de la tâche influent-elles sur la motivation et les performances des participants ?

Certaines tâches associées à un certain public et un certain type de rétribution peuvent en effet conduire certains participants à tricher. Ainsi, dans le cas de jeux ayant un but, « trouver le degré maximal de « ludification » est crucial au regard de la motivation suscitée chez le joueur »[60] (Lee *et al.*, 2013). Les travaux de Morschheuser *et al.* (2019) montrent par ailleurs qu'il est plus rentable dans certaines situations de jouer sur la collaboration que sur la compétition entre participants pour les motiver. Attirer et retenir les participants sur ces plate-formes constituent un exercice complexe (Munro, 2013; Tuite, 2014), encore qualifiable d'« alchimie ».

Répondre à ces questions afin que le concepteur et le participant tirent le meilleur parti possible de chaque participation fait donc appel à des compétences multiples. Ces compétences incluent

59. « *a paradigm for utilizing human processing power to solve problems that computers cannot yet solve* », définition proposée par L. von Ahn dans sa thèse (von Ahn, 2005).

60. « *Finding the maximum safe level of gamification is crucial for motivational design.* » (Lee *et al.*, 2013).

la conception au sens large (c'est-à-dire de la tâche, de la plateforme, du système de rétribution) mais aussi le recrutement des participants et le dialogue avec eux, et l'analyse des participations.

3.1.2 Agrégation des contributions

Nous nous limitons dans cette section à présenter les techniques d'agrégation pour des tâches d'annotation ou de catégorisation telles que l'annotation en parties du discours ou la reconnaissance d'entités nommées. En particulier, nous ne présentons pas ici les méthodes qui peuvent être envisagées dans le cadre de traductions collaboratives.

En fonction de l'objectif poursuivi et de la méthodologie mise en place, il peut être nécessaire d'agréger des contributions redondantes produites par différents participants.

C'est particulièrement le cas dans des activités de type « externalisation d'une tâche linguistiques », en particulier des activités d'annotation séquentielle ou de classification qui supportent peu d'ambiguïté et pour lesquelles une valeur unique objective est généralement attendue. Plusieurs modèles ont été proposés pour permettre au concepteur d'une telle tâche de trancher en faveur d'une catégorie lorsqu'un désaccord est observé parmi les participants. Dans le cas d'une tâche d'annotation pour laquelle les étiquettes sont non ambigues et non subjectives cela revient à trouver comment donner plus de poids aux annotateurs fiables et moins de poids à ceux qui le sont moins.

Les méthodes traditionnelles d'agrégation consistent à prendre l'option majoritaire (« *majority voting* ») ou d'opérer des traitements statistiques comme le calcul de la valeur moyenne des propositions lorsque c'est possible. Ce sont par exemple les méthodes employées dans la grande majorité des travaux comparés par Sabou *et al.* (2014), qui concernent une diversité de tâches d'annotations, d'entités nommées, d'opinions etc.

Certains modèles probabilistes plus élaborés cherchent à intégrer des connaissances sur les contributeurs ou la tâche elle-même afin de déterminer la catégorie la plus probable et éventuellement d'identifier les biais des annotateurs. Le modèle MACE (Hovy *et al.*, 2013) propose d'utiliser des informations comme l'expertise du participant, la distribution attendue des étiquettes, afin de déterminer pour chaque annotation quel annotateur est le plus digne de confiance. D'après Paun *et al.* (2018), qui propose une comparaison de différents modèles bayésiens d'agrégation, MACE est la méthode la plus précise pour inférer la bonne étiquette lorsque la tâche peut être mesurée à une référence.

L'agrégation d'annotations myriadisées pour le TAL est un domaine actif, en témoigne le premier atelier AnnoNLP ayant eu lieu en 2019 [61], dont l'objectif a visé en particulier à questionner l'intégration de certaines propriétés des annotations linguistiques (par exemple la dépendance entre certaines étiquettes ou l'ambiguïté acceptée dans certains cas précis) aux techniques d'agrégation.

Dans le cas des activités de myriadisation visant à obtenir des données linguistiques variées, les contributions sont agrégées par simple accumulation des propositions formulées par les participants.

61. Voir : http://dali.eecs.qmul.ac.uk/annonlp.

3.2 Externalisation de tâches linguistiques : la myriadisation pour pallier le manque d'experts

Dans cette section, nous présentons la myriadisation comme moyen d'accéder à une « foule » de participants nécessaire à la réalisation d'une tâche linguistique précise par le biais de son externalisation.

La production participative s'est ainsi imposée depuis une quinzaine d'années comme l'une des solutions aux freins que constituent les coûts important de constitution de ressources linguistiques (Wang *et al.*, 2013).

Les tâches linguistiques concernées sont variées. La myriadisation est ainsi invoquée aussi bien pour résoudre des tâches presque triviales pour l'humain mais très ardue pour la machine (voir par exemple le tour d'horizon de l'utilisation de la myriadisation pour la reconnaissance de caractères sur des documents numérisés (Andro et Saleh, 2017)), que des tâches linguistiques complexes telles que l'annotation en syntaxe en dépendance (Guillaume *et al.*, 2016).

Les trois dimensions présentées dans la section 3.1.1 offrent de nombreuses possibilités pour la mise en place d'une activité de myriadisation, c'est-à-dire la mise en place des *conditions* permettant de faire réaliser une *tâche* à un *public* choisi.

La capacité à motiver des locuteurs à participer est cruciale : il convient dans un premier temps d'attirer et dans un second temps de retenir les participants. Quelle que soit la tâche à effectuer et les communautés linguistiques visées, des mécanismes incitatifs doivent ainsi être mis en place. Dans le cas de productions linguistiques pour le TAL, l'ensemble du spectre de rétribution proposé par Fort (2016) est exploré. Nous proposons ici de présenter les activités de myriadisation pour le TAL en les distinguant selon deux groupes principaux correspondant à des stratégies incitatives différentes. Nous présentons comment le cadre dans lesquelles elles sont mise en place influe sur la tâche à laquelle elles peuvent être appliquées, sur les communautés linguistiques qu'elle permet de toucher et sur la flexibilité quant aux conditions de réalisation.

Nous présentons ainsi d'abord en section 3.2.1 les applications misant sur une compensation financière, c'est-à-dire qui passent par le recrutement de participants en échange d'une (micro) rémunération. Puis nous présentons en section 3.2.2 les applications qui misent sur une compensation autre que financières, des applications bénévoles recourant à des mécanismes de récompenses variées.

Enfin, en section 3.2.3 nous présentons comment ces deux stratégies, correspondant à des choix techniques différents, peuvent être appliquées dans le cas de langues peu dotées.

3.2.1 Recrutement de participants et plateformes de *microworking*

À notre connaissance, le travail parcellisé (*microworking*) est mis en place *via* des plateformes dédiées telles que `CrowdFlower`[62] ou Amazon Mechanical Turk (Callison-Burch et Dredze, 2010).

Le travail parcellisé à la `Amazon Mechanical Turk` ou `CrowdFlower` permet pour certaines langues d'accéder à une importante masse de travailleurs et de faire réaliser très rapidement des microtâches (HIT, *Human Intelligence Tasks*), ce type de plateforme ne permet pas de

62. Désormais `Figure Eight`, voir : https://www.figure-eight.com/.

trouver plus facilement des experts pour toutes les langues (Callison-Burch et Dredze, 2010). Par ailleurs, outre les problèmes éthiques qu'elles peuvent soulever, ces plateformes posent des problèmes de qualité produite (Fort *et al.*, 2011).

En effet, il est actuellement impossible, sur ces plateformes, de former les travailleurs à la tâche souhaitée (il n'est possible que de les évaluer). Or, une tâche complexe comme l'annotation en parties du discours requiert une formation : en témoignent les résultats obtenus par Hovy *et al.* (2014) pour l'annotation *via* `CrowdFlower`, de *tweets* en anglais (84 % d'exactitude). De manière similaire, Jamatia et Das (2014) et Zaghouani et Dukes (2014) déplorent la faible qualité des annotations obtenues *via* `Amazon Mechanical Turk`, respectivement sur des *tweets* en hindi (moins de 60 % d'exactitude) et sur un chapitre du Coran (63,91 % d'exactitude).

Cette limitation n'existe pas dans les autres modes de production participative et de nombreuses plateformes imposent une formation (plus ou moins longue) préalable à la participation effective. C'est notamment le cas des `Distributed Proofreaders`[63], de `Phrase Detectives` ou de `ZombiLingo`. Les résultats ainsi obtenus en termes de qualité et de quantité de données produites sont tout à fait satisfaisants (Poesio *et al.*, 2013; Guillaume *et al.*, 2016; Madge *et al.*, 2019b).

3.2.2 Myriadisation bénévole

Le choix de mécanismes incitatifs autres pousse au développement d'applications dédiées qui offrent davantage de flexibilité dans les fonctionnalité ludiques mises en place.

L'intégration de tâches au sein d'applications ludiques misant sur le divertissement des participants a été largement exploré. Certaines applications développent des fonctionnalités ludiques autour de la tâche à réaliser, comme dans le cas de `ZombiLingo` (Guillaume *et al.*, 2016) ou `Phrase Detective` où le participant voit chacune de ses contributions récompensées par des points. D'autres sont conçus comme des jeux à part entière, la contribution des participants devenant idéalement un effet de bord du temps passé à jouer. C'est le cas de `JeuxDeMots` (Lafourcade et Joubert, 2008) ou de `Puzzle Racer` et `Ka-boom!` qui vont jusqu'à reprendre les codes du jeu d'arcade (Jurgens et Navigli, 2014). Comme nous l'avons vu dans la section 3.1, il est important de trouver la bonne quantité d'éléments ludiques. Il est également important que la « jouabilité » n'impacte pas la qualité des données produites. C'est ce constat qui amène Dziedzic (2016) a développer `RoboCorp` qui propose des jeux F2P (*free to play*, soit « gratuits à jouer »), c'est-à-dire des jeux éprouvés par des concepteurs jeux-vidéo dans lesquels des fonctionnalités peuvent être débloquées grâce à des micro-paiements. Dans le cas de `RoboCorp`, les concepteurs remplacent ces micro-paiements par des tâches de reconnaissance d'entités nommées du polonais. L'activité principale du participant n'est donc plus dans ce cas la réalisation de la tâche mais le divertissement que celle-ci permet. Ce jeu a inspiré Kicikoglu *et al.* (2019) à développer le « paradigme motivation-annotation » dans lequel les phases de divertissement et d'annotation s'alternent. La plateforme `Wormingo` illustre ce paradigme : des mini-jeux du type du jeu du pendu sont couplés avec des phases d'annotation d'anaphores en anglais.

Au cours des dernières années les recherches concernant l'intégration de fonctionnalité ludiques pour l'annotation de tâches linguistiques se sont en effet diversifiées. Notamment, Madge *et al.* (2019a) proposent une plateforme d'annotation inspirée de la notion de « jeu incrémental » : plus le participant joue, plus il accumule des points ou des monnaies virtuelles qu'il peut investir pour

63. Les `Distributed Proofreaders` corrigent les livres numérisés du `Projet Gutenberg` de mise à disposition de livres libres de droits : https://www.pgdp.net/c/.

débloquer de nouvelles fonctionnalités, celles-ci n'influant pas nécessairement sur la progression dans le jeu. Parmi ces types de jeux, les *clicker games* proposent une très forte interactivité au joueur dont la tâche se limite à cliquer ou non sur un élément de jeu. La répétitivité de la tâche est camouflée par une évolution artificielle dans le jeu.

À l'instar des plateformes génériques de sciences participatives telles que Crowd4U [64] ou Zooniverse [65], le portail Lingo Boingo [66] recense depuis 2017 les activités de myriadisation ludique de ressources linguistiques.

3.2.3 Application au cas des langues peu dotées

Les entreprises de production participative en termes d'externalisation d'une tâche linguistique ayant été mis en place pour la production de ressources pour des langues peu dotées sont peu nombreuses à notre connaissance. Lewis et Yang (2012) montrent qu'il est possible de produire de manière participative un corpus de phrases parallèles à partir duquel un outil de traduction anglais / Hmong est développé. Peu d'informations sont données sur les profils et degrés d'implication des différents participants. Munro (2013) décrit un projet de traduction collaborative de SMS de l'haïtien vers l'anglais mise en place en 2010 suite au tremblement de terre à Haïti. Les ressources produites ont été valorisées mais les volontaires se sont épuisés (à tous les sens du terme) au bout de quelques semaines.

Packham et Suleman (2015) présentent une expérience de production participative de corpus textuels à partir de traduction en isiXhosa, langue d'Afrique du Sud parlée par plus de 8 millions de personnes.

Concernant la myriadisation d'annotations en parties du discours pour une langue peu dotée la plateforme `sloWCrowd` permet la *validation* de ressources annotées (par exemple, le `sloWNet` (Fišer *et al.*, 2014)) mais elle n'est pas conçue pour permettre la résolution de tâches relativement complexes d'annotation.

Notons enfin que, depuis 2020, le Projet Masakhane [67] (∀ *et al.*, 2020; Nekoto *et al.*, 2020) vise à coordonner les recherches concernant la production de ressources et le développement d'outils de traitement automatique pour les langues d'Afrique. Outre son ambition de fédérer les recherches concernées à travers la construction d'une communauté de chercheurs internationale, l'initiative encourage l'investissement de tout locuteur de langue présente sur le continent africain à rentrer en contact direct avec les universités du ou des pays concernés afin de contribuer au projet, notamment en évaluant des systèmes de traduction automatique. Si la majorité des 400 participants semblent provenir de l'univers académique à ce stade, il est probable que l'ouverture à un ensemble plus large de locuteurs se poursuive, l'importance de la connaissance et de la prise en compte de graphies multiples ayant par exemple été mises en avant par les auteurs (Nekoto *et al.*, 2020).

64. Voir : https://crowd4u.org, juin 2020.
65. Voir : https://www.zooniverse.org.
66. Voir : https://lingoboingo.org/game-landing-page/, juin 2020.
67. Voir https://www.masakhane.io/.

3.3 Des locuteurs variés pour collecter des ressources variées

Dans cette section, nous présentons la production participative comme moyen pour les chercheurs de recueillir des données rendant compte de la variation pouvant être observée dans une langue.

3.3.1 Myriadisation de ressources orales

Le premier axe concerne la production participative pour rendre compte la variabilité dialectale d'une aire linguistique.

Traditionnellement, cette collecte est menée par les linguistes de terrain eux-mêmes. Depuis quelques années, cette collecte peut être assistée par la technologie grâce, par exemple à l'utilisation d'applications facilitant la transcription tels qu'Aikuma (Bettinson et Bird, 2017) et Lig-Aikuma (Blachon *et al.*, 2016).

Mais les dernières années ont aussi vu naître nombre d'outils permettant aux locuteurs de contribuer directement à travers des sites Internet et applications mobiles dédiés qui tirent parfois parti de la possibilité de géo-localiser les locuteurs directement *via* leurs appareils mobiles.

C'est le cas par exemple des applications Dialäkt Äpp (Leemann *et al.*, 2015), Voice äpp (Leemann *et al.*, 2016) ou Dialectos del Espanol (Bouzouita *et al.*, 2018), pour la collecte de ressources et la documentation du changement linguistique. Les applications ont été implémentées respectivement pour l'aire dialectale suisse-allemande [68], les dialectes de l'anglais [69] et les espagnols dans le monde. L'application Paldaruo [70] (Cooper *et al.*, 2019) a pour objectif de recueillir la plus grande diversité possible pour les dialectes gallois (pour lesquels une orthographe unifie six zones dialectales). La myriadisation de ressources orales a également été expérimentée pour le développement d'outils de transcription robustes à la variation dans le cas de l'espagnol (Guevara-Rukoz *et al.*, 2020) : 174 locuteurs provenant de 6 pays hispanophones d'Amérique latine ont été impliqués dans la collecte de près de 40 heures de corpus oral transcrit et distribués sous licence CC BY-SA 4.0.

3.3.2 Lexicographie collaborative

L'implication de locuteurs pour la lexicographie collaborative a fait l'objet de nombreux travaux et a permis la production de nombre de ressources, notamment en ce qui concerne de dictionnaires en ligne (Abel et Meyer, 2013).

En effet, il existe de nombreux projets impliquant la construction de ressources lexicales pour les langues régionales et/ou la documentation de variantes locales. Nombre d'entre eux sont basés sur une documentation préexistante de la variation dialectale.

L'intégration de la variabilité au sein de la ressource fait l'objet de différentes stratégies.

Par exemple, dans le cas du *Dictionnaire des mots de base du francoprovençal*, une orthographe supra-dialectale normalisée est utilisée pour les entrées (Stich *et al.*, 2003). Suivant une autre

68. Voir Dialäkt Äpp, http://dialaektaepp.ch/.
69. Voir English Dialect App, http://www.englishdialectapp.com/.
70. Voir : https://apps.apple.com/fr/app/paldaruo/id840185808.

approche, le « *Lexique dialectal suisse italien* » [71] dispose d'une entrée par variante, chacune d'entre elles étant liée à un « terme principal » (« *capolemma* »). Les concepteurs de cette ressource en ligne semblent travailler en étroite collaboration avec des locuteurs locaux (Zoli et Randaccio, 2016), mais la contribution de ceux-ci à l'enrichissement de la ressource n'est pas explicitée.

Duijff *et al.* (2016) rapporte une expérience où des locuteurs contribuent à la construction d'un dictionnaire de dialecte néerlandais-frison, et souligne en particulier la capacité des locuteurs à combler les « lacunes lexicales ».

Citons en outre les multiples éditions de projets Wiktionnaire [72], dont 150 instances sont actives, et qui contiennent des entrées lexicales produites de manière participative.

Ces exemples montrent que des communautés linguistiques de toute taille peuvent être mobilisées pour produire des ressources linguistiques pour leurs langues.

Enfin, citons les travaux d'Avanzi et Stark (2017), qui concernent quant à eux la documentation de la variation du français.

3.3.3 La production participative comme effet de bord

La production participative de corpus peut se concevoir comme un effet de bord de toute production langagière, dès lors que celle-ci est rendue accessible. C'est le cas par exemple du corpus formé par les conversations ayant lieu sur les salons de conversations mis en place par `Ubuntu` sur le réseau de discussion relayée par Internet (*Internet Relay Chat*, ou IRC) `Freenode` (Uthus et Aha, 2013) et disponibles librement [73]. Notons par ailleurs le projet SMS4Science (Dürscheid et Stark, 2011), qui a permis de recueillir jusqu'en 2016 des corpus de SMS dans une quinzaine de pays sur la base de volontariat des participants acceptant de faire « don de leurs SMS à la science ». Nous ignorons délibérément les corpus issus des réseaux sociaux dans le cadre de notre recherche, ceux-ci ne pouvant être redistribués librement (voir section 2.2.4).

La production participative de corpus bruts requiert la mise en place de mécanismes de suggestion pour inspirer les contributeurs, comme par exemple la demande de descriptions. Dans un tel cas, la myriadisation est explicite (Fort, 2016), le but de l'activité étant clairement affiché au participants.

Afin de remédier au caractère rébarbatif et artificiel d'une telle activité et afin de recueillir des contenus divers, Niculae et Danescu-Niculescu-Mizil (2016) et Prys *et al.* (2016) ont proposé deux approches originales permettant la collecte de corpus de manière *implicite*.

Le premier article présente `Street Crowd`, un jeu en ligne dont l'objectif est d'identifier l'endroit du monde où une photo a été prise. Cette recherche vers le bon emplacement se fait en collaboration, plusieurs participants donnant leur avis et éventuellement débattant de la solution. Le corpus recueilli est ici composé des conversations entre les participants.

Le deuxième article présente un correcteur en ligne pour l'orthographe et la grammaire du gallois, utilisé comme tel par les locuteurs. Le corpus recueilli est ici le texte dont le locuteur souhaite

71. *Lessico dialettale della Svizzera Italiana*, voir : http://lsi.ti-edu.ch/lsi/.
72. Voir : https://meta.wikimedia.org/wiki/Wiktionary.
73. Voir : https://irclogs.ubuntu.com/.

vérifier l'orthographe et la grammaire. Cette stratégie apparaît comme particulièrement efficace pour collecter des données diverses en termes de forme et de contenu.

3.4 Conclusions

La myriadisation a été utilisée pour impliquer des membres de communautés linguistiques dans la construction de ressources pour leurs langues. Nous défendons que cette méthode participative est d'autant plus légitime lorsqu'il s'agit de construire des ressources linguistique pour des langues non standardisées.

Si cette pratique est usuelle dans le cadre de langues à tradition principalement orale, elle l'est moins, à notre connaissance, s'agissant de la collecte de données écrites.

Cela n'est pas étonnant étant donné l'étendue du matériau linguistique que constituent les langues orales. Par ailleurs, le statut davantage officiel de la langue écrite laisse sans doute moins d'espace à l'expression de la diversité linguistique, même si, dans la pratique, l'orthographe dans n'importe quelle langue est sujette à des variations.

Étant donné que les langues multi-variantes sont par définition variées, nous pensons que le processus de collecte doit impliquer une diversité de locuteurs. C'est ce qui est observé dans les travaux concernant le contenu généré par les internautes (*User Generated Content (UGC)*), qui s'appuient sur des corpus produits par de nombreux locuteurs pour capturer la diversité des pratiques et étudier l'écart à la norme linguistique dans certains contextes du Web. Nous pensons qu'une approche similaire devrait être envisagée pour les langues à multi-variantes.

L'utilisation de la myriadisation comme moyen de produire des corpus de texte résout le problème de l'identification de la langue, car la langue est contrainte en premier lieu. De plus, le contact direct avec les participants permet la production de méta-données supplémentaires telles que la variante dialectale ou l'habitude d'orthographe utilisée. Bien que dans certains contextes, les locuteurs ne soient pas en mesure de nommer la variante qu'ils utilisent, il peut leur être demandé de pointer la zone géographique sur une carte. De même, lorsque la convention d'orthographe leur est inconnue, nous pouvons demander aux locuteurs d'indiquer leur préférence pour une graphie suggérée par rapport à une autre.

De plus, la myriadisation de ressources pour des langues orales à l'écrit permet de confier la tâche de transcription de leur langue directement aux locuteurs, la transcription de l'oral étant en elle-même une annotation (Falbo, 2005; Eshkol-Taravella, 2015; Gadet, 2017; Niemants, 2018), donc un travail d'interprétation. Ainsi, dans un contexte de collecte de ressources textuelles se voulant au plus près de la pratique réelle des locuteurs, la construction de corpus par transcription d'enregistrements par un tiers nous semble à exclure.

Dans le cas de la construction de corpus écrit, nous sommes par ailleurs davantage intéressée par la manière dont les locuteurs écrivent leur langue que par une transcription précise de la manière dont ils prononcent.

En effet, si les outils développés se veulent performants pour traiter les productions réelles des locuteurs, il est nécessaire que les ressources reflètent effectivement la pratique de ceux-ci. Étant donné que de nombreuses langues ne bénéficient pas de corpus de texte facilement accessibles (au sens large) en dehors de leur utilisation numérique, il est donc nécessaire de réfléchir à la manière appropriée de collecter des corpus durables pour chacune de ces langues.

Alors que certains textes entrés dans le domaine public car suffisamment anciens peuvent représenter une source de corpus intéressante [74], ceux-ci sont peu susceptibles d'être représentatifs des pratiques actuelles. Nous ne considérons par conséquent pas pour notre part la collecte de documents littéraires ou correspondant à des versions anciennes de la langue.

Enfin, si nous voulons pouvoir impliquer les locuteurs dans la participation à d'autres traitements linguistiques tels que l'annotation, la traduction, etc., nous devons leur permettre de contribuer sur des contenus dont la langue et les conventions orthographiques leur sont familières.

74. Voir par exemple le cas d'un corpus développé pour le quechua décrit par Monson *et al.* (2006), composé par exemple de deux textes littéraires publiés au début du 20^{e} siècle.

Conclusion

Les trois premiers chapitres de cette thèse nous ont permis de présenter le contexte au sens large dans lequel s'inscrit notre travail. Nous avons observé dans un premier temps que si la diversité globale s'amenuise, celle qu'on constate dans les TICs et dans le domaine du TAL s'accroissent.

Les TICs peuvent agir comme catalyseur ou comme inhibiteur pour la présence en ligne des langues des internautes. On peut supposer que plusieurs facteurs permettent d'influer dans un sens ou dans l'autre, comme par exemple les politiques linguistiques locales ou le bilinguisme des populations concernées. Concernant le rôle pouvant être joué par le TAL dans ces mécaniques, nous supposons ici que le développement d'outils adaptés à l'utilisation d'une langue donnée dans les TICs facilite son emploi, favorise la production de contenus et augmente l'intérêt pour les locuteurs d'en faire un usage numérique.

Si le TAL s'emploie à répondre aux besoins de technologies pour le traitement d'un nombre croissant de pratiques linguistiques, la marge de progression est importante. En outre, il semble qu'alors que nombre de technologies récentes soient agnostiques de la langue de traitement, leur adaptabilité à de nouvelles pratiques est mise à mal par le manque de ressources nécessaires pour mener cette adaptation.

La construction de telles ressources dans le contexte de pratiques scripturales qui ne sont pas stabilisées ou qui ne font pas consensus parmi les locuteurs présentent une difficulté additionnelle. Cette difficulté se traduit tant d'un point de vue de la représentativité des données que des performances des outils développés lorsque appliqués aux pratiques réelles des locuteurs.

Ces pratiques n'étant pas décrites formellement pour beaucoup, et dépendant notamment du bagage linguistique ou des habitudes graphiques de ses locuteurs, il nous paraît dès lors que le traitement écrit ces langues ne peut s'envisager sans que ceux-ci soient mis à contribution pour documenter collaborativement leurs pratiques et partager leurs connaissances linguistiques.

Le troisième chapitre nous a permis de montrer que la myriadisation peut être invoquée dans des contextes linguistiques variés et pour des tâches diverses. Les travaux proposant la myriadisation d'annotation, ou la production de ressources *pour le TAL*, s'ils ont permis d'obtenir des résultats satisfaisants, concernent à notre connaissance en très grande majorité l'anglais et le français. En effet, les travaux de myriadisation pour d'autres langues concernent principalement des activités de documentation permettant par exemple d'établir des dictionnaires de prononciation ou des lexiques. Ces travaux montrent néanmoins que la mobilisation de locuteurs dans des communautés linguistiques de tailles moindres peut être envisagée.

La myriadisation de ressources *pour le TAL* pour des langues autres que les langues majoritaires n'a en revanche à notre connaissance pas été exploitée. C'est dans l'objectif de tester si la myriadisation pouvait s'appliquer à de tels cas que nous avons développé les plateformes présentées

dans la partie suivante. Les ressources produites grâce à elles ainsi que l'exploitation que nous en avons faite sont présentées dans la partie 3 de cette thèse.

Deuxième partie

Plateformes de myriadisation pour des langues non standardisées

Introduction

À l'instar de Berment (2004), dont la thèse s'intitule *Méthodes pour informatiser les langues et les groupes de langues « peu dotées »*, nous avons choisi un intitulé de thèse évoquant une méthode « générale » : *Myriadisation de ressources linguistiques pour le traitement automatique de langues non standardisées*. Comme lui, nous avons fait des choix contraignant notre étude à un cadre réaliste et permettant l'expérimentation pratique.

Nous consacrons la seconde partie de ce document à une présentation de ces choix, qui concernent dans notre cas les langues des expériences de myriadisation ainsi que les paramètres de celle-ci : l'environnement de collecte (plateformes) et les tâches qui y sont instanciées.

C'est aux trois langues d'intérêt que nous consacrons le premier chapitre (chapitre 4). Nous y décrivons les caractéristiques de l'alsacien, du créole guadeloupéen et du créole mauricien ainsi que les enjeux existants quant à leur traitement automatique.

N'étant locutrice d'aucune de ces langues, nos différents travaux ont été le fruit de collaborations successives. Dans un premier temps, nous avons testé notre méthodologie sur l'alsacien. Ce choix a été favorisé par la disponibilité de Delphine Bernhard et de Lucie Steiblé en 2016 : toutes deux chercheuses au LiLPa de Strasbourg, elles ont construit une partie des ressources nécessaires à l'implémentation de notre méthode (voir section 5.2.1).

Au cours de ce travail, nous avons développé deux plateformes : P_Ann pour l'annotation de corpus textuel en parties du discours, et P_Prod_Var pour la myriadisation de corpus, l'annotation participative de ceux-ci, et la collecte de variantes graphiques. Nous avons mené les développements de façon à ce que ces plateformes puissent être adaptées facilement à une nouvelle langue : les paramètres à modifier pour toute nouvelle langue sont clairement identifiés dans la documentation distribuée avec le code source [75].

Pour valider cette adaptabilité théorique de la méthodologie et des plateformes en contexte réel, nous avons développé des instances pour de nouvelles langues. Cela a été rendu possible par le co-encadrement avec Karën Fort de deux mémoires de Master 1 : celui de Gwladys Feler en 2017, qui nous a permis d'adapter notre première plateforme au créole guadeloupéen, et celui de Harmonie Begue en 2019, au cours duquel la seconde plateforme a été adaptée au créole mauricien.

75. Voir : https://github.com/alicemillour/Bisame/blob/recipes/README.md.

Ainsi, la plateforme P_ANN a été instanciée pour l'alsacien et le créole guadeloupéen :

- `Bisame`[76] pour l'alsacien : *bisame*, ou *bisanme*, *bisàmme*, *bisamme* signifie « ensemble », et est particulièrement utilisé dans l'expression *Salü bisame !* signifiant « Bonjour à tous ! ».
- `Krik !`[77] pour le créole guadeloupéen : *Krik* est un terme intraduisible utilisé dans la tradition créole par les conteurs en guadeloupéens avant leur prise de parole dans l'échange « — *Krik ?* —*Krak !* » avec le public.

La plateforme P_PROD_VAR a également été instanciée pour deux des langues, d'abord l'alsacien, puis le créole mauricien :

- `Recettes de Grammaire`[78] pour l'alsacien.
- `Ayo !`[79] pour le créole mauricien : « *Ayo !* est une interjection difficilement traduisible, qui exprime l'étonnement, la joie, la douleur entre autres. Cette expression proviendrait du tamoul *ayyo*, signifiant « hélas » » (Begue, 2019).

Le second chapitre de cette partie (chapitre 5) dresse les contours des développements que nous avons effectués au cours de cette thèse. Nous y présentons comment les contraintes que nous nous sommes imposées ont conditionné nos choix de conception et de développements à proprement parler pour les trois tâches de myriadisation auxquelles nous nous sommes limitée : la collecte de corpus annoté, de corpus brut et de variantes scripturales.

76. Voir : http://bisame.paris-sorbonne.fr.
77. Voir : http://krik.paris-sorbonne.fr.
78. Voir : http://bisame.paris-sorbonne.fr/recettes.
79. Voir : http://krik.paris-sorbonne.fr/ayo.

Chapitre 4

Communautés linguistiques visées

Sommaire

Dans ce chapitre, nous présentons les trois langues pour lesquelles nous avons implémenté des tâches de myriadisation. En parallèle de ces développements, nous avons mené deux enquêtes en ligne sur l'utilisation de l'alsacien et du créole mauricien sur Internet. Nous présentons ces deux enquêtes dans la première section de ce chapitre (section 4.1) afin de pouvoir en utiliser les résultats pour étayer notre propos dans les sections suivantes.

Les sections suivantes de ce chapitre sont consacrées à la présentation des caractéristiques spécifiques de chaque langue, présentant des profils variés mais se caractérisant toutes trois par une absence d'orthographe consensuelle.

Le tableau 4.1 rappelle les plateformes instanciées et, le cas échéant, les enquêtes menées pour chacune des langues.

Langue (ISO 639-3)	P_ANN	P_PROD_VAR	**Enquête linguistique**
Alsacien (gsw)	Bisame	Recettes de Grammaire	oui
Créole guadeloupéen (gcf)	Krik !	/	non
Créole mauricien (mfe)	/	Ayo !	oui

TABLEAU 4.1 – Plateformes développées pour chacune des langues.

4.1 Enquêtes sur l'utilisation en ligne de l'alsacien et du créole mauricien

Nous avons mené deux enquêtes afin de mieux cerner les pratiques en ligne et les attentes des locuteurs en termes de ressources et d'outils numériques. L'instance créée pour l'alsacien est intitulée « L'alsacien, Internet, et vous », et celle créée pour le créole mauricien « Le créole mauricien et sa présence en ligne ».

Une partie des résultats obtenus pour l'alsacien a été présentée à la 9^{e} *Language & Technology Conference : Human Language Technologies as a Challenge for Computer Science and Linguistics* (LTC 2019) (Millour, 2019a), tandis que les résultats pour le créole mauricien ont été publiés dans (Begue, 2019). Les questionnaires complets sont disponibles en annexe G.1 et G.2.

4.1.1 Motivations et objectifs

Au cours de nos expérimentations, nous avons rencontré trois obstacles majeurs quant à la nature participative du projet. Ces obstacles, que nous avons déduit de nos difficultés ou qui ont été soulevés par les participants, sont les suivants :

1. Rentrer en contact avec la communauté linguistique est une tâche difficile, d'autant plus lorsque celle-ci est répartie sur un territoire où une autre langue est majoritaire.
2. Nous ne sommes pas parvenue à trouver une manière efficace de faire la publicité de nos plateformes : être trop descriptif tend à décourager les participants, mettre l'accent sur l'aspect ludique des plateformes est trompeur et la motivation des locuteurs pour participer à un projet collaboratif profitant à la langue s'est révélé insuffisante dans le cas de l'alsacien.
3. Même quand il existe un ensemble de locuteurs motivés, maintenir leur motivation sur le long terme requiert la mise en place de fonctionnalités adaptées au contexte linguistique.

Pour comprendre quelles actions étaient à entreprendre pour surmonter ces obstacles, et pour avoir un meilleur aperçu des pratiques en ligne, nous nous sommes adressée directement aux locuteurs et diffusé des enquêtes poursuivant les objectifs suivants :

- Connaître le profil des internautes, en mettant l'accent sur leur relation avec la pratique écrite de leur langue (lisent-ils, écrivent-ils, sont-ils à l'aise avec le contenu écrit ?) ;
- comprendre à quel(s) type(s) de contenu(s) écrit(s) les locuteurs aimeraient d'une part avoir accès et qu'ils seraient d'autre part à même de partager dans un contexte de production collaborative de corpus ;

- profiter de l'établissement d'un contact par le biais de l'enquête (plus susceptible d'être propagée que des informations sur une plateforme universitaire de myriadisation) pour
  - sensibiliser les locuteurs sur la nécessité d'inclure leur langue dans le monde numérique,
  - faire de la publicité sur les plateformes de myriadisation existantes pour leur langue,
  - collecter les e-mails de contact des locuteurs qui montrent un intérêt pour les projets que nous développons ;
- s'enquérir de la motivation des locuteurs quant à leur participation potentielle à un projet visant à construire en collaboration des ressources pour leur langue ;
- donner aux locuteurs un espace pour exprimer leur opinion sur l'utilisabilité d'Internet pour leur langue, et les inclure dans le développement de nos projets ;
- recueillir des commentaires supplémentaires sur les plateformes de myriadisation existantes auprès des répondants y ayant d'ores et déjà participé.

4.1.2 Structure des enquêtes

Les deux enquêtes développées s'inspirent largement des études menées par le Digital Language Diversity Project (DLDP) (Soria *et al.*, 2018) [80]. Afin de permettre la comparaison entre nos enquêtes et celles du projets DLDP, nous avons gardé comme telles les questions concernant l'auto-évaluation de la langue et son utilisation numérique.

Les enquêtes sont divisées en quatre parties : (i) Profil des répondants (voir section 4.1.4), (ii) Auto-évaluation des compétences (voir section 4.3.3), (iii) Opinion sur les outils numériques existants [81], (iv) Opinion sur le « *crowdsourcing* » [82](voir section 6.1.3).

4.1.3 Moyens de diffusion

Les enquêtes ont été créées sur `Framaform`, un service français gratuit respectueux de la vie privée [83]. Les réponses ont été collectées uniquement sur le Web, ce qui signifie que l'ensemble de nos répondants sont effectivement des internautes. Pour la diffusion de l'enquête concernant l'alsacien, nous avons transmis l'enquête aux participants de nos projets et invité les organismes officiels, les antennes radio locales et les groupes Facebook alsaciens à partager le lien. Même si nous n'avons pas retracé la provenance des répondants, nous pouvons affirmer que la publication la plus efficace a été réalisée par une boutique de costumes traditionnels à Strasbourg [84]. Le *post* effectué par la page Facebook de la boutique a en effet été partagé plus de 130 fois en quelques jours. C'est intéressant à souligner car les organismes officiels vers lesquels nous nous sommes naturellement tournés dans un premier temps se sont révélés toucher un public beaucoup plus restreint et avoir par conséquent un impact moindre.

80. Toutes les études sont disponibles sur : http://wp.dldp.eu/reports-on-digital-language-diversity-in-europe/.
81. Nous ne présentons pas ici les résultats obtenus à ces questions, dont nous nous sommes rendue compte trop tard qu'elles étaient mal posées et ne permettent par conséquent pas d'analyse pertinente.
82. Bien que nous employions dans ce manuscrit le terme « myriadisation », nous avons utilisé celui de « *crowdsourcing* », en le définissant, dans le cadre des enquêtes linguistiques.
83. Voir : https://framaforms.com, juin 2020.
84. La boutique *Geht's in* : voir https://gehts-in.com/, juin 2020.

« L'alsacien, Internet et moi » a reçu 1 224 réponses, et « Le créole mauricien et sa présence en ligne » en a reçu 143. À titre de comparaison, les enquêtes menées par le Digital Language Diversity Project (DLDP) ont recueilli respectivement 200 réponses pour le breton, 428 réponses pour le basque, 156 réponses pour le karélien, et 596 réponses pour le sarde (Soria *et al.*, 2018).

4.1.4 Profils des répondants

Dans les deux cas, les groupes de répondants sont déséquilibrés en termes de sexe : 55,1 % des répondants sont des femmes dans le cas de l'alsacien, 74,1 % dans le cas du créole mauricien.

Langue	**<20**	**20 à 29**	**30 à 39**	**40 à 49**	**50 à 59**	**60 à 69**	**>70**	**ND**
Alsacien	2,0 %	14,4 %	15,1 %	18,5 %	22,8 %	20,7 %	6,3 %	0,2 %
Créole mauricien	10,5 %	**39,2** %	23,1 %	10,5 %	13,3 %	2,1 %	0,7 %	0,7 %

Tableau 4.2 – Répartition des répondants selon leur âge.

Le tableau 4.2 montre dans le cas du créole mauricien une sur-représentation des individus de la tranche d'âge « 20 à 29 ans ». Nous expliquons ce déséquilibre par le fait que la communication de l'enquête a principalement été menée sur Facebook à partir du compte d'Harmonie Begue que se trouve elle-même dans cette tranche d'âge.

La répartition des répondants dans le cas de l'alsacien est plus étalée. Si elle peut s'expliquer par une population de locuteurs de l'alsacien vieillissante, un âge médian autour de 50 ans montre également que de jeunes locuteurs ont répondu à l'enquête.

Notons que l'alsacien est très peu enseigné en dehors du cadre associatif, alors que le créole mauricien a fait son apparition dans les écoles dans le début des années 2010 (Miller, 2015).

À l'instar des enquêtes du DLDP, nous avons questionné les répondants quant à leur investissement professionnel ou associatif vis-à-vis de la langue concernée. Dans le cas des quatre études menées, entre 48,7 % et 69,9 % des répondants présentaient un tel investissement. Comme cela est signalé par les auteurs, cela introduit un biais important quant à leurs réponses [85]. Dans le cas de l'alsacien, 25 % des répondants déclarent un tel investissement et 18 % dans le cas du mauricien. Ces chiffres restent élevés mais montre que nos enquêtes ont davantage été partagées en dehors des cercles de locuteurs œuvrant activement pour la défense de leur langue.

Enfin, concernant leurs langues maternelles, un tiers des répondants de l'enquête alsacienne déclare que leur première langue est le `français`, un tiers déclare que c'est l'`alsacien` et le dernier tiers déclare que `les deux langues` le sont [86]. 10 % des répondants de l'enquête sur le créole mauricien considèrent que ce n'est pas leur langue maternelle.

85. « Les activistes linguistiques ont tendance à être intentionnellement plus déclaratifs quant à l'utilisation de leur langue et, par conséquent, ils ne peuvent être considérés comme représentatifs des locuteurs. » (« *Language activists tend to be intentionally more assertive in their use of the language and, as a consequence, they can't represent average speakers.* ») (Soria *et al.*, 2018).

86. 1.5 % des répondants ont choisi une autre langue comme première langue, et pour la plupart d'entre eux, l'allemand.

4.2 Trois langues aux profils variés

Comme présenté brièvement dans le tableau 4.3, les trois langues étudiées présentent des profils variés, tant d'un point de vue linguistique que d'un point de vue de leurs statuts.

Langue (ISO 639-3)	**Famille**	**Statut**
Alsacien (`gsw`)	germanique occidentale	langue de France (régionale)
Créole guadeloupéen (`gcf`)	créole à base lexicale française	langue de France (des Outre-mer)
Créole mauricien (`mfe`)	créole à base lexicale française	pas de statut officiel

Tableau 4.3 – Types et statuts des langues de travail.

Un des points communs que présentent ces trois langues est qu'elles évoluent toutes trois dans un contexte de diglossie plus ou moins prononcé avec le français : le français s'est imposé comme langue majoritaire en Alsace depuis les années 1970 (Huck *et al.*, 2007) ; d'après le document produit par l'Inspection générale de l'éducation, du sport et de la recherche (IGÉSR) concernant la Guadeloupe, « la grande majorité de la population peut [...] être qualifiée de bilingue avec des compétences variées dans chaque langue » (IGÉSR, 2018) ; enfin, on rencontre à Maurice un contexte de *polyglossie* où le créole mauricien coexiste principalement avec le bhodjpouri, l'anglais et le français, la pratique de chacune de ces langues étant favorisée par certains contextes d'énonciation spécifiques (Thomson, 2006).

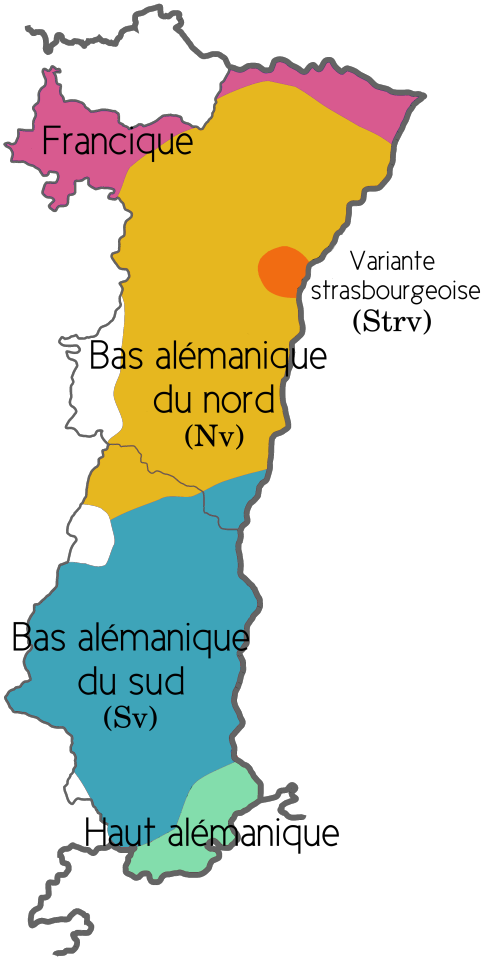


FIGURE 4.1 – Aires dialectales en Alsace[87].

Sachant ce contexte de diglossie, nous avons choisi d'utiliser le français comme langue de nos interfaces. Outre la dimension pratique de ce choix, cela nous permet également de ne pas avoir à préférer une variété dialectale ou scripturale à une autre, évitant ainsi d'exclure une partie des locuteurs de chaque langue.

Par ailleurs, notons qu'alsacien et créole guadeloupéen coexistant avec le français, les locuteurs ont tendance à remplir leurs éventuels trous lexicaux par des termes français. En créole mauricien, c'est l'anglais qui est invoqué principalement, et l'AKADEMI KREOL MORISIEN recommande de conserver la graphie anglaise et de placer le mot entre guillemets dans de tels cas (Hookoomsing, 2004).

87. Cette figure a été adaptée en français à partir de la carte *Aires dialectales en Alsace*, créée par MlibFR et distribuée sous licence CC BY-SA, voir : https://fr.m.wikipedia.org/wiki/Fichier:Carte_Alsace_dialectes.svg.

4.2.1 L'alsacien, langue régionale de France

L'alsacien est un terme générique se référant au continuum de sous-systèmes dialectaux germaniques (Malherbe, 1983) parlés en Alsace et dans une partie de la Moselle. En dépit du déclin de la transmission familiale de l'alsacien, une étude comptabilise 550 000 locuteurs en 2004 (Barre et Vanderschelden, 2004).

La figure 4.1 présente les aires linguistiques principales. Le bas alémanique, la variété principale de l'alsacien, peut lui-même être divisé en deux sous-ensembles : le bas alémanique du nord (Nv) et du sud (Sv). On trouve à Strasbourg une variété légèrement différente du bas alémanique du nord teintée de francique (Strv).

Un des traits linguistiques permettant de distinguer bas alémanique du nord et bas alémanique du sud est la tendance à marquer en fin de mot le son /e/ en bas alémanique du nord (Nv) et le son /a/ en bas alémanique du sud (Sv) (Crévenat-Werner et Zeidler, 2016). La méthode Orthal, présentée en détail dans la section 4.3.2, décrit d'autres caractéristiques telles que le « *relâchement articulatoire (ou de palatalisation) < g-j > du Sud au Nord de l'Alsace* », se traduisant par les graphies suivantes pour le verbe « scier » : `saga` (Mulhouse, Sv), `saja` (Colmar, Sv), `säje` (Strasbourg, Nv).

L'atlas des langues en danger établi par l'UNESCO [88] ne donne pas le degré de vitalité de l'alsacien pris isolément, mais donne celui du groupe des « langues alémaniques », comprenant l'alsacien (`gsw`), le souabe (`swg`) et le haut valaisan (`wae`). Ce groupe, annoncé comme ayant un nombre de locuteurs inconnu mais de l'ordre du million, est classé comme vulnérable par l'UNESCO.

À la question `Utilisez-vous l'alsacien sur Internet (même rarement)?`, que nous avons posée au cours de notre enquête, 47 % des répondants répondent `Oui`, la majorité d'entre eux l'utilisant à la fois pour produire et accéder à des contenus (articles, publications, commentaires). Cela confirme l'étude menée par la DGLFLF qui faisait état en 2014 d'une bonne présence de l'alsacien sur Internet (Pimienta et Prado, 2014). Ce rapport classe l'alsacien parmi les « langues parlées et avec une bonne présence sur la Toile, maintenue de manière équilibrée par tous les secteurs ».

4.2.2 Le créole guadeloupéen, langue de France et des Outre-mer

Le créole guadeloupéen est un créole à base lexicale française parlé principalement dans le département français de la Guadeloupe, île française de l'archipel des Antilles. On dénombre autour de 600 000 locuteurs, environ 400 000 en Guadeloupe et 200 000 sur d'autres territoires (Colot et Ludwig, 2013).

Le créole guadeloupéen se caractérise par une morphologie flexionnelle et dérivationnelle réduite : le pluriel n'est indiqué que par la particule « *sé* » (« *timoun-la* » (« l'enfant »), « *sé timoun-la* » (« les enfants »)), le lexème verbal est principalement invariable, et les temps et les aspects sont marqués par des combinaisons de particules. Il est impossible d'identifier la partie du discours de certains mots indépendamment du contexte : par exemple, « *manjé* » peut à la fois signifier « manger » (à la forme infinitive et conjuguée) et « nourriture ».

88. Voir : http://www.unesco.org/languages-atlas/fr/atlasmap.html.

Notons que le créole guadeloupéen est très proche de l'autre principale variété de créole antillais : le créole martiniquais. Certaines caractéristiques lexicales et morphologiques les distinguent néanmoins, par exemple, les pronoms personnels « *man* » / « *an* » en créole martiniquais, « *moin* » / « *mwen* » en créole guadeloupéen pour le pronom singulier à la première personne « je », ou les pronoms possessifs « *fidji* ***'w*** » en créole martiniquais, « *figi* ***a'w*** » en créole guadeloupéen (« **ton** visage »). De plus, le créole guadeloupéen présente une plus grande variation linguistique en raison de sa géographie moins compacte (Observatoire des pratiques linguistiques, 2005).

FIGURE 4.2 – Géographie de la Guadeloupe[89].

Concernant la présence du créole guadeloupéen en ligne, Pimienta et Prado (2014) font état en 2014 de 90 000 usagers de FACEBOOK ayant déclaré parler le créole guadeloupéen. Le rapport classe le créole guadeloupéen parmi les « langues parlées couramment par la population concernée mais avec une faiblesse marquée sur la toile, maintenues par le secteur académique et recevant un faible soutien institutionnel ».

4.2.3 Le créole mauricien, langue majoritaire sans statut officiel

La république de Maurice est un état insulaire dont les deux îles principales sont Maurice et Rodrigues. Plus de 13 langues sont parlées sur le territoire, les quatre langues majoritaires étant l'anglais, le français, le bhojpuri et le créole mauricien (Thomson, 2006). Si la Constitution de Maurice ne désigne aucune langue officielle, c'est l'anglais qui est la langue utilisée par l'administration et est considérée comme la langue officielle de l'Assemblée.

Le créole mauricien, également créole à base française, présente une flexion verbale qui se traduit par l'existence d'une forme courte et d'une forme longue dont les mécanismes morphologiques et sémantiques respectifs sont étudiés en détail par Bonami et Henri (2010).

Mamode (2013) montre également que le créole mauricien n'est pas dépourvu de morphologie dérivationnelle en mettant en lumière notamment des phénomènes :

89. La figure contient, à gauche, *Location of Guadeloupe in the small antils and the World* créée par TUBS, et distribuée sous licence CC BY-SA, voir : https://commons.wikimedia.org/w/index.php?title=File:Guadeloupe_in_France.svg&lang=fr&uselang=fr et à droite *Carte de lieux touristiques de Guadeloupe, un département français des Antilles*, créé par Sémhur, et distribuée sous licence CC BY-SA, voir : https://fr.m.wikipedia.org/wiki/Fichier:Guadeloupe_Places_of_interest_map-fr.svg.

- d'affixation, par exemple avec (i) le préfixe *de-* dans le sens d'annulation de l'action, donnant les verbes *dekouloute* (« dé-clouer »), *demaye* (« dé-mailler »), *dekuyone* (« dé-couilloner », « déconcerter ») ou (ii) le suffixe *-er* dans par exemple *fezer* (« fais-eur », « vantard »), *dominer* (« domin-eur », « bourreau ») ;
- de réduplication, atténuant l'intensité de l'élément original, par exemple *delo-delo*, littéralement « eau-eau » signifiant « liquéfié », *kol-kole*, littéralement « colle-colle » signifiant « un peu collant ».

D'après le recensement de 2011 publié par la république de Maurice concernant les îles de Maurice et de Rodrigues, le créole mauricien est la langue la plus répandue dans le contexte familial (Republic Of Mauritius Ministry Of Finance And Economic Development Statistics Mauritius, 2011). En effet lorsqu'il est demandé aux habitants d'indiquer la ou les langues habituellement parlées à la maison, le créole mauricien arrive en première position des 52 choix possibles avec 86,5 % des répondants, 0,7 % des répondants ayant répondu parler majoritairement le créole mauricien et une autre langue. Nous reprenons certains des résultats publiés par le Ministère des Finances et du Développement Économique dans le tableau 4.4.

Langue	**Proportion de répondants**
Créole mauricien	86,5 %
Bhojpuri	5,3 %
Français	4,1 %
Hindi	0,7 %
Anglais	0,5 %
Autre	2,9 %

TABLEAU 4.4 – Répartition des répondants selon la « langue habituellement parlée à la maison ».

À la question `Utilisez-vous le créole mauricien sur Internet (même rarement) ?`, que nous avons posée au cours de notre enquête, 68 % des répondants répondent `Oui`, la majorité d'entre eux l'utilisant à la fois pour produire et accéder à des contenus (articles, publications, commentaires).

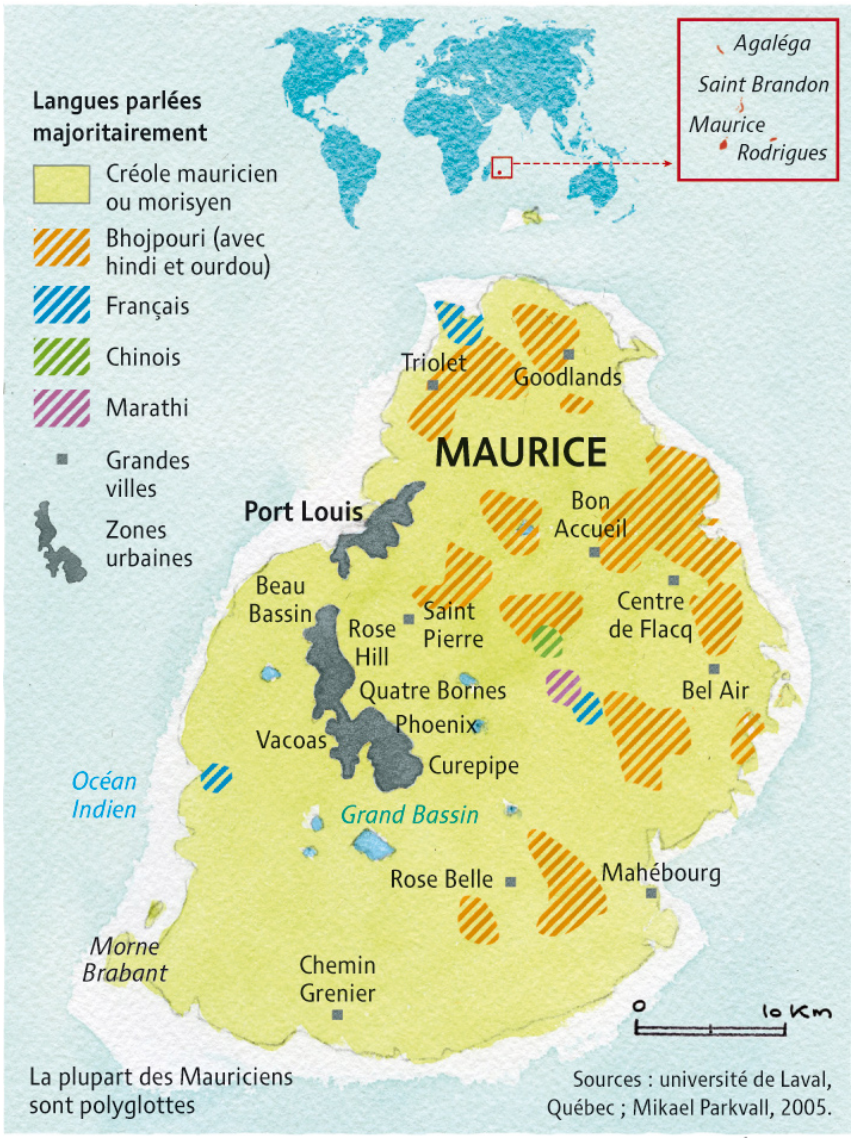


FIGURE 4.3 – Géographie de Maurice et paysage linguistique [90].

90. Cette carte est le fruit du travail d'A. Stienne, publié dans LE MONDE DIPLOMATIQUE, voir : https://www.monde-diplomatique.fr/cartes/maurice-langues. Elle s'appuie sur la carte linguistique élaborée par M. Parvall dont l'originale est disponible ici : http://www.axl.cefan.ulaval.ca/afrique/maurice.htm. Nous avons obtenu l'autorisation du MONDE DIPLOMATIQUE et de M. Parvall pour la publication de cette carte.

4.3 Une absence d'orthographe consensuelle en commun

Après avoir présenté dans la section 4.3.1 comment les notions de variation et d'orthographe s'articulent, nous présentons comment celles-ci se réalisent dans les contextes linguistiques auxquels nous nous sommes intéressée.

4.3.1 Orthographe et variation à l'écrit

On distingue théoriquement les contextes (ou axes) d'expression des variations observées. Les travaux de Coseriu (1973) ont permis de formaliser la variation comme évoluant selon cinq axes : les axes diachronique (c'est-à-dire selon le *temps*), diatopique (selon le *lieu*), diastratique (selon le *groupe social*), diaphasique (selon la *situation d'expression*), et diamésique (selon le mode, *oral* ou *écrit*, d'expression, voir également (Mioni, 1983)). À ces axes de variation peuvent s'ajouter d'autres paramètres tels que le sexe ou l'âge des locuteurs (Moreau, 1997; Androutsopoulos, 1999; Nneka et Okitikpi, 2017).

La variation, quelle qu'en soit sa nature, peut s'exprimer à différents niveaux de l'analyse linguistique : phonétique, phonologique, lexical, morphologique, syntaxique, et sémantique. Si l'impact de chacune de ces variations peut être étudié d'un point de vue du traitement automatique des langues, nous nous intéressons particulièrement ici à la relation entre variations phonétique et phonologique, et l'existence, ou non, d'une orthographe consensuelle.

D'après le CNRTL [91], Le terme *orthographe* se définit d'une part comme suit :

> A. Manière, considérée comme correcte, d'écrire un mot.
> B. Ensemble des règles fixées par l'usage, la tradition, qui régissent l'organisation des graphèmes, la manière d'écrire les mots d'une langue ; connaissance et application de ces règles.

La définition *par extension* est donnée dans un second temps :

> A. Manière, quelle qu'elle soit, d'écrire un mot.
> B. Système de représentation des sons par des graphies, qui est propre à une époque, à un pays, à un auteur, etc.

Bien que son étymologie soit sans équivoque, le préfixe *ortho-* portant le sens de conformité à un ensemble de règle, le terme *orthographe* est donc ambigu : si la première définition contient l'idée d'une norme, la seconde s'en éloigne.

Dans la suite de ce manuscrit et afin d'éviter toute confusion, nous emploierons le terme *graphie* au sens B. proposé par le CNRTL [92] :

> A. (Mode de) représentation du phonème dans le code graphique.
> B. Façon d'écrire un mot.

91. Voir : https://www.cnrtl.fr/definition/orthographe, juin 2020.
92. Voir : https://www.cnrtl.fr/definition/graphie, juin 2020.

Nous utilisons *convention orthographique* ou *orthographe* pour désigner un ensemble de règles destiné à standardiser la pratique écrite d'une langue en prescrivant une (ou plusieurs) graphie(s) correcte(s) en opposition à d'autres graphies considérées comme fautives. Notons que pour qu'une convention orthographique devienne une *norme*, c'est-à-dire un « *état habituel* [de la langue], *régulier, conforme à la majorité des cas* » [93], il faut non seulement qu'une orthographe existe, mais aussi qu'elle soit connue, acceptée et employée par les locuteurs.

Si la question de l'orthographe nous intéresse ici, c'est parce que l'existence d'une convention orthographique permet habituellement de dissimuler à l'écrit les variations phonétique et phonologique derrière une forme graphique *unique*. *A contrario*, l'absence de convention consensuelle laisse la possibilité à de multiples graphies de coexister. Par exemple, et bien que l'Académie française qualifie la prononciation du *s* final dans certaines régions de France d'« exceptions par rapport à la norme » [94], il existe deux prononciations en français du mot *moins* selon un phénomène identifié de variation diatopique (Avanzi *et al.*, 2016; Thibault, 2017). Si la prononciation du *s* final n'est pas rare, c'est son omission à l'écrit qui l'est. La figure 4.4 permet de constater l'effet de l'acceptation de `moins` comme graphie unique par les personnes écrivant le français.

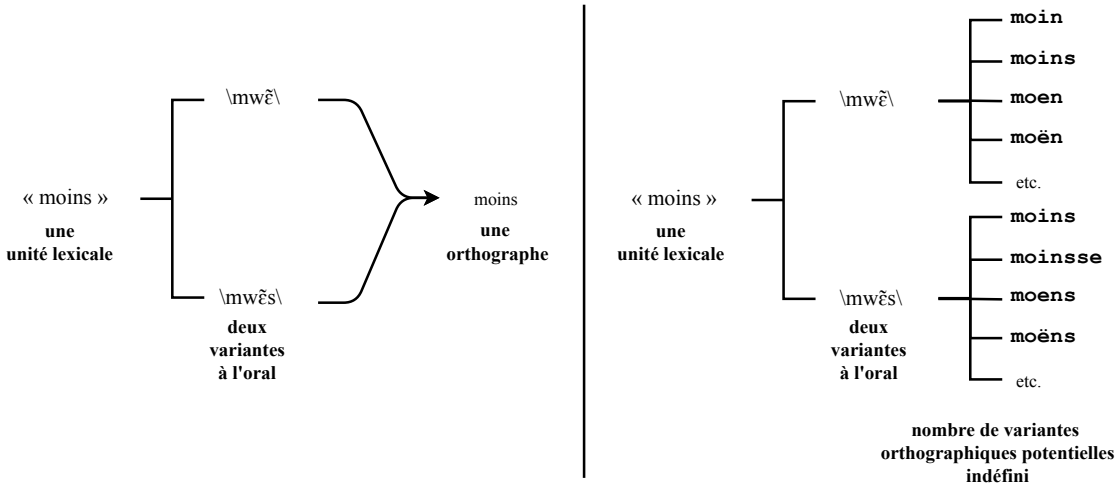


FIGURE 4.4 – Écriture(s) possible(s) de l'entité sémantique *moins* selon qu'il existe (à gauche) ou non (à droite) une convention orthographique consensuelle.

Outre le cas bien identifié de la variation diatopique de l'alsacien (voir section 4.2.1), nous n'avons pas cherché dans le cadre de notre travail à analyser les causes des phénomènes de variation observés pour les trois langues d'intérêt. Il est certain que ceux-ci procèdent de causes variées et propres à chaque communauté linguistique.

Nous nous sommes en effet davantage intéressée à la variation graphique en tant que phénomène observable à ne pas négliger lors de la construction d'une chaîne de traitements linguistiques.

4.3.2 Initiatives de standardisation orthographiques

Les trois langues auxquelles nous nous sommes intéressée ont fait l'objet de plusieurs tentatives de standardisation de l'écrit. Les propositions de conventions orthographiques qui ont été

93. Voir : https://www.cnrtl.fr/definition/norme, juin 2020.
94. Voir : http://www.academie-francaise.fr/sebastien-d-france, juin 2020.

proposées pour les trois langues sont les suivantes :

- pour l'alsacien : les premières conventions orthographiques sont publiées en 2003 par le Groupe d'Études et de Recherches Interdisciplinaires sur le Plurilinguisme en Alsace et en Europe (GERIPA), dirigé par A. Hudlett, sous le nom de *Charte de la graphie harmonisée des parlers alsaciens*. Cette charte donne naissance à la graphie de l'Association pour la Culture et le Patrimoine d'Alsace (ACPA), créée par André Nisslé [95]. Enfin, la méthode Orthal (Crévenat-Werner et Zeidler, 2008), mise à jour en 2016 (Crévenat-Werner et Zeidler, 2016) formalise un ensemble de règles de *transcription* permettant d'instaurer une « orthographe flexible ». Les particularités régionales sont maintenues à l'écrit mais les règles de transcription sont unifiées. L'engagement poursuivi par l'orthographe est le suivant :

  > « Il ne s'agit en aucun cas d'uniformiser les différents parlers pour créer une forme standard artificielle, mais de faciliter l'écriture et la lecture de textes en dialecte avec un système orthographique cohérent qui utilise, du nord au sud de l'Alsace, les mêmes lettres pour les mêmes sons. » (Crévenat-Werner et Zeidler, 2016)

  Le guide contient ainsi des recommandations de transcription du type :

  > « Par exemple, la tendance pour les parlers de Haute Alsace est de prononcer /a/ pour l'article indéfini (« un », « *ein* ») et de marteler davantage la syllabe finale où l'on entend /a/ [a] même si la syllabe est atone [ɐ]. C'est pourquoi Orthal propose d'écrire : `a Wort` (« *ein Wort* », « un mot ») ; `màcha` (« *machen* », « faire ») ; `Maidla` (« *Mädchen* », « jeunes filles »). NB : ceci ne signifie pas que tous les locuteurs de Haute Alsace doivent s'y conformer. S'ils entendent ou réalisent /e/ réduit, ils écriront « e ». » (Crévenat-Werner et Zeidler, 2008)

  Cette flexibilité graphique conduit ainsi à trouver dans un lexique compilant différentes sources (Bernhard et Steiblé, 2015) les quatre graphies suivantes pour le mot « coude » : `Elleboje` (ACPA), `Ällabooga` (ACPA), `Elleböje` (Orthal) et `Ellaboja` (Orthal).
- pour le créole guadeloupéen : l'orthographe du GEREC-F (Groupe d'Études et de Recherches en Espace Créolophone et Francophone) (Ludwig *et al.*, 1990) modifiée plus tard par Bernabé (2001), coexiste avec le système introduit par Hazaël-Massieux (2000). En se basant sur les conventions utilisées dans les écrits informels et l'affichage public, Jeannot-Fourcaud (2017) observe que le système utilisé majoritairement aujourd'hui semblerait être celui correspondant au *Dictionnaire Créole-Français* (Ludwig *et al.*, 2002).
- pour le créole mauricien (Olivas Alguacil, 2019) : la Grafi-larmoni (Hookoomsing, 2004), vise, comme son nom l'indique, à *harmoniser* l'écriture du créole mauricien, faisant jusqu'alors l'objet de débats ne rendant pas service au développement de la pratique écrite. Elle a été proposée suite à une commande du ministère de l'éducation et de la recherche, qui a plus tard donné naissance à l'Akademi Kreol Morisien qui publie en 2011 *Lortograf Kreol Morisien* et *Gramer Kreol Morisien* (Carpooran, 2011).

Nous nous trouvons donc dans trois situation linguistiques où des orthographes ont été introduites récemment, et dont l'intégration aux pratiques scripturales des locuteurs est très inégale.

95. Le lexique bilingue décrivant ces conventions était hébergé jusqu'à récemment sur la page Web d'une association locale qui n'est malheureusement plus en ligne aujourd'hui.

4.3.3 Manifestations de la variation à l'écrit

Dans cette section, nous présentons dans un premier temps les situations rencontrées dans le cas de l'alsacien et du créole mauricien, langues pour lesquelles nous pouvons nous appuyer sur les enquêtes présentées dans la section 4.1. Les observations proposées dans le cas du créole guadeloupéen sont issues de travaux existants ou de notre propre étude des corpus utilisés (voir section 5.2.1).

4.3.3.1 Cas de l'alsacien et du créole mauricien

Perception de la variation

D'abord, nous nous sommes interrogée sur la perception de la variation par les locuteurs de l'alsacien et du créole mauricien. Les résultats que nous obtenons montrent, qu'en majorité, ceux-ci ont conscience de phénomènes de variation dans leurs langues respectives.

En effet, dans le cas de l'alsacien, nous leur avons demandé de renseigner la `variété d'alsacien` qui était la leur. 41,5 % des répondants ont choisi `le bas alémanique du nord` (NV), 20 % `le bas alémanique du sud` (SV), 19,2 % `l'alsacien de Strasbourg` (STRV), 2,5 % le `Francique lorrain`, ou *Plàtt*, 1,7 % le `Palatin allemand` et 1,3 % on proposé d'autres variantes[96]. Ainsi, seuls 5 % des répondants ont signalé `ne pas savoir la variante qu'ils parlent`. Notons que 12,5 % des répondants ont choisi au moins deux réponses.

Dans le cas du créole mauricien, pour lequel la variation n'est pas aussi formellement décrite et identifiable qu'en alsacien, nous avons posé la question suivante : « `Pensez-vous que le créole mauricien possède des variantes ?` », à laquelle 76 % des répondants ont répondu « `Oui, le créole parlé varie selon les régions de l'île` ».

Auto-évaluation

Nous avons ensuite proposé aux répondants de s'auto-évaluer en leur demandant s'ils estiment avoir une compétence `bonne`, `moyenne` ou `faible` en (i) `compréhension orale`, (ii) `expression orale`, (iii) `compréhension écrite` et (iv) `expression écrite`. Nous donnons dans le tableau 4.5 la proportion des répondants ayant auto-évalué leur compétence à `bonne` ou `moyenne`. Nous présentons les résultats par tranche d'âge dans le cas de l'alsacien, et globalement pour le créole mauricien compte tenu du nombre moindre de réponses.

Dans le cas de l'alsacien, les chiffres illustrent la baisse de la transmission de la langue (près de 90 % des répondants de plus de 50 ans estiment à `moyenne` ou `bonne` leur expression orale, contre 71 % des répondants ayant entre 30 et 50 ans, et 57 % des moins de 30 ans). Néanmoins, ces résultats montrent tout de même qu'une nouvelle génération de locuteurs est présente sur Internet (22 % des moins de 30 ans évaluant à `bonne` leur expression orale).

On note dans les deux cas une chute de l'auto-évaluation entre la compétence orale et écrite, et cette tendance est d'autant plus marquée dans le cas de l'alsacien, ce qui est intéressant sachant qu'il n'existe pas de directive officielle pour sanctionner une graphie fautive. En effet, si les répondants dans le cas de l'alsacien estiment globalement leur expression orale à bonne ou

96. Les autres variantes renseignées par les répondants sont : `l'alsacien de Sundgau`, `le haut alémanique`, `l'alsacien de Mulhouse`, `le suisse alémanique`, et des variantes spécifiques à des villages d'Alsace.

		Alsacien				Créole
		<30	30 à 50	50 à 70	>70	mauricien
Compréhension orale	bonne	58 %	76 %	89 %	95 %	92 %
	moyenne	31 %	16 %	19 %	1 %	6 %
Expression orale	bonne	22 %	42 %	74 %	94 %	80 %
	moyenne	35 %	29 %	21 %	3 %	15 %
Compréhension écrite	bonne	29 %	34 %	55 %	71 %	50 %
	moyenne	31 %	44 %	32 %	22 %	36 %
Expression écrite	bonne	7 %	13 %	20 %	32 %	35 %
	moyenne	16 %	20 %	39 %	26 %	37 %

TABLEAU 4.5 – Auto-évaluation des répondants, par tranche d'âge pour l'alsacien, globale pour le créole mauricien.

moyenne à 95 %, seuls 46 % d'entre eux évaluent leur expression écrite comme telle. Dans le cas du créole mauricien l'évaluation passe de 98 % à l'oral à 72 % à l'écrit.

Rapport à l'écriture

Nous avons également demandé aux répondants s'ils écrivaient la langue concernée et, en cas de réponse négative, pourquoi. Le tableau 4.6 donne les proportions de réponses obtenues dans les deux cas. On remarque ici et dans les deux cas que le taux de répondants Oui à la question « Écrivez-vous [l'alsacien|le créole mauricien] ? » est plus élevé que le taux de répondants s'auto-évaluant positivement en expression écrite. Les locuteurs des deux langues écrivent donc davantage qu'ils ne se sentent performants à l'écrit.

La part de locuteurs ayant répondu Non, je ne saurais pas comment l'écrire (13 % dans le cas de l'alsacien, 10 % dans le cas du créole mauricien) montre en revanche qu'une part des locuteurs présente une insécurité bloquante à l'écrit, et qu'une marge de progression existe par l'enseignement d'une orthographe pouvant y remédier dans ces cas.

Oui	Alsacien	69 %
	Créole mauricien	71 %
Non, je ne saurais pas comment l'écrire	Alsacien	13 %
	Créole mauricien	10 %
Non, je n'en ai pas l'occasion	Alsacien	12 %
	Créole mauricien	9 %
Non, c'est une langue orale, je ne souhaite pas l'écrire	Alsacien	2 %
	Créole mauricien	7 %
Non, pour une autre raison	Alsacien	3 %
	Créole mauricien	2 %

TABLEAU 4.6 – Relation à l'écriture en alsacien et en créole mauricien – réponses obtenues à la question : « Écrivez-vous [l'alsacien|le créole mauricien] ? ».

Utilisation de l'orthographe

Nous nous sommes enfin intéressée à la relation entretenue par les locuteurs avec les orthographes principalement en usage. Le tableau 4.7 présente les réponses obtenues aux questions « Lorsque vous écrivez, utilisez vous [l'orthographe Orthal | la graphie prescrite par l'Akademi Kreol Morisien dans *Lortograf Kreol Morisien*] ? ».

Oui, toujours	Alsacien	8 %
	Créole mauricien	15 %
Oui, parfois	Alsacien	11 %
	Créole mauricien	27 %
Non, j'aimerais, mais je ne la maîtrise pas	Alsacien	7 %
	Créole mauricien	13 %
Non, je refuse de l'utiliser	Alsacien	5 %
	Créole mauricien	7 %
Non, je ne connaissais pas cette graphie	Alsacien	69 %
	Créole mauricien	38 %

TABLEAU 4.7 – Relation à l'orthographe en alsacien et en créole mauricien – réponses obtenues à la question : « Lorsque vous écrivez, utilisez vous [l'orthographe Orthal |la graphie prescrite par l'Akademi Kreol Morisien dans *Lortograf Kreol Morisien*] ? ».

Dans le cas du créole mauricien, nous avons également proposé aux répondants de choisir la graphie qui leur paraissait la plus familière pour la phrase « Je suis allé acheter du pain ». Nous donnons dans le tableau les options choisies par les répondants. La graphie n°1 correspond aux recommandations officielles, la graphie n°2 est une graphie plus proche de la graphie française, les trois dernières versions ont été proposées par des locuteurs ne se reconnaissant dans aucune des deux premières.

Au sujet de la variation de graphies proposées par les répondants, Begue (2019) affirme :

> « [...] nous pouvons définir assez assurément que cette variation n'est pas régionale (les deux premières alternatives ont été proposées par des répondants habitant le centre de l'île [97], et la dernière de l'ouest), elle est purement une variation orthographique, qui représente bien la situation du [créole mauricien] aujourd'hui – où plusieurs graphies co-existent. »

Enfin, nous avons demandé aux répondants de tenter de nous décrire les règles qu'ils utilisent lorsqu'il ne sont pas sûrs de l'orthographe d'un mot. 50 % d'entre eux l'écrivent comme ils l'entendent, 30 % l'écrivent avec la graphie française, et 20 % vérifient l'orthographe dans le dictionnaire du créole mauricien ou l'écrivent tantôt comme ils l'entendent, tantôt avec la graphie française.

Les commentaires suivants ont été donné par les participants pour expliquer le mécanisme permettant d'arrêter leurs choix sur une ou l'autre des orthographes :

97. 56 % des répondants de l'enquête proviennent du centre de l'île, où se situent les principales agglomérations.

	Graphie	% Répondants
1	monn al aste dipin	61 %
2	monne alle acheté dipain	36 %
3	mo'nn al aste dipin	-
4	mone al aster dipain	-
5	mone ale acheté du pain	-

TABLEAU 4.8 – Différentes graphies en créole mauricien pour la phrase correspondant à « J'ai été acheter du pain ».

- « Si le même mot écrit en français transmet la même intention, il sera utilisé. Dans l'autre cas, ce sera comme je l'entends. »
- « Si le mot est couramment utilisé en créole, je ferai l'effort d'écrire. Si c'est un mot plutôt français, non. (ex : cancer du sein). »
- « La sonorité et comment je peux intégrer (le mot français) en créole. »
- « Je vérifie dans le dictionnaire ou selon mes connaissances en morphosyntaxe. »

4.3.3.2 Cas du créole guadeloupéen

Dans le cas du créole guadeloupéen, aucun accord n'a pu être trouvé quant au positionnement envers l'orthographe française lorsqu'elle peut être invoquée. Par exemple, les graphies `chien` et `chyen` peuvent être trouvées en créole guadeloupéen.

D'autre part, la segmentation fait l'objet de variations, et nous avons rencontré dans notre corpus pourtant relativement petit plusieurs cas de graphies concurrentes impliquant la ponctuation et le découpage des unités lexicales. Par exemple, « *latè* », issue de l'agglutination du déterminant français « la » et du nom propre « Terre », est généralement perçue comme une entité unique, signifiant « Terre » dans son ensemble, et est par conséquent écrite comme telle. Pourtant, nous avons trouvé des occurrences de la forme séparée `la tè`, qui est considérée comme erronée par les créolistes. On constate également une utilisation inconstante de l'espace et du trait d'union entre les noms et les déterminants postposés, avec par exemple `tifi-la` et `tifi la` (« la fille »), ou entre adjectifs et noms, avec par exemple `jenn fi`, `jenn-fi`, et `jennfi` (« jeune fille ») (Delumeau, 2006).

Un autre exemple présent dans le corpus illustre la mauvaise pénétration des normes parmi les locuteurs : il s'agit de la coexistence de deux graphies « `a pa` » et « `apa` ». Alors que Ludwig *et al.* (1990) ont introduit une convention graphique pour distinguer « *a pa* » (existentiel négatif) trouvé dans un contexte tel que « *A pa pas ou ni lajan* [...] » [98] (« Ce n'est pas parce que tu as de l'argent [...] »), de « *apa* » (« à part »), deux des trois locuteurs du créole guadeloupéen avec lesquels nous avons travaillé n'avaient jamais rencontré la forme séparée.

Une étude menée sur un corpus de SMS en créole guadeloupéen permet de constater cet écart à ce que l'auteure appelle *l'orthographe usuelle du créole*, c'est-à-dire correspondant au *Dictionnaire Créole-Français* (Ludwig *et al.*, 2002) :

> « [...] on pourrait s'attendre à ce que les messages rédigés en créole soient systématiquement graphiés en *orthographe usuelle du créole* du fait de son potentiel

98. Cet exemple est tiré de (Delumeau, 2006).

> économique. Or, l'étude du corpus montre plutôt une tendance à la divergence, qui se manifeste généralement par une exploitation de normes orthographiques du français. » (Jeannot-Fourcaud, 2017)

L'étude donne notamment des exemples de graphies francisées suivant une écriture qualifiée par l'auteure d'« étymologisante » : `apres` au lieu de `aprè` (« après »), `movai` au lieu de `mové` (« mauvais »), `caz` au lieu de `kaz` (« maison »), etc. Notons que, par rapport à l'exemple de « moins » présenté dans la figure 4.4, Jeannot-Fourcaud (2017) note que si seule l'« unité orale » \ mwɛ̃ \ existe en créole guadeloupéen pour l'unité lexicale française correspondante « moins », trois graphies sont en revanche attestées dans le corpus de SMS : `mwen` (graphie majoritaire correspondant aux conventions), `mwin`, et `moin`.

4.4 Travaux de recherche existants

Nous présentons ici les travaux de recherche et outils pour l'annotation en morphosyntaxe des trois langues d'intérêt. Lorsque le cas se présente, nous donnons les tailles des corpus en « *tokens* » tels que fournis par les auteurs des travaux présentés. Nous revenons sur cette dénomination dans la section 5.2.2.1.

4.4.1 Annotation en morphosyntaxe de l'alsacien

Dans le cas de l'alsacien, les expériences d'étiquetage en parties du discours existant en 2016 sont très exploratoires, la seule expérience dans ce sens étant celle de Bernhard et Ligozat (2013). La méthode proposée consiste à remplacer les « mots-outils » alsaciens par leurs équivalents allemands, à l'aide d'un lexique bilingue de petite taille, pour ensuite appliquer des `taggers` existant pour l'allemand (en l'occurrence, `TreeTagger` (Schmid, 1997) et `Stanford POS Tagger` (Toutanova *et al.*, 2003)). Le lexique bilingue de « mots-outils » (déterminants, pronoms, prépositions, conjonctions, particules, adverbes et verbes fréquents) $L_{\text{MO\_gsw}}$ a été construit manuellement par les auteures à partir d'un corpus d'évaluation de 2 292 *tokens*. Notons que les formes ambiguës d'un point de vue morphosyntaxique ont été éliminées. La moitié du corpus utilisé, qui n'est pas distribué, est constituée d'exemples issus d'un dictionnaire multilingue français-allemand-anglais-alsacien (Adolf, 2006).

Les mesures d'exactitude obtenues grâce à cette méthode vont de 79 à 89 % en fonction du corpus d'évaluation et du `tagger` utilisé. Les tailles des corpus d'évaluation variant entre 166 et 1 180 *tokens*, il est difficile de dire à quoi est liée cette grande amplitude de performances. On peut supposer que la méthode proposée est peu robuste à la variation existant entre les différents textes (voir section 5.2.3).

Il existe un deuxième lexique, L_{gsw}, contenant environ 40 000 entrées issues de divers lexiques bilingues : lexiques thématiques du site Web de l'OLCA (Office pour la Langue et la Culture d'Alsace) [99], dictionnaire bilingue du site de l'Association Culture et Patrimoine d'Alsace (ACPA) [100] et du dictionnaire multilingue français-allemand-anglais-alsacien (Adolf, 2006). L'utilisation de différentes sources permet de couvrir des graphies variées ; le verbe « jouer » y est par exemple présent sous les sept formes `spiel`, `spìel`, `spìela`, `spìele`, `spiele`, `speele`, `schpeela`.

99. Voir : http://www.olcalsace.org/, juin 2020.
100. Le site n'est malheureusement plus en ligne.

4.4.2 Annotation en morphosyntaxe des créoles guadeloupéen et mauricien

Dans le cas des créoles guadeloupéen et mauricien, aucun corpus annoté ni outil d'annotation n'avait été développé auparavant.

À notre connaissance, les seuls travaux ayant envisagé le traitement automatique du créole guadeloupéen sont ceux de Delumeau (2006), qui introduit une description linguistique du créole guadeloupéen dans une perspective de génération, de Carrión Gonzalez et Cartier (2012) qui détaillent les ressources lexicales existantes pour divers créoles français, de Schang (2013), qui présente une méta-grammaire pour le créole guadeloupéen et de Schang *et al.* (2017), introduisant un corpus annoté en co-référence [101].

Le créole mauricien ne faisant pas partie des langues proposées par Google Translate pour la traduction, Dabre *et al.* (2014) explorent diverses stratégies de développement d'un outil de traduction automatique sur la base d'un corpus aligné anglais–français–mauricien issu d'un dictionnaire. Les corpus utilisés ne sont pas distribués.

4.5 Conclusion

Les trois communautés linguistiques auxquelles nous nous sommes intéressée sont à la fois présentes en ligne et peu représentées dans les technologies du langage. En outre, il s'agit de communautés dont la pratique écrite n'est pas stabilisée par une orthographe consensuelle parmi ses locuteurs. Pour les trois langues, plusieurs graphies peuvent donc coexister pour un élément de lexique, les conventions existantes pour chacune des langues ne faisant pas office de norme consensuelle au sein des locuteurs. Cette variation graphique procède d'habitudes d'écriture variées, pouvant s'entremêler à une variation dialectale notamment identifiée dans le cas de l'alsacien.

En cela, l'alsacien, le créole guadeloupéen et le créole mauricien constituent à la fois un terrain propice à l'expérimentation de la myriadisation et un enjeu intéressant pour le traitement automatique des langues.

Il n'existait pas au moment où nous avons démarré nos expérimentations, de corpus annoté en parties du discours permettant d'entraîner un modèle d'annotation pour aucune des trois langues. C'est dans l'optique de remédier à ce manque que nous avons conçu les tâches de myriadisation présentées dans le chapitre suivant.

101. Le corpus utilisé par les auteurs est le même que celui que nous présentons dans la section 5.2.1.2, c'est-à-dire un corpus d'oral transcrit.

Chapitre 5

Tâches de myriadisation implémentées

Sommaire

Nous consacrons le deuxième volet de cette partie consacrée aux plateformes développées aux aspects techniques de celles-ci et à nos divers choix d'implémentation. Nous y présentons une chronologie argumentée des développements réalisés.

Cette chronologie est illustrée dans la figure 5.1 : comme évoqué dans l'introduction de cette partie, la première plateforme P_Ann est d'abord instanciée pour l'alsacien (plateforme `Bisame`), puis adaptée au créole guadeloupéen (plateforme `Krik !`). La seconde plateforme P_Prod_Var, est elle aussi d'abord instanciée pour l'alsacien (plateforme `Recettes de Grammaire`), puis pour le créole mauricien (plateforme `Ayo !`).

Si l'évolution des fonctionnalités a été influencée par les besoins rencontrés pour les langues pour lesquelles nous les avons successivement instanciées, elles sont conçues pour être adaptables

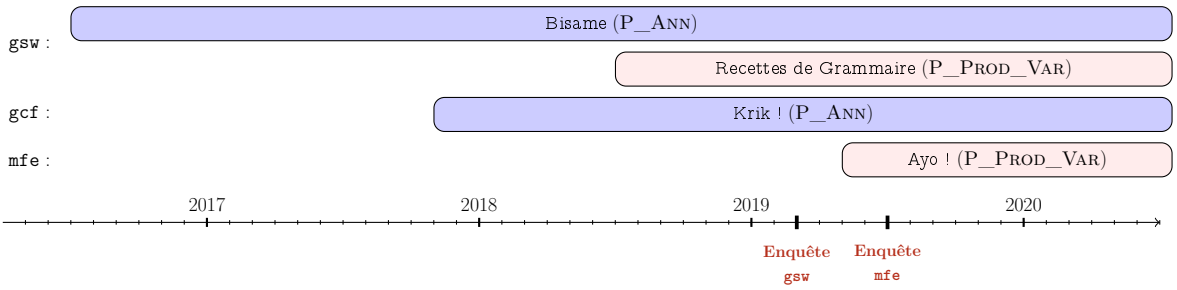


Figure 5.1 – Chronologie de la mise en ligne des plateformes instanciées et des enquêtes réalisées pour l'alsacien (`gsw`), le créole guadeloupéen (`gcf`), et le créole mauricien (`mfe`).

à toute langue candidate à la myriadisation, c'est-à-dire présentant un nombre suffisant de locuteurs en ligne. Les ressources minimales requises ainsi que les stratégies mises en place pour les construire pour chacune des trois langues d'intérêt sont détaillées dans la section 5.2. Nous décrivons également dans cette section les choix de traitements que nous avons dû opérer en amont des expériences de myriadisation.

Nous présentons ensuite les trois tâches de myriadisation que nous avons implémentées et diffusées *via* les plateformes décrites dans la section 5.1. Dans un premier temps, dans la section 5.3, nous présentons comment nous avons conçu la myriadisation d'annotations en parties du discours, en premier lieu sur des corpus existants, puis sur des corpus saisis directement par les participants. Nous présentons dans la section 5.4 la collecte de corpus textuels qui nous a permis de mettre en place l'annotation « à la volée » par les participants de leurs propres textes. Enfin, la section 5.5 est consacrée à la présentation d'une troisième tâche visant à collecter des variantes scripturales sur du lexique présent dans les textes proposés.

Les choix de pré-traitements et d'implémentation présentés ont été guidés par les contraintes C1 (« Pérennité des ressources produites »), C2 (« Réplicabilité de la méthodologie ») et C3 (« Myriadisation bénévole ») présentées dans l'introduction.

5.1 Conditions de collecte

5.1.1 Le choix du contrôle sur les fonctionnalités

Conformément à la contrainte « C2 : Réplicabilité de la méthodologie » présentée dans l'introduction, nous avons choisi de réaliser nos expériences de myriadisation sans recourir à des solutions logicielles pré-existantes.

Nous n'avons pas souhaité utiliser d'incitation financière à la participation telles que présentées dans la section 3.2.1, notamment pour assurer la reproductibilité de nos travaux à moindre frais. L'absence d'incitation financière nous contraignant à susciter la motivation par un autre biais, il était primordial de pouvoir avoir la main sur les fonctionnalités présentes sur la plateforme. Nous avons ainsi pu développer des mécanismes d'incitation basés sur d'autres leviers que le gain financier, comme le divertissement, la valorisation de la connaissance, ou le sentiment d'appartenance à une communauté spécifique (Poesio *et al.*, 2013).

Le contrôle que nous avons eu sur les fonctionnalités nous a permis d'adapter librement les

tâches de myriadisation au fur et à mesure de nos expériences.

Par ailleurs, le développement de plateformes spécifiques nous permet d'assurer la distribution du code source sur `GitHub` [102] sous licence CeCILL v2.1 [103]. Les codes source des deux plateformes P__Ann et P__Prod__Var correspondent à deux branches distinctes sur le dépôt Git. Pour chacune des plateformes, le passage d'une instance de langue à une autre se fait par modification d'une variable globale. Le code source est distribué avec un guide des étapes à suivre pour fournir les éléments nécessaires à l'instanciation d'une nouvelle langue.

Nous présentons dans la section 5.1.2 les spécifications techniques découlant de ce choix, puis, dans la section 5.1.3 les précautions éthiques qu'il nous a permis de mettre en place sur nos différentes plateformes.

5.1.2 Spécifications techniques

Les spécifications techniques des plateformes ont été choisies compte tenu de nos propres compétences en développement web et gestion de bases de données. Notamment nous avons choisi de développer le site en lui-même en PHP (version 5.5.9 puis 7.7.1) et les bases de données en MySQL (version 5.7.20).

Afin de pouvoir bénéficier des nombreux modules mis à disposition (authentifications, gestion des sessions, des migrations de bases de données, de routage etc.), nous avons utilisé le *framework* Laravel (version 5.0 puis 5.6) qui fonctionne sur le patron Modèle-Vue-Contrôleur (MVC). Les sites ont été déployés par nos soins sur deux serveurs hébergés par la Sorbonne : `bisame.paris-sorbonne.fr` et `krik.paris-sorbonne.fr`, sur lesquels a été installée et configurée la version 7 de CentOS.

Nous avons bénéficié au cours de notre thèse d'un financement de la délégation générale à la langue française et aux langues de France [104] (DGLFLF), dans le cadre de l'appel à projet *Langues et numérique 2018*. Le projet correspondant au financement obtenu a été conçu avec Karën Fort, Delphine Bernhard, André Thibault, Bruno Guillaume, et Nicolas Lefèbvre est le projet Production ludique de ressources annotées pour les langues de France (PLURAL).

Le financement obtenu nous a permis de recruter Nicolas Lefèbvre, alors ingénieur Inria. Il a apporté son soutien technique durant quelques mois lors du développement de la plateforme Recettes de Grammaire notamment concernant la fonctionnalité d'ajout d'une variante graphique pour un mot donné de la plateforme.

Les financements obtenus ne nous ont en revanche pas permis de développer la version mobile de ces plateformes, par conséquent accessibles à ce jour uniquement *via* un navigateur. Néanmoins, un travail a été fait pour que la plateforme P__Prod__Var corresponde aux spécifications portées par *responsive Web design*, ou le développement de sites Web *réactifs*, qui s'ajustent à la taille du terminal utilisé pour les consulter.

102. Voir : https://github.com/alicemillour/Bisame.
103. Voir : http://www.cecill.info/, juin 2020.
104. Voir : https://www.culture.gouv.fr/Sites-thematiques/Langue-francaise-et-langues-de-France, juin 2020.

5.1.3 Considérations éthiques

Afin de préserver l'anonymat et la vie privée des participants, nous nous sommes appuyée sur certaines recommandations de la conception numérique responsable et durable (*Ethics by design*). À l'inscription, la seule information obligatoire à fournir est un pseudonyme et un mot de passe. Nous recommandons aux participants de choisir des pseudonymes ne permettant pas de les identifier.

Les autres informations sont facultatives et peuvent être modifiées ou supprimées à tout moment par le participant.

Une de nos plateformes permettant de saisir du texte, une fenêtre *pop-up* demande au participant de confirmer avant l'enregistrement du texte que celui-ci ne contient pas d'information pouvant nuire à sa vie privée ou à celle d'autrui. Tout texte présent sur la plate-forme peut être retiré immédiatement s'il est considéré comme inapproprié, et ce par n'importe quel utilisateur. Un texte signalé comme tel est examiné manuellement pour être supprimé de la base de données si le caractère inapproprié est avéré.

Par ailleurs, nous demandons aux participants au moment de leur inscription de confirmer qu'ils acceptent la distribution des données qu'ils produisent.

Notons que le règlement général sur la protection des données (RGPD) étant entré en vigueur en 2018 nous ne l'avons pas pris en compte au début de nos développements. Par ailleurs, nous n'avons pas demandé d'avis au Comité d'éthique de la recherche (CER), celui-ci n'ayant été mis en place à Sorbonne Université qu'en novembre 2019.

5.1.4 Évolution des plateformes

La démarche de myriadisation dans laquelle nous nous inscrivons est un processus cyclique dans lequel le dialogue avec le locuteur fait partie intégrante du développement. Après avoir développé une première plateforme, nous avons pris en compte les retours des participants ainsi que les besoins identifiés pour développer une seconde plateforme.

5.1.4.1 Première expérience de myriadisation avec P_Ann

La première plateforme P_Ann a été développée dans le seul but de proposer une interface d'annotation en parties du discours légèrement ludifiée à destination d'annotateurs non linguistes.

Cette plateforme peut-être instanciée pour toute langue pour laquelle il existe *a minima* un corpus brut, un outil de tokenisation, un petit corpus annoté de référence et un guide d'annotation. Son implémentation, pour l'alsacien puis le créole guadeloupéen a permis de mettre au jour les limites de cette première démarche.

Notre première expérience a été conçue dans l'optique d'être instanciée pour certaines langues de France. Cet objectif a conditionné certains de nos choix. En particulier, la défense des langues régionales est un enjeu de mobilisation chez une partie des locuteurs, attachés à la protection du patrimoine culturel que constituent leur langue. Notre hypothèse de départ a donc été qu'il n'était pas nécessaire de développer un jeu autour de la tâche pour motiver les locuteurs à y participer. En effet, bien que les jeux ayant but aient montré leur efficacité, le développement

de ceux-ci nécessite une expertise particulière et des développements coûteux que nous n'avons pas estimés indispensables dans le cadre de cette expérience. Cette hypothèse, forte, nous a ainsi conduit à limiter la ludification de la plateforme à un simple système de points permettant de classer les participants. Rapidement, cette hypothèse s'est révélée insuffisante. En effet, si participer à la création de ressources langagières pour leur langue motive certains à venir participer, ils ne restent pas (voir la section 6.1.1).

Par ailleurs, les variétés scripturales et dialectales inhérentes aux langues non standardisées ont posé plusieurs problèmes. D'une part, il est plus facile pour un locuteur (qu'il soit linguiste ou non) d'annoter la variété d'une langue qui lui est la plus familière. Dans le cadre d'un projet de production participative tel que le nôtre, il apparaît que proposer différentes variétés est indispensable, afin de ne pas décourager les contributeurs potentiels. En témoignent les commentaires reçus par mail et *via* le formulaire de contact mis en place sur la plateforme :

> « *J'ai dernièrement envoyé le lien vers le site à des membres de ma famille d'origine alsacienne... ils me demandent maintenant s'il faut contribuer en haut-rhinois ou en bas-rhinois... auriez-vous une idée ?* »

> « *C'est de l'alsacien haut-rhinois, pas toujours facile pour les gens du Bas-Rhin ! On a fait ce qu'on a pu.* »

D'autre part, les évaluations menées sur les ressources myriadisées (voir la section 6.2) montrent qu'il est indispensable d'avoir accès à suffisamment de corpus bruts, de qualité suffisante, pour assurer la qualité des annotations produites, et que l'absence de prise en compte de la variation peut conduire à une stagnation voire une dégradation des performances de l'outil entraîné. Or, pour certaines langues peu dotées, en particulier non standardisées, il existe peu voire pas de corpus disponible. Les ressources myriadisées sur la plateforme P_Prod_Var visent à pallier ce manque.

5.1.4.2 Vers une prise en compte de la variation avec P_Prod_Var

La plateforme P_Prod_Var a pour objectif premier de combler nos besoins en ressources brutes. Nous écartons ici l'utilisation de la transcription pour produire des textes à partir de données orales. La transcription est en effet un processus non seulement extrêmement coûteux mais aussi un travail d'interprétation (voir la section 3.4). Dans le cadre dans lequel nous nous plaçons, des données orales variées transcrites par un seul transcripteur ne rentraient en outre pas compte de la réalité des pratiques scripturales des locuteurs eux-mêmes.

Cette plateforme permet donc de collecter du corpus brut, mais aussi des variantes graphiques pour les mots des textes déposés. En outre, alors que l'annotation était réalisée « en séquence » (voir section 5.3.1) dans la première implémentation proposée, nous avons opté pour une annotation « par étiquette » (voir section 5.3.2), que nous avons imaginée moins rébarbative, lors de la seconde implémentation.

Cette seconde plateforme est plus stylisée et personnalisée que P_Ann, notamment grâce à un profil de participant plus complet avec un choix d'avatar, un accès aux profils publics des participants et à leurs contributions, et la possibilité de commenter les contenus et d'interagir entre participants au sein de la plateforme.

Aux fonctionnalités existantes (nombre de points et classement des joueurs), nous avons ajouté un ensemble de badges récompensant l'activité des participants sur la plateforme (voir figures 5.2 et 5.3). Outre les badges s'accumulant au fur et à mesure que le participant ajoute des recettes, des anecdotes, des variantes et des annotations, nous avons introduit des badges de compétence obtenus à l'issue des formations réalisées sur les catégories identifiées comme *difficiles*.

FIGURE 5.2 – Liste des badges obtenus apparaissant dans les profils privé et public du participant.

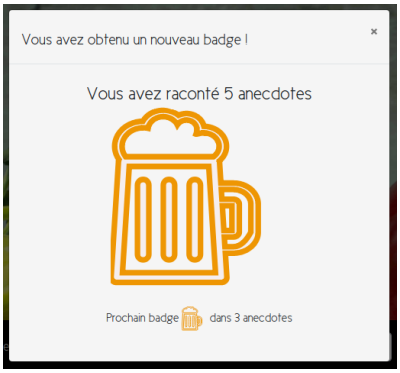


FIGURE 5.3 – Exemple de *pop-up* apparaissant lorsqu'un badge est gagné.

5.2 Préparation des ressources nécessaires

Parmi les outils nécessaires au traitement automatique d'une langue écrite quelle qu'elle soit figurent les outils de découpage en phrases, puis en « *tokens* » (voir section 5.2.2.1), et d'analyse morphologique. Pour nombre de langues aucun de ces outils n'est disponible, et ce notamment en raison du manque de ressources brutes et annotées que leur développement requiert.

Les étapes de découpage en phrase puis en « *tokens* » ainsi que la construction du jeu d'étiquettes nous ont semblé être deux tâches trop complexes pour être réalisées par des locuteurs non spécialistes de la question linguistique. Par ailleurs, en leur qualité de choix linguistiques à arrêter, ce sont des tâches se prêtant moins à la myriadisation.

Nous avons choisi de mettre à profit la participation de locuteurs pour construire une ressource annotée dynamique, en nous focalisant donc sur l'annotation en parties du discours. Le choix de cette tâche particulière a été conditionné par son caractère prioritaire dans la chaîne de traitement automatique. Nous avons conçu cette activité de classification de façon à en limiter la complexité.

Nous avons envisagé deux approches pour l'annotation de corpus en parties du discours : la première consiste à annoter une *phrase* « en séquence », la seconde consiste à annoter un *texte* « par étiquette ».

Dans les deux cas, nous avons appliqué la méthodologie utilisée avec succès pour la syntaxe en dépendances du français dans `ZombiLingo` (Guillaume *et al.*, 2016). Notamment, nous avons transformé la tâche d'annotation en tâche de *correction* des annotations produites par un outil imparfait. La pré-annotation des corpus de texte par deux outils que nous intégrons à la plate-

forme est facultative mais permet de réduire la complexité de la tâche proposée au participant en ne lui proposant que les étiquettes les plus probables.

Lorsque aucun outil d'annotation n'est disponible, il est possible de développer des outils simples, tirant parti des caractéristiques linguistiques de la langue considérée (nous en donnons des exemples dans la section 5.2.4.1). Une autre possibilité est d'entraîner un premier outil supervisé avec un corpus d'entraînement de taille très réduite. Cet outil sera ensuite remplacé par un outil plus performant, dès lors que la taille du corpus annoté *via* la plateforme le permettra. Ce processus est illustré dans la figure 5.4.

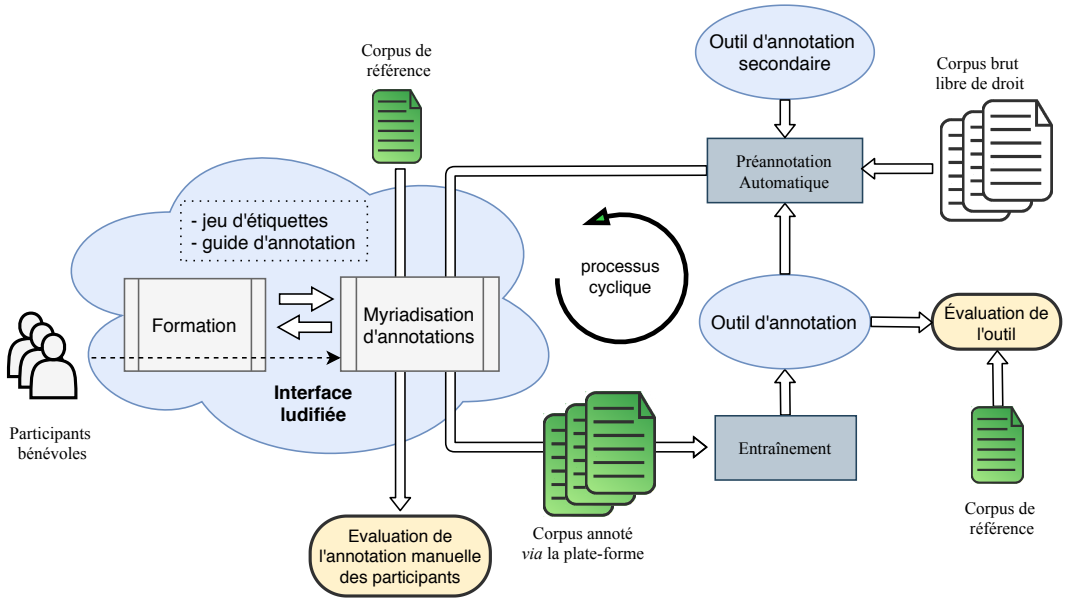


FIGURE 5.4 – Pré-annotation et intégration continue des annotations myriadisées.

Les éléments nécessaires sont donc les suivants :

- un corpus textuel alimentant la plateforme C_{Brut},
- un outil de tokenisation,
- un ensemble de conventions de découpage des tokens en *mots morphosyntaxiques* accompagné d'un jeu d'étiquettes adapté au découpage arrêté,
- un (ou plusieurs) outil(s) de pré-annotation,
- un corpus de référence permettant de former et d'évaluer les participants C_{Ref}.

Nous présentons dans la section 5.2.1 comment nous avons mené la collecte manuelle de corpus textuels. Nous poursuivons avec le détail du développement de tous les autres éléments mentionnés ci-dessus dans la section 5.2.2. Une fois les choix de tokenisation et d'étiquetages arrêtés, nous présentons les corpus de référence que nous avons construits dans la section 5.2.3.

Enfin nous consacrons la section 5.3 à la présentation de l'implémentation des deux tâches d'annotation sur nos plateformes.

5.2.1 Collecte manuelle de corpus textuels

Suivant la contrainte « C1 : Pérennité des ressources produites » présentée dans l'introduction, nous nous sommes attachée à n'alimenter la plateforme d'annotation qu'en corpus libres de droit afin de pouvoir les redistribuer une fois annotés. Nous présentons ici les corpus collectés classés par source et par langue. La source la plus importante de corpus sous licence claire est Wikipédia. Ensuite, nous présentons l'exploitation d'une ressource produite par des chercheurs. Enfin nous présentons d'autres sources de corpus. Nous avons suivi une approche pragmatique (McEnery et Hardie, 2011) pour la collecte de ces corpus. Pour chacune des sources nous décrivons la ressource obtenue pour les différentes langues auxquelles elles s'appliquent.

Dans le cas de l'alsacien et du créole guadeloupéen, la collecte de corpus s'est effectuée dans le but d'alimenter la plateforme d'annotation. Dans le cas du créole mauricien, langue pour laquelle seule la plateforme `Recettes de Grammaire` a été déployée, il s'est seulement agi de collecter un petit corpus provenant du maximum de sources possible afin qu'il soit annoté manuellement pour former une référence.

Dans toute cette section et en première approximation, nous appelons dans cette section « *tokens* » l'ensemble des entités linguistiques séparées par des espaces et « types » l'ensemble des « *tokens* » dédoublonnés.

5.2.1.1 Wikisphère

Section alsacienne de la Wikipédia alémanique :

Nous décrivons ici le processus mis en place pour extraire un corpus de l'alsacien à partir de la Wikipédia alémanique. Les chiffres présentés dans cette section correspondent à la situation en janvier 2018. Ils ont été obtenus grâce aux données distribuées par Wikipédia [105]. En particulier, les liste de catégories, liens entre articles et catégories sont disponibles sous format SQL. Nous mettons à disposition sur GitHub les requêtes SQL et scripts développés pour obtenir les données ci-dessous [106].

La Wikipédia alsacienne fait partie de la Wikipédia alémanique dont la page d'accueil annonce cinq sections correspondant à cinq familles de dialectes alémaniques : le *Schwyzerdütsch* (suisse alémanique), le *Badisch* (badois), l'*Elsassisch* (dialectes alsaciens), le *Schwäbisch* (allemand souabe) et le *Vorarlbergisch* (dialecte autrichien parlé au Vorarlberg)).

Une partie des articles (18 % environ) est associée à une catégorie de la forme `Artikel_uf_-dialecte`, littéralement « article_en_*dialecte* », spécifiant le dialecte dans lequel il a été rédigé.

La liste de catégories de la forme `Artikel_uf_dialecte` contient 38 catégories, par exemple : `Artikel_uf_Aargauerdüütsch` ou `Artikel_uf_Oberrhiinalemannisch`. Parmi ces catégories, la catégorie `Artikel_uf_Elsassisch` (littéralement « article_en_alsacien ») compte 1 888 articles. Il est probable que d'autres pages de la Wikipédia alémanique soient également écrites en alsacien, mais seul un examen manuel des pages permettrait de les identifier.

105. Voir : https://dumps.wikimedia.org/alswiki/latest/, juin 2020.
106. Voir : https://github.com/alicemillour/export_alswiki, juin 2020.

Un premier export de ces articles a été réalisé grâce à l'outil d'export par catégorie ou par article fourni par Wikipédia permettant de télécharger des *dumps* de l'encyclopédie au format xml [107].

89 % de ces articles appartiennent également à une catégorie commençant par `Ort_`, c'est-à-dire les catégories spécifiant qu'il s'agit d'un article traitant d'un lieu géographique. Nous avons donc créé un deuxième export sans ces articles dont un examen manuel a montré qu'il s'agissait de pages très similaires entre elles [108]. Ce deuxième export se réduit à 196 articles.

L'extraction des textes contenus dans les articles à partir des *dumps* fournis par Wikipédia a été réalisée grâce au script `WikiExtractor.py` développé par G. Attardi [109]. Nous avons gardé les paramètres par défaut du script ne gardant que le texte des articles Wikipédia et excluant en particulier les titres de sections, sous-sections et les tableaux pouvant être présents dans l'article initial.

Ressources Wikimédia pour le créole guadeloupéen :

Seule une version *bêta* [110] de la Wikipédia pour le créole guadeloupéen existe dans l'incubateur Wikimédia. Celui-ci héberge tous les projets de la wikisphère avant qu'ils n'arrivent à maturation. L'incubateur du créole guadeloupéen contient 17 articles (mars 2018). Nous avons également recueilli les proverbes publiés sur la page Créole guadeloupéen de la Wikipédia française [111]. Ce corpus contient 74 phrases et totalise 638 *tokens* et 363 types.

Ressources Wikimédia pour le créole mauricien :

De la même manière que le créole guadeloupéen, le créole mauricien ne dispose que d'un incubateur Wikimédia de sa Wikipédia. Celui-ci contient 78 articles, 373 phrases, 7 490 *tokens* et 1 318 types.

Les tailles des corpus obtenus collectés sur les différents projets de la Wikisphère sont données dans le tableau 5.1. Concernant les statistiques de la Wikipédia alémanique, nous avons utilisé les informations fournies par Wikipédia [112].

Export	**Nb. articles**	**Nb. phrases**	**Nb. *tokens***	**Nb. « types »**
Wikipédia alémanique	27 229	inconnu	11 412 150	inconnu
$CWiki_{gsw\_complet}$	1 888	11 139	230 270	33 174
$CWiki_{gsw\_sans\_lieux}$	196	2 236	57 769	16 127
$CWiki_{gcf}$	17 + 27 proverbes	101 (74 + 27)	825 (638 + 187)	503 (363 + 140)
$CWiki_{mfe}$	78	373	7 490	1 318

Tableau 5.1 – Taille des corpus issus de la Wikisphère pour l'alsacien (`gsw`), le créole guadeloupéen (`gcf`) et le créole mauricien (`mfe`).

107. voir : https://als.wikipedia.org/wiki/Spezial:Exportiere, juin 2020.
108. Voir par exemple les articles des communes de Wirminge, Wisches, Wittelsheim, Wittene, Wittre etc.
109. Le script et sa documentation sont disponibles sur GitHub : https://github.com/attardi/wikiextractor.
110. Voir : https://incubator.wikimedia.org/wiki/Wp/gcf, juin 2020.
111. Voir : https://fr.wikipedia.org/wiki/Creole_guadeloupeen, juin 2020.
112. Voir : https://als.wikipedia.org/wiki/Spezial:Statischtik, juin 2020.

5.2.1.2 Plateformes de ressources linguistiques

Une autre source de corpus importante sont les ressources développées par les chercheurs et archivées par des bases de données linguistiques. Peu de travaux étant menés sur les langues d'intérêt, la seule ressource dont nous avons pu tirer parti est un corpus transcrit ayant été développé pour le créole guadeloupéen et distribué sous licence CC BY-NC-SA [113] sur CoCoON.

Nom	**Genre**	**Nb. gpes de s.**	**Nb. « tokens »**	**Nb. « types »**
creole1	Monologue	42	289	133
creole2	Monologue	82	626	264
enfance_en_guadeloupe	Monologue	79	1 010	337
journal	Conversation	242	1 969	530
langues_des_signes	Récit	196	1 802	563
marie_Galante_1	Conversation	213	1 782	712
marie_Galante_2	Conversation	80	684	334
recit_d_enfance	Récit	146	1 019	358
Total	Mixte	1 180	9 181	2 873

Tableau 5.2 – Description du corpus collecté sur CoCoON (« gpes de s. » : groupes de souffle).

CoCoON est « une plateforme technique qui accompagne les producteurs de ressources orales, à créer, structurer et archiver leurs corpus » [114]. Cette ressource, destinée à recueillir principalement des données audio, contient 12 304 enregistrements pour 245 langues déposés par 55 éditeurs totalisant 5 281,7 heures d'écoute. Outre les enregistrements sont parfois fournies les transcriptions.

C'est le cas de la ressource ayant été collectée pour le créole guadeloupéen par le Laboratoire de Linguistique Formelle (LLF, Paris) et le Laboratoire Ligérien de Linguistique (LLL, Orléans), qui contient 11 enregistrements transcrits. Cette ressource n'est pas distribuée comme un ensemble, chaque enregistrement faisant l'objet d'une entrée différente dans la plateforme, comme par exemple l'entrée *Histoire de l'âne en créole guadeloupéen* (Soare *et al.*, 2016).

Nous avons utilisé huit des 11 enregistrements, les trois derniers contenant majoritairement des énoncés en français. Le détail des enregistrements est donné dans le tableau 5.2.

Notons qu'une ressource d'intérêt pour l'alsacien a été déposée en mai 2020 (Ruiz Fabo *et al.*, 2020) sur GitLab [115]. Cette ressource contient cinq pièces de théâtre datant d'entre 1816 et 1905 encodées en TEI et distribuées sous licence CC0 1.0 ou CC BY 2.0 selon les cas. Étant donné sa date de publication récente, nous n'avons pas pu exploiter cette ressource dans le cadre de ce travail.

113. Voir : https://creativecommons.org/licenses/by-nc-sa/3.0/, juin 2020.
114. Voir : https://cocoon.huma-num.fr/, juin 2020.
115. GitLab est « un logiciel libre de forge basé sur git proposant les fonctionnalités de wiki, un système de suivi des bugs, l'intégration continue et la livraison continue. », page Wikipédia https://fr.wikipedia.org/wiki/GitLab, juin 2020.

5.2.1.3 Autres sources de corpus

Les autres sources de corpus sont celles produites par des organismes publics ou des particuliers. Dans le cas de l'alsacien, nous avons ainsi eu accès à des textes distribués par l'OLCA, l'Office pour la langue et la culture d'Alsace [116], ou gracieusement fournis par des participants. En particulier, un participant, Raymond W., nous a transmis une nouvelle intitulée *E Hochzit in de 50er Johre* (« *Un mariage dans les années 50* »), que nous avons intégrée à notre corpus de l'alsacien.

Dans le cas du créole mauricien, nous avons eu l'autorisation d'utiliser et de redistribuer certains textes poétiques et en prose de l'homme politique et de lettres mauricien D. Virahsawmy [117]. Ont également été intégrées au corpus des phrases d'exemples du dictionnaire en ligne *Morysien Dictionary* d'Andras Rajki [118] et de *Lortograf Kreol Morisien* (Carpooran, 2011), un conte produit par une locutrice du créole mauricien, et un extrait d'un poème de Bertolt Brecht traduit par Lindsey Collen [119].

5.2.1.4 Résumé des corpus textuels

Le tableau 5.3 résume les corpus textuels à disposition lorsque nous avons démarré nos expériences sur les trois langues.

Langue	**Nom**	**Nb. phrases (ou gpe. de s.)**	**Nb. « tokens »**
Alsacien	$CBrut_{gsw}$	2 302	59 537
Créole guadeloupéen	$CBrut_{gcf}$	1 281	10 006
Créole mauricien	$CBrut_{mfe}$	79	1 024

Tableau 5.3 – Résumé des corpus collectés pour l'alsacien (`gsw`), le créole guadeloupéen (`gcf`) et le créole mauricien (`mfe`) (« gpes de s. » : groupes de souffle).

5.2.2 Premiers traitements sur les corpus

5.2.2.1 Tokenisation

Les systèmes d'écriture de l'alsacien et des créoles guadeloupéen et mauricien sont des systèmes à séparateurs typographiques. Dans ce contexte, la première opération que nous avons réalisée consiste donc à définir les séparateurs et à isoler les *tokens* qu'ils délimitent :

> « On peut [...] définir une unité typographique, *mot typographique* ou *token*, de la façon suivante : un token est une séquence contiguë de caractères délimités de

116. Voir : https://www.olcalsace.org/.
117. Voir : https://boukiebanane.com/table-of-contents-konteni/poezi-dev-devs-poetry/ et https://boukiebanane.com/table-of-contents-konteni/proz-literer-dev-devs-literary-prose/.
118. Voir : http://web.archive.org/web/20101010120405/http://www.freeweb.hu/etymological/Morisyenweb.htm
119. Voir : https://www.lalitmauritius.org/modules/docpool/files/x-doc-lindsey-pu-bann-ki-puvinn-apre-nu.pdf.

> part et d'autre par un séparateur typographique ou par un signe de ponctuation ; par ailleurs, un signe de ponctuation (ou, dans certaines définitions, une séquence contiguë de signes de ponctuation) est token en soi. (Sagot, 2018, p. 33) »

Notons que cette opération, réalisée sur une langue non standardisée, peut produire des tokenisations multiples pour une forme donnée. C'est ce qui est observé lorsque certaines formes peuvent être agglomérées, ou non, et que leurs graphies peuvent ainsi contenir un séparateur. Par exemple, en alsacien, les trois graphies suivantes sont acceptées : `ìnere␣Stund`, `ìn␣ere Stund`, `ìn're Stund` (« dans une heure ») (Bernhard, 2018b).

Par ailleurs, la tâche de tokenisation est affectée par la présence dans ces langues de marques de l'oralité. La première de ces marques est l'élision. Dans le cas de l'alsacien ou du créole mauricien, celle-ci peut être transcrite par une apostrophe. Il faut donc distinguer les cas de l'apostrophe signalant une frontière de *token* (par exemple, en alsacien, `d'Junga` doit être tokénisé `d'␣Junga` (« le jeune »), tandis que `d'r` (« le ») doit être maintenu tel quel, l'apostrophe y marquant l'élision de la voyelle `e`). La seconde est l'inscription de l'épenthèse. C'est le cas notamment du <n> euphonique en alsacien (par exemple : `fânga-n-â`, « commencé à »), ou le <z> en créole mauricien (par exemple : `de-z-er` (« deux heures »).

Si la tokenisation de langues non standardisées représente une gageure c'est que l'ensemble des pratiques scripturales n'est pas connu en amont de la conception du tokéniseur. L'augmentation en taille d'un corpus collecté auprès de locuteurs aux pratiques variées permet de remettre en question les conventions de tokenisation régulièrement et d'améliorer la couverture des outils.

Nous avons utilisé comme point de départ le tokéniseur de l'alsacien fourni par D. Bernhard en 2016 que nous avons adapté aux créoles guadeloupéen et mauricien. Le tokéniseur pour l'alsacien a depuis été complété et distribué (Bernhard, 2018b). Y sont distingués six « types » de *tokens* : `abbreviation`, `url`, `email`, `number`, `separator` et `word`. Les cinq premières sont définies explicitement comme des expressions régulières.

Les trois tokéniseurs `tokeniseur_gsw`, `tokeniseur_gcf` et `tokeniseur_mfe` partagent leurs définitions des types `abbreviation`, `url`, `email`, `number`. Nous avons en revanche défini pour chacun les expressions régulières du type `separator`, lui-même divisé en deux sous catégories `begin_sep` et `end_sep`. Nous présentons ces séparateurs dans le tableau 5.4 [120].

Dans les cas des créoles guadeloupéen et mauricien, nous n'avons considéré le tiret comme étant un séparateur que dans le cas du déterminant postposé (par exemple en créole mauricien : *-la* dans *deziem-la* (« la deuxième »), qui peut également être rencontré sous la forme graphique séparée `deziem␣la`). Nous n'avons pas en revanche séparé les cas dont le découpage produisait des *tokens* amalgamés : par exemple, toujours en créole mauricien, *sink-er* (« cinq heures ») est la contraction de *sink* (« cinq ») et de ***l**er* (« heure »), mais *er* seul n'a pas d'existence autonome.

5.2.2.2 Découpage en segments pour l'annotation

Afin de préparer la tâche d'annotation, nous avons découpé nos corpus en « segments annotables ». La définition de ces segments est étroitement liée dans le cas de nos expériences aux jeux d'étiquettes que nous avons choisi d'utiliser pour chaque langue, notamment concernant le

120. Le tokéniseur distribué depuis 2018 par Bernhard (2018b) a été complété depuis pour l'alsacien avec de nouveaux cas de séparateurs. Un troisième type y est défini : `mid_sep` utilisé uniquement dans le cas des épenthèses « -n- » et « -w- » qui marquent la liaison phonétique.

Langue	begin_sep	end_sep
Alsacien	[dsz]['‘´`](? !r)	(?< !d)(-|,| :|['‘´`][mrs])
Créole guadeloupéen		(-lasa|-la)
Créole mauricien		(-la)

TABLEAU 5.4 – Expressions régulières des séparateurs spécifiques de début et fin de *token* pour les trois langues.

découpage et l'annotation des amalgames. Il ne s'agit pas pour nous de proposer ici un point de vue prescriptif sur la bonne manière de préparer un texte pour une tâche d'annotation en parties du discours. Nous croyons en revanche à la nécessité de documenter les choix de découpage effectués.

Une fois arrêtées les règles de *tokenisation* nous avons défini les règles de découpages en *mots morphosyntaxiques* tels que définis par Sagot (2018) :

> « Nous parlerons de *mot morphosyntaxique* pour dénoter les unités (minimales) auxquelles on peut conférer une partie du discours. » (Sagot, 2018, p. 30)

La définition proposée dépend donc de celle du jeu de parties du discours utilisé. Il est néanmoins d'usage d'accompagner la tokenisation de deux opérations :

- le **regroupement des formes composées** devant être associée à une unique partie du discours, c'est-à-dire les locutions lexicales du type « pomme de terre », « à propos » etc.,
- et le **découpage des amalgames**, par exemple dans le cas du français « du », amalgame de « de » et « le », associés à leurs parties du discours respectives.

Dans la suite de ce manuscrit, nous utiliserons *mot morphosyntaxique* pour nous référer aux entités pouvant être annotées en parties du discours compte tenu des opérations de regroupement et de découpage effectuées et des jeux d'étiquettes arrêtés.

Le jeu d'étiquettes que nous avons utilisé comme base est celui proposé par le projet *Universal Dependencies* (UD) (Petrov *et al.*, 2012), soit un jeu formé par 17 étiquettes et conçu pour harmoniser les annotations morphosyntaxiques sur des corpus dans plusieurs langues. Le choix de ce jeu d'étiquettes a donc notamment été motivé par notre volonté de proposer une méthodologie qui soit facilement réplicable à un nouveau contexte linguistique [121]. Notre choix de nous limiter en première instance à ce jeu d'étiquettes sans y intégrer les traits proposés par le projet UD (par exemple le `mode` des verbes ou le `genre` des noms communs) a été motivé par un enjeu de complexité. L'annotation étant destinée à être réalisée par des locuteurs non linguistes, il nous a semblé raisonnable en première instance de procéder à la classification des étiquettes, l'ajout des traits pouvant être réalisée dans un second temps.

Nous avons suivi la sémantique de ce jeu d'étiquettes [122] sauf dans le cas de deux étiquettes que nous avons jugées comme ambiguës.

La première est l'étiquette AUX (*Auxiliary*) :

> « [un auxiliaire] est souvent un verbe (qui peut être utilisé avec son sens plein par ailleurs) mais de nombreuses langues ont des marqueurs de temps, mode, aspect,

121. Pour plus de détails sur le terme « réplicable », voir (Cohen *et al.*, 2018) et la section 7.2.
122. La documentation de chaque étiquette est disponible ici : https://universaldependencies.org/u/pos/index.html.

Classes ouvertes	Classes fermées	Autres
ADJ	ADP	PUNCT
ADV	AUX	SYM
INTJ	CCONJ	X
NOUN	DET	
PROPN	NUM	
VERB	PART	
	PRON	
	SCONJ	

TABLEAU 5.5 – Liste des étiquettes utilisées selon le classement de ses créateurs (Petrov *et al.*, 2012).

> voix, et ceux-ci doivent aussi être étiquetés comme AUX. » [123]

Les créoles guadeloupéen et mauricien présentant des particules de ce type, nous avons décidé de proposer une étiquette supplémentaire aux participants pour ces cas spécifiques, PART_TMA, dont la signification nous a paru plus claire que celle d'« auxiliaire » [124]. Il suffit de remplacer ces étiquettes par AUX dans le corpus une fois annoté pour devenir conforme aux recommandations UD.

La seconde étiquette dont nous avons modifié la sémantique est l'étiquette X (*Other*), présentée comme l'étiquette à utiliser lorsque aucune autre étiquette ne convient. Nous avons restreint son usage aux cas d'alternance codique [125], à différencier de la présence d'emprunts qui doivent être annotés avec l'étiquette qui leur correspond. Par exemple, en alsacien, la présence de la séquence en français dans un dialogue : « *Toi*/X *tais-toi*/X » (extrait d'un texte alsacien) est bien annotée avec la catégorie X. En revanche, les occurrences telles que « plafond » dans *bis an de plafond*/NOUN (« jusqu'au plafond ») ne le sont pas.

Les autres écarts à ces recommandations que nous documentons ci-dessous ont été motivés par les limitations de nos outils de tokenisation, et par la contrainte C3 présentée dans l'introduction : le caractère participatif et bénévole de la production d'annotation. En particulier, il s'est agi dans notre cas de proposer un découpage en mots morphosyntaxiques qui privilégie la bonne lisibilité du texte. Une fois le découpage déterminé, nous y avons fait correspondre un jeu d'étiquettes adapté.

Regroupement des formes composées :

La documentation d'UD préconise le seul regroupement des « mots multitokens » (*multitoken words*), c'est-à-dire des expressions numériques et des abréviations. Ces cas ont été traités au moment de la tokenisation. Les formes composées ne sont pas regroupées, l'annotation de celles-ci s'effectuant au niveau syntaxique :

> « Notez, cependant, que [le regroupement de tokens] ne doit pas être généralisé aux expressions multimots tels que *in spite of* et *by and large* (sans parler d'expres-

123. « *It is often a verb (which may have non-auxiliary uses as well) but many languages have nonverbal TAME markers and these should also be tagged aux.* », voir : https://universaldependencies.org/u/pos/aux_.html, juin 2020.
124. L'utilisation de l'étiquette AUX plutôt que PART pour les marqueurs de temps, mode, aspect et voix fait débat, voir par exemple la discussion menée ici : https://github.com/UniversalDependencies/docs/issues/625.
125. Voir : https://universaldependencies.org/u/pos/X.html, juin 2020.

> sions multimots plus flexibles tels que les mots composés ou les verbes à particules), qui doivent être annotés par le biais de relations de dépendances spécifiques. » [126]

Nous avons suivi ces recommandations à l'exception du cas des adverbes *ki jan* (« comment »), *ki tan / ki lè* (« quand »), *ki koté* (« où »), *ki moun* (« qui ») en créole guadeloupéen. Si un locuteur francophone pourrait être tenté d'analyser la séquence en l'annotant *ki*/PRON *jan*/NOUN, ce découpage se traduisant littéralement « quel genre » en français, cela va à l'encontre de l'intuition d'un locuteur natif.

Découpage des amalgames :

La documentation d'UD préconise le découpage systématique des amalgames ou « tokens multimots » (*multiword tokens*) :

> « [...] l'unité d'annotation sont les mots syntaxique (pas les mots phonologiques ou orthographiques), ce qui signifie le découpage systématique des clitiques, comme dans le cas espagnol « *dámelo* » = « *da me lo* », et le découpage des contractions comme dans le cas français « au » = « à le ». » [127]

Nous avons décidé de ne pas suivre ces recommandations pour deux raisons. D'abord, en raison notamment de la variété graphique observée, nous ne disposions pas de ressources permettant d'assurer un découpage automatique de qualité. Ensuite, ne pas altérer la forme graphique originale facilite la lisibilité pour les locuteurs, et évite de faire apparaître au moment de l'annotation des segments n'ayant pas d'usage isolé.

Nous avons donc étendu ce jeu d'étiquettes, à la manière de Hollenstein et Aepli (2014) qui ont adapté le Stuttgart-Tübingen-TagSet (STTS) (Schiller *et al.*, 1995), standard pour l'allemand, aux caractéristiques du suisse allemand, c'est-à-dire en ajoutant la possibilité de signaler la concaténation de catégories grammaticales lors de l'annotation d'un mot morphosyntaxique.

Le découpage peut être effectué *a posteriori* : tout segment résultant de la contraction de deux segments et ayant été annoté avec une étiquette composée CAT_1+CAT_2 peut être transformé en « segment_1/CAT_1 segment_2/CAT_2 » une fois les règles de découpage connues.

Le tableau 5.6 liste pour les trois langues les catégories amalgamées ajoutées au jeu d'étiquettes.

5.2.2.3 Rédaction du guide d'annotation pour la myriadisation

Pour chacune des langues, un guide d'annotation simplifié à destination des participants a été rédigé. Chaque guide est rédigé en français et est organisé, pour chaque catégorie, en deux parties :

- une liste d'exemples servant d'aide-mémoire,

126. « *Note, however, that [combining tokens] should not be generalized to multiword expressions like* in spite of *and* by and large *(let alone to more flexible multiword expressions like compounds or particle verbs), which should instead be annotated using special dependency relations.* », voir : https://universaldependencies.org/u/overview/tokenization.html, juin 2020.

127. « *[...] the basic units of annotation are syntactic words (not phonological or orthographic words), which means that we systematically want to split off clitics, as in Spanish dámelo = da me lo, and undo contractions, as in French au = à le.* », voir : https://universaldependencies.org/u/overview/tokenization.html, juin 2020.

Langue	**Étiquette**	**Exemple**
Alsacien	ADP+DET	***am*** *(an+dem)* (« au/à la »)
Créole mauricien	ADV+PART_TMA	*Personn* ***pa'nn*** *dir twa tap laport. (pa+finn)* (« On ne t'a pas dit de toquer. »)
	NUM+NOUN	*Ziska* ***sink-er*** *tanto. (sink+ler)* (« Jusqu'à cinq heures de l'après-midi. »)
	PART_TMA+PART_TMA	***Ti'a*** *bon gagn zot lopinion. (ti+ava)* (« Nous espérons avoir leur opinion. »)
	PRON+ADV	***Mo'si*** *mo enn bouro ? (mo+osi)* (« Suis-je également un bourreau ? »)
	PRON+PART_TMA	***Mo'nn*** *konpran. (mo+finn)* (« J'ai compris. »)
	PRON+VERB	Toutswit si **to'le**. *(to+oule)* (« Dès maintenant si tu le veux. »)
Créole guadeloupéen	ADP+PRON	*pati* ***evè'y*** *(evè+i)* (« avec lui/elle »)
	PART_TMA+VERB	*Yo* ***k'ay*** *Gozié (ka+ay)* (« Je vais à Gozié »)
	PRON+PRON	*A pa* ***sa'w*** *té di mwen (sa+ou)* (« Ce n'est pas ce que tu m'avais dit »)
	VERB+PRON	*I manké* ***trapé'y*** *(trapé+i)* (« Il ne l'a pas attrapé. »)

Tableau 5.6 – Liste des étiquettes ajoutées par langue.

- une section « **ATTENTION** » visant à attirer l'attention de l'annotateur sur les sources d'erreur qui ont été identifiées. Nous avons tâché d'identifier les mots morphosyntaxiques ambigus, et d'alerter sur les catégories pouvant être confondues entre elles (nous en donnons des exemples dans la figure 5.5).

Le guide pour l'annotation de l'alsacien est fortement inspiré de l'édition 2016 du guide d'annotation morphosyntaxique pour les dialectes alsaciens (Bernhard *et al.*, 2016).

5.2.3 Corpus annotés de référence

Les corpus de référence $CRef_{gsw}$, $CRef_{gcf}$ et $CRef_{mfe}$ ont été annotés manuellement respectivement par :

- D. Bernhard et L. Steiblé, chercheuses du laboratoire LiLPa de Strasbourg,
- une étudiante guadeloupéenne, deux expertes de l'annotation et A. Thibault, professeur en linguistique à Sorbonne Université et créolophone,

- une étudiante mauricienne et une experte de l'annotation, sous la supervision de F. Henri, chercheuse en linguistique mauricienne en post-doctorat à l'université du Kentucky (*University of Kentucky*, États-Unis).

L'annotation de ces corpus a permis d'examiner les choix à effectuer en termes de tokenisation et de jeu d'étiquettes correspondants présentés dans les sections 5.2.2.1, 5.2.2.2 et 5.2.2.3.

Le tableau 5.7 détaille les contenus de ces corpus, tous composés d'extraits provenant de diverses sources, conformément à la méthodologie « opportuniste » décrite dans la section 5.2.1.

Langue	**Nom**	**Nb. phrases (ou gpe. de s.)**	**Nb. mots morpho.**
Alsacien	$CRef_{gsw}$	102	1 468
Créole guadeloupéen	$CRef_{gcf}$	100	1 623
Créole mauricien	$CRef_{mfe}$	79	1 024

TABLEAU 5.7 – Résumé des corpus de références pour l'alsacien (`gsw`), le créole guadeloupéen (`gcf`) et le créole mauricien (`mfe`) (« gpes de s. » : groupes de souffle).

5.2.4 Annotation manuelle de référence

Cas de l'alsacien :

Nous avons eu la possibilité de calculer l'accord inter-annotateur entre les deux chercheuses ayant annoté le corpus de référence de l'alsacien. Ce corpus comprend quatre sous-corpus : `Hoflieferant_p53`, `recettes`, `wikipedia1`, et `wikipedia2`. Les résultats des calculs des coefficients κ de Cohen (Cohen, 1960) et π (Scott, 1955) sont présentés par sous-corpus dans le tableau 5.8.

Corpus	Hoflieferant_p53	recettes	wikipedia1	wikipedia2
Accord observé	0,92	0,87	0,91	0,89
Coefficient κ de Cohen	0,91	0,85	0,90	0,88
Coefficient π	0,91	0,85	0,90	0,88

TABLEAU 5.8 – Accords inter-annotateur calculés pour les annotations manuelles fournies pour l'alsacien.

Les coefficients calculés sont tous supérieurs à 0,8, c'est-à-dire, selon le critère défini par Artstein et Poesio (2008), que l'annotation manuelle fournie est de bonne qualité. On constate également que les coefficients κ de Cohen et π sont quasiment égaux, ce qui signifie que le biais des deux annotatrices est très faible.

Les matrices de confusion des deux annotations manuelles fournies sont disponibles en annexe A.

Les désaccords portent en majorité sur les catégories suivantes :

- ADJ et ADV : par exemple *licht* (« légèrement ») ;
- ADP et ADV : par exemple *zamme* (« ensemble ») ;
- ADV et CONJ : par exemple *àwwer* (« mais ») ;

- AUX et VERB : par exemple *ìsch* (« est »).

Les divergences entre les deux annotatrices, principalement dues à des contextes ambigus, ont été résolues dans $CRef_{gsw}$, l'adjudication fournie par les deux expertes que nous utilisons comme référence.

Cas du créole guadeloupéen :

Dans le cas du créole guadeloupéen, le corpus de référence est composé de 100 phrases extraites de $CBrut_{gcf}$, qui contient de extraits de l'incubateur Wikipédia du créole guadeloupéen et des transcriptions provenant du corpus CoCoON.

Alors que les phrases tirées de l'incubateur peuvent être immédiatement annotées, nous avons dû effectuer un pré-traitement sur C_{Speech} pour obtenir des séquences grammaticalement correctes de mots morphosyntaxiques pouvant effectivement être annotés. Notamment, la présente de achoppements, transcrites dans le texte brut, rendaient certains énoncés difficiles à comprendre. Certains « groupes de souffle » sortis de leur contextes devenaient des énoncés agrammaticaux et impossibles à annoter.

En conséquence, nous avons été contraints de modifier le corpus original de deux manières :

- en procédant à l'élimination des achoppements telles que les ellipses générant des découpages inattendus au milieu d'un mot morphosyntaxique. C'est le cas dans l'extrait suivant du mot morphosyntaxique *gwoka*, un genre de musique guadeloupéen :
  1. « *sé pou sa jodijou nou ka respékté gwo...* »
     (« c'est pourquoi nous respectons le gros... »)
  2. « *ka* » (« tambour »)

  En fait, et bien que *gwo ka* existe sous cette forme séparée, ce n'est pas la convention utilisée dans cet extrait. Par ailleurs, cette séparation rend les énoncés incomplets. *ka* pris isolément est par ailleurs ambigu et peut être NOUN ou PART, en fonction du contexte ;
- en regroupant les tours de parole afin de produire des énoncés complets. C'est la cas par exemple des énoncé :
  1. « *Lagwadloup dévlopé pli* »
     (« La Guadeloupe s'est développée plus »)
  2. « *vit sé on grand tè* »
     (« rapidement, c'est un gros territoire »)

Au final, le corpus annoté contient un échantillon de phrases déclaratives, interrogatives, impératives, simples ou complexes, de tailles et de type de discours variés.

5.2.4.1 Outils de pré-annotation

Comme présenté dans la méthodologie illustrée par la figure 5.4, nous avons intégré deux outils de pré-annotation à chacune des instances développées.

Dans le cas de l'alsacien, nous avons utilisé le `Stanford POS Tagger` (Toutanova *et al.*, 2003) pour l'allemand, employé selon la méthodologie définie par Bernhard et Ligozat (2013), ainsi

que `MElt` (Denis et Sagot, 2010), entraîné au fur et à mesure de la croissance du corpus d'entraînement annoté *via* la plateforme.

Dans le cas du créole guadeloupéen, aucun outil d'annotation n'étant disponible à notre connaissance, nous avons développé un script `Python` tirant parti de la faible flexion du créole guadeloupéen et de l'importante fréquence absolue des mots morphosyntaxiques les plus fréquents : par exemple, la particule *ka* représente 4,6 % du corpus brut, le pronom *an* (« je »), 3,6 %, le verbe *sé* (verbe « être », sous ses formes infinitive et conjuguées), 2,8 % etc. Nous avons extrait du corpus de référence une liste des 100 couples [mot morphosyntaxique - étiquette] non ambigus les plus fréquents que nous avons utilisés pour annoter le corpus brut. Cette liste n'est pas représentative des mots les plus fréquents en créole guadeloupéen, mais nous a néanmoins permis d'annoter 37 % du corpus.

Notre deuxième outil de pré-annotation est `MElt` (Denis et Sagot, 2012), entraîné sur le corpus de référence $CRef_{gcf}$ et utilisé sans lexique additionnel : MElt_{gcf}. Nous avons procédé de la même manière pour le créole mauricien, et utilisé $CRef_{mfe}$ pour entraîner un unique modèle pour le créole mauricien : MElt_{mfe}.

Étant donné la taille des données dont nous disposons, nous avons choisi d'évaluer ces outils en procédant à une validation croisée sur ces corpus. Cette technique permet d'utiliser l'intégralité du corpus à disposition à la fois pour l'entraînement et pour l'évaluation de l'outil. Si cette technique peut être utilisée pour l'ajustement (*fine tuning*) des paramètres au cours du *développement* d'un modèle, nous l'utilisons bien dans notre cas pour l'*évaluation* du modèle (Vabalas *et al.*, 2019).

Pour chaque corpus d'entraînement de taille N mots, nous choisissons un paramètre k correspondant au nombre de blocs. Le modèle est entraîné sur un corpus de $k-1$ blocs (et de taille $\frac{(k-1)*N}{k}$ mots) et évalué sur le bloc restant (de taille $\frac{N}{k}$ mots). Ce processus est répété k fois de manière à entraîner k modèles. L'exactitude que nous donnons correspond à la moyenne des exactitudes obtenues sur chacun des k blocs d'évaluation. La valeur donnée est donc une estimation de la qualité du modèle, chaque outil entraîné présentant des performances variées.

Les chiffres que nous proposons sont indicatifs, cette méthode d'évaluation présentant de nombreux biais. Notamment, lorsque les classes ne sont pas distribuées uniformément dans le corpus, la méthode de la validation croisée présente de sérieuses limites. Notamment, il est possible que les classes les plus rares soient absentes de certains des blocs d'entraînement ou d'évaluation. Dans ces cas, le modèle ne peut être entraîné ou évalué que sur les classes majoritaires. Une solution pour remédier à ce biais consiste à ne pas découper aléatoirement les blocs mais à les *construire* de façon à assurer une distribution équivalente de chaque classe dans chacun des blocs. Dans le cadre de l'annotation séquentielle de phrases en parties du discours un tel découpage ne peut être effectué, les classes étant dépendantes entre elles.

Nous avons donc évalué nos outils de pré-annotation grâce à une validation croisée avec $k = 10$: l'exactitude moyenne de MElt_{gcf} ainsi calculée est de 82,28±0.005 %, celle de MElt_{mfe} de 76±0.011 %.

Nous savons que la pré-annotation introduit un biais (Fort et Sagot, 2010), auquel les utilisateurs les moins formés sont les plus sensibles (Dandapat *et al.*, 2009). Il est donc probable que celle-ci impactent nos participants. Nous observons néanmoins dans le cas de l'alsacien que si les outils proposent la même étiquette pour un mot morphosyntaxique sur deux en moyenne, celle-ci est rejetée par les participants dans 12 % des cas.

catégorie `Auxiliaire` (AUX) :

La catégorie regroupe les verbes hànn / hàn (avoir), sìnn / sín / sìì / sii (être), wère / waere / werre / war(d)e / wùrre / wurra (devenir), tüen / duen / düe :

Exemples :
Sallamols ìsch nur z' Mainz druckt **worra**/AUX.
ìch **tüa**/AUX a Kugelhupf assa

ATTENTION : ces verbes peuvent prendre l'étiquette VERB lorsqu'ils ont un sens lexical plein :
S Gschaft vum Johannes Mentelin hàt schnall Erfolg **bikumma**/VERB, ar ìsch a riicher Mànn **worra**/VERB.
Charles , wenn du wüescht , wie ich dich gern **hab**/VERB !
's **tüat**/VERB m'r leid, ar **tüat**/VERB nix !

catégorie `Nom commun` (NOUN) :

Un nom peut être accompagné ou précédé d'un déterminant et occupe la plupart du temps une fonction sujet, objet ou complément :

Exemples :
sa **konser**/NOUN yer la mari ti top
mo'nn atan li kot **laboutik**/NOUN Zan

ATTENTION : Un nom peut aussi être un adjectif ! Aidez vous du contexte pour trouver la bonne catégorie :
enn **kabri**/NOUN
enn lavwa **kabri**/ADJ

catégorie `Préposition` (ADP) :

Exemples :
Aka/ADP manman
Vini'w **an**/ADP bra an mwen
En ké goumé **pou**/ADP péyi en mwen
i palé **pendan**/ADP conbyen tan
twa zè **avan**/ADP match-la fin

ATTENTION : Une préposition peut être collée à un pronom
I té ka palé ban mwen é an té ka palé **ba'y**/ADP+PRON
Profité **dè'y**/ADP+PRON !

FIGURE 5.5 – Extraits des guides d'annotation pour les catégories `Auxiliaire` (AUX) en alsacien, `Nom commun` (NOUN) en créole mauricien et `Préposition` (ADP) en créole guadeloupéen.

5.3 Produire du corpus annoté en parties du discours

5.3.1 Annotation en séquence

5.3.1.1 Implémentation

Notre première expérience d'annotation a consisté à proposer aux participants des *phrases* pré-annotées, les participants pouvant valider, invalider ou proposer une autre étiquette.

Une séquence d'annotation comprend quatre phrases, dont trois sont tirées aléatoirement de C_{Brut} et une provient de C_{Ref}. Cette dernière nous permet d'évaluer le participant à l'issue de chaque séquence annotée.

Selon les résultats produits par les outils de pré-annotation, l'annotation peut-être réalisée de deux manières :

- par attribution d'une étiquette à partir de la suggestion des deux propositions données par les *taggers* lorsqu'ils sont en désaccord (voir figure 5.6.),
- par validation ou rejet de l'étiquette proposée par les deux *taggers* lorsqu'ils sont d'accord (voir figure 5.7.). En cas de rejet de l'étiquette, les deux catégories ayant le meilleur score de confiance (défini dans la section 5.3.1.3) sont proposées au participant.

Nous laissons toujours la possibilité de corriger l'annotation suggérée ou de choisir une tierce catégorie.

FIGURE 5.6 – Annotation directe.

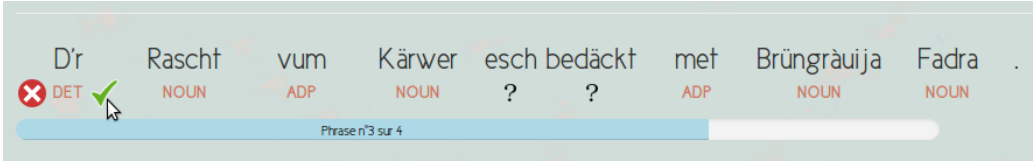


FIGURE 5.7 – Annotation par validation de l'étiquette suggérée.

5.3.1.2 Formation

La formation est réalisée sur quatre phrases de C_{Ref} à annoter complètement : cette phase est obligatoire avant de pouvoir entrer dans la phase effective de production d'annotations. Le

participant ne peut pas passer à la phrase suivante tant que toutes les étiquettes ne sont pas correctes. En cas d'erreur, un rappel reprenant les informations de l'aide-mémoire pour chaque catégorie erronée choisie est affiché, mais les étiquettes attendues ne sont pas divulguées. Cette première phase est destinée à familiariser le participant aux catégories existantes, ainsi qu'à le confronter aux difficultés de la tâche. Les annotations produites au cours de cette phase ne sont pas enregistrées.

5.3.1.3 Évaluation

La phase de production d'annotations est constituée d'une séquence de trois phrases à annoter issues du corpus brut, auxquelles s'ajoute une phrase issue du corpus de référence. Les $NbAnn_{Ref}$ annotations produites par un participant P sur cette phrase de référence permettent de calculer, à l'issue de chaque séquence, le score de confiance du participant :

$$Score_P = \frac{NbAnn_{Ref,Correctes}}{NbAnn_{Ref}}$$

Ce score est ainsi mis à jour régulièrement et reporté sur toute annotation produite par le participant P sur un mot morphosyntaxique M avec la catégorie C_i : $ScoreAnn_{M,P,C_i}$ vaut $Score_P$ au moment de l'annotation.

Nous utilisons ce score de confiance pour filtrer les annotations de mauvaise qualité et pour identifier l'étiquette la plus probable parmi les éventuelles annotations concurrentes réalisées sur un mot morphosyntaxique M par plusieurs participants. Nous déterminons ainsi pour chaque étiquette attribuée au mot morphosyntaxique M un score de confiance $Score_{M,C_i}$ correspondant à la moyenne des scores des annotations Ann_{M,P_j,C_i} produites par différents participants :

$$Score_{M,C_i} = \frac{\sum_j ScoreAnn_{M,P_j,C_i}}{\sum_{i,j} ScoreAnn_{M,P_j,C_i}}$$

Nous choisissons enfin l'étiquette unique la plus probable pour chaque mot morphosyntaxique :

$$C_M = \arg\max_i(Score_{M,C_i})$$

Le corpus ainsi annoté est ensuite utilisé pour entraîner des *taggers* qui seront utilisés à leur tour comme outils de pré-annotation, dès lors que leurs performances dépassent celles de l'outil précédent.

5.3.2 Annotation par étiquette

Ce modèle d'annotation est destiné à réduire la complexité de la tâche d'annotation en obligeant le participant à se focaliser sur une catégorie à la fois. Il n'a été implémenté que dans le cas de la plateforme `Recettes de Grammaire` pour l'alsacien.

En pratique, nous avons conçu cette tâche pour qu'elle soit réalisée à la suite de la tâche de production de corpus bruts présentée dans la section 5.4, selon le cheminement suivant :

1. Le participant ajoute un texte.

2. Le texte est pré-annoté par un outil état de l'art et le résultat de cette annotation est montré au participant.
3. Si le participant accepte de corriger ces pré-annotations, il est renvoyé vers l'interface d'annotation correspondante.

5.3.2.1 Implémentation

En cliquant sur une catégorie, le participant fait apparaître les pré-annotations correspondantes qu'il peut valider ou rejeter. Trois niveaux de difficulté ont été établis pour les catégories grâce à l'étude des annotations produites par les participants :

- niveau *facile*, ayant obtenu plus de 0,95 de F-mesure : dans le cas de l'alsacien, il s'agit des catégories NOUN et DET ;
- niveau *intermédiaire*, ayant obtenu une F-mesure comprise entre 0,90 et 0,95 : dans le cas de l'alsacien, il s'agit des catégories ADP+DET, NUM, INTJ, PROPN, VERB et SYM ;
- niveau *difficile*, contenant les catégories restantes nécessitant une formation : dans le cas de l'alsacien, il s'agit des catégories ADJ, ADV, ADP, AUX, CONJ, PRON, PART, SCONJ et X.

Ce découpage permet de diminuer la complexité de la tâche par rapport à une annotation séquentielle. Il permet aussi une meilleure intégration de la formation dans la plateforme : le « coût d'entrée » pour participer est amoindri par le fait que les formations se font au fur et à mesure que le participant progresse dans l'annotation.

La production d'annotation est structurée en trois étapes correspondant aux trois niveaux de difficulté définis ci-dessus :

1. Dans un premier temps, seules les catégories *faciles* sont visibles. Cette étape est destinée à mettre en confiance le participant quant à sa capacité à participer à l'amélioration des performances de l'outil entraîné.
2. Une fois les catégories *faciles* annotées, la liste complète des étiquettes apparaît à la droite du texte saisi (voir la figure 5.8). Les catégories blanches sont les catégories *intermédiaires* entre lesquelles le participant peut naviguer librement.
   Les catégories hachurées sont les catégories *difficiles* qui requièrent une formation.
   Les catégories grisées sont celles qui n'ont pas été utilisées dans l'annotation par l'outil de pré-annotation.
3. Une fois que toutes les étiquettes issues de la pré-annotation ont été examinées par le participant, il lui reste à annoter les mots morphosyntaxiques dont l'étiquette a été rejetée. C'est la phase d'annotation libre au cours de laquelle le participant doit choisir une catégorie parmi la liste complète.

Enfin, le participant peut naviguer entre les catégories, accéder au guide d'annotation pour chacune d'entre elles sous la forme d'un menu déroulant, et corriger ses annotations s'il le souhaite.

5.3.2.2 Formation

Dans ce modèle d'annotation, la formation est également réalisée *par étiquette* : elle consiste à présenter au participant une séquence de phrases faisant apparaître des occurrences de mots

FIGURE 5.8 – Extrait de l'interface d'annotation pour la catégorie VERB .

morphosyntaxiques appartenant à la catégorie en question ainsi que d'autres susceptibles de générer une confusion avec celle-ci. Les mots morphosyntaxiques sont pré-annotés avec la catégorie en cours et le participant doit identifier en validant ou rejetant l'étiquette proposée, les mots morphosyntaxiques appartenant ou non à la catégorie. Le participant doit valider toutes les pré-annotations correctes et rejeter toutes celles qui sont erronées pour pouvoir valider sa formation. La figure 5.9 illustre l'interface de formation.

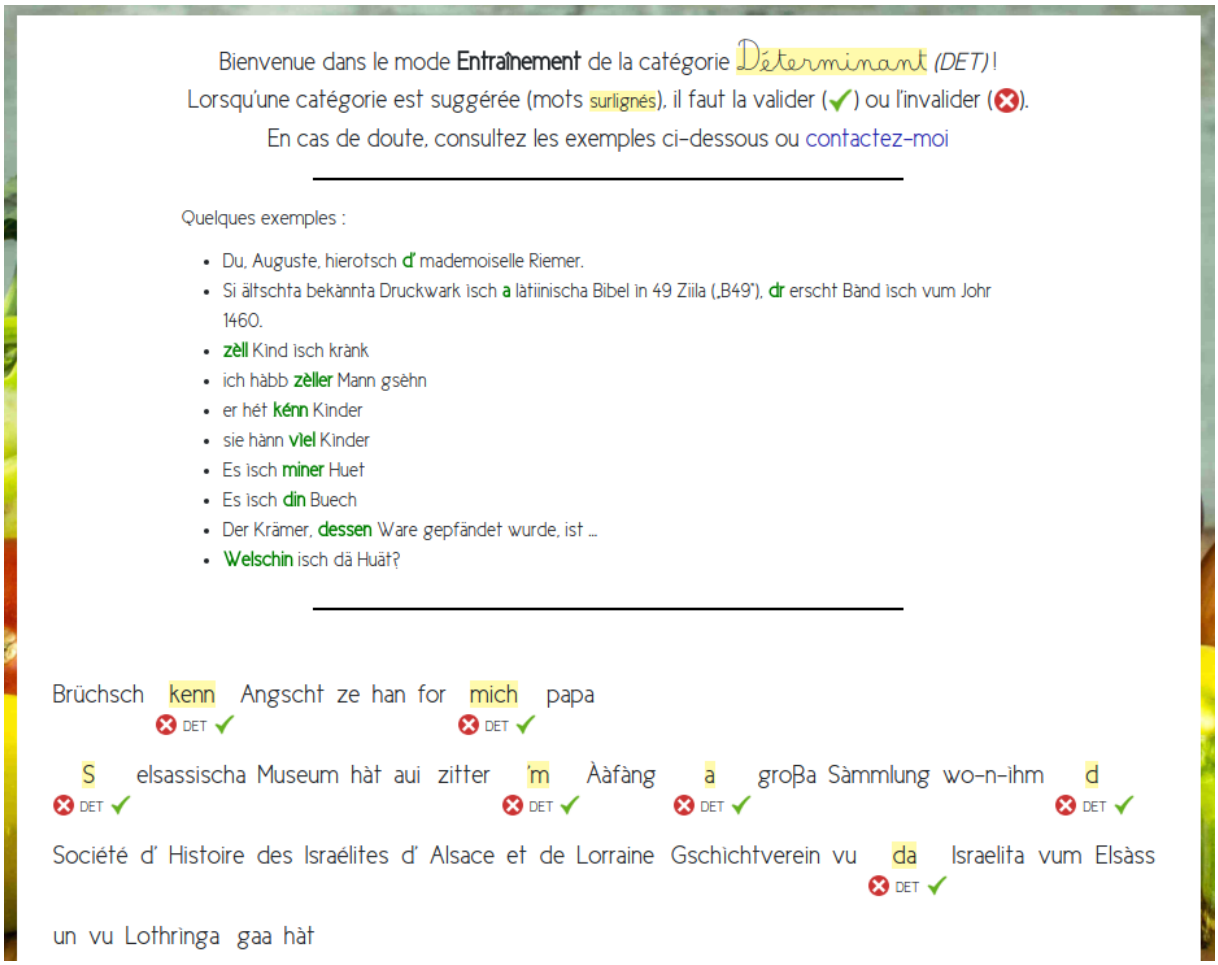


FIGURE 5.9 – Extrait de l'interface de formation pour la catégorie DET .

5.3.2.3 Évaluation

La conception de la tâche, où les textes des participants sont annotés à la volée (voir section 5.3.2) ne permet pas d'introduire des phrases de référence sur lesquelles évaluer les participants. Il nous a donc été impossible de reproduire la méthodologie, pourtant efficace, d'évaluation présentée dans la section 5.3.1.3. Néanmoins, nous pouvons évaluer les participants grâce à l'introduction, dans les pré-annotations à corriger, d'étiquettes volontairement erronées pour un jeu de mots morphosyntaxiques connus et non ambigus. Par exemple, la pré-annotation du mot « avec » *mìt*/ADV devra être corrigée en *mìt*/ADP. Les performances atteintes par le participant sur ces mots morphosyntaxiques nous permettent de définir un niveau de confiance semblable à celui

mis en place précédemment. Cependant, cette méthode sans doute moins performante devra faire l'objet d'une attention particulière pour être améliorée au besoin.

5.4 Produire du corpus textuel

L'évaluation et l'exploitation des ressources obtenues lors de nos premières expériences de myriadisation d'annotations, détaillées dans la partie 6.2.1, nous ont menée à conclure :

1. qu'il est indispensable d'avoir accès à suffisamment de corpus bruts, de qualité suffisante, pour assurer la qualité des annotations produites ;
2. que l'absence de prise en compte de la variation peut conduire à une stagnation voire à une dégradation des performances des outils qui en découlent.

Or, nous l'avons vu, accéder à des corpus en taille suffisante n'est pas une tâche aisée et c'est pour cette raison que nous avons décidé d'ajouter la tâche de production de corpus et de variantes pour compléter les tâches d'annotation présentées dans la section 5.3.

Notre conception de la myriadisation de corpus textuels suit donc à la fois :

- Un objectif de *cumul* : plus le nombre de participants est élevé, plus le nombre de textes produits et la taille du corpus correspondant sont importants.
- Un objectif de *diversité* : plus les locuteurs participant ont des pratiques variées, plus le corpus collecté pourra prétendre être représentatif des variétés scripturales et dialectales en usage.

5.4.1 Le choix des recettes de cuisine

Dans un premier temps, nous avons décidé de faire produire aux participants des recettes de cuisine. Ce choix nous a permis d'apporter une coloration culturelle à notre projet, et de faciliter la promotion du projet, la gastronomie étant un domaine plus accessible que la linguistique.

Le projet Recettes de Grammaire a ainsi été présenté lors d'une conférence de presse organisée conjointement avec l'OLCA en juin 2018. Ce lancement a donné lieu à un certain nombre de publications papier [128] et radiophonique [129].

Si ces efforts de communication ont permis une augmentation du nombre d'inscriptions, les recettes de cuisine se sont en revanche révélées être un genre peu adapté à la production de corpus. Nous détaillons ces observations dans la section 6.3.1.

128. Publication dans les Dernières Nouvelles d'Alsace, voir https://www.dna.fr/actualite/2018/06/11/audio-comment-automatiser-l-alsacien-a-travers-des-recettes-de-cuisines, dans 20 minutes, voir : https://www.20minutes.fr/strasbourg/2287547-20180612-alsace-site-internet-mele-wikipedia-marmiton-developper-langue-regionale, sur le site de l'OLCA, voir http://www.olcalsace.org/fr/actualite/des-recettes-et-des-lettres, dans Tchapp, voir : https://www.tchapp.alsace/articles/decouvertes/des-recettes-en-alsacien-pour-promouvoir-la-langue-r%C3%A9gionale.html, juin 2020.
129. Sur France Bleu Elsass, voir : https://www.francebleu.fr/emissions/billet-d-humeur/elsass/billet-d-humeur-76, juin 2020.

5.4.2 Production réelle et diversification des genres

Nous avons profité des enquêtes linguistiques présentées dans la section 4.1 pour demander aux participants quels types de textes ils écrivent habituellement. Les réponses obtenues à la question « Lorsque vous écrivez [alsacien|créole mauricien] (hors ligne ou en ligne), qu'écrivez-vous (plusieurs réponses sont possibles) ? » sont présentées dans le tableau 5.9. Ces réponses montrent que l'usage principal de l'écrit est conversationnel et nous incitent à diversifier les types de textes pouvant être produits sur la plateforme.

Commentaires à des publications	Alsacien	26 %
	Créole mauricien	36 %
Discussion sur les réseaux sociaux	Alsacien	25 %
	Créole mauricien	64 %
Lettres ou e-mails	Alsacien	16 %
	Créole mauricien	14 %
Blagues	Alsacien	14 %
	Créole mauricien	19 %
Contenu littéraire	Alsacien	4 %
	Créole mauricien	7 %
Opinions politiques	Alsacien	4 %
	Créole mauricien	9 %
Contenu informatif (par ex. Actualités, blogs)	Alsacien	3 %
	Créole mauricien	8 %
Recettes de cuisine	Alsacien	3 %
	Créole mauricien	4 %

Tableau 5.9 – Productions écrites habituelles en alsacien et créole mauricien.

Suite à ces observations, nous avons par conséquent ajouté sur nos plateformes la possibilité de contribuer grâce à :

- des phrases ou proverbes,
- des poèmes,
- des textes libres.

Nous avons ainsi souhaité libérer les participants de la contrainte formelle que constituaient les recettes de cuisine, mais nous cherchons aussi à pousser les participants à proposer des contenus courts.

5.5 « Moi, j'aurais dit ça comme ça ! » : collecter la diversité

Nous présentons dans cette section la collecte de lexiques alignés multi-variantes.

L'interface « Moi j'aurais dit ça comme ça ! » est accessible *via* un onglet pour tout contenu publié sur la plateforme. Elle permet à tout participant d'ajouter une variante graphique de tout mot présent dans un des textes de la plateforme. Cet ajout n'édite pas le corpus existant, toutes les versions restant accessibles à tous les participants. La liste des variantes ainsi obtenues nous renseigne sur les mécanismes de variation à l'œuvre. L'interface d'ajout en contexte est présentée dans la figure 5.10.

FIGURE 5.10 – Ajout de la variante *Kugelhof* pour le mot *Kugelhopf*.

Nous avons également implémenté cette fonctionnalité dans une seconde interface accessible *via* les nuages de mots générés automatiquement à partir des contenus et publiés sur la page d'accueil (voir par exemple la figure 5.11).

L'ajout de variante graphique se fait après avoir cliqué sur un mot. La figure 5.12 est un exemple de l'interface correspondante, qui donne les variantes existantes, les contextes d'apparitions, et le cas échéant les annotations proposées.

Chaque participant a en outre la possibilité dans son profil de renseigner son lieu d'apprentissage de l'alsacien, en le plaçant sur une carte de l'Alsace découpée en cinq aires dialectales (voir la figure 5.13) : nous utilisons cette information pour faire correspondre les variantes ajoutées aux variétés d'alsacien identifiées.

5.6 Conclusion

Les tâches de myriadisation que nous avons proposées recouvrent deux besoins : un besoin en ressources et un besoin en annotateurs humains pouvant enrichir ces ressources. L'état de l'art pour la myriadisation de ressources pour le TAL pour des langues non standardisées étant quasi inexistant, nous avons tenté plusieurs approches permettant de produire trois ressources différentes : des corpus textuels bruts, des corpus annotées en morphosyntaxe et des lexiques de variantes dialectales.

Figure 5.11 – Nuage de mot généré automatiquement.

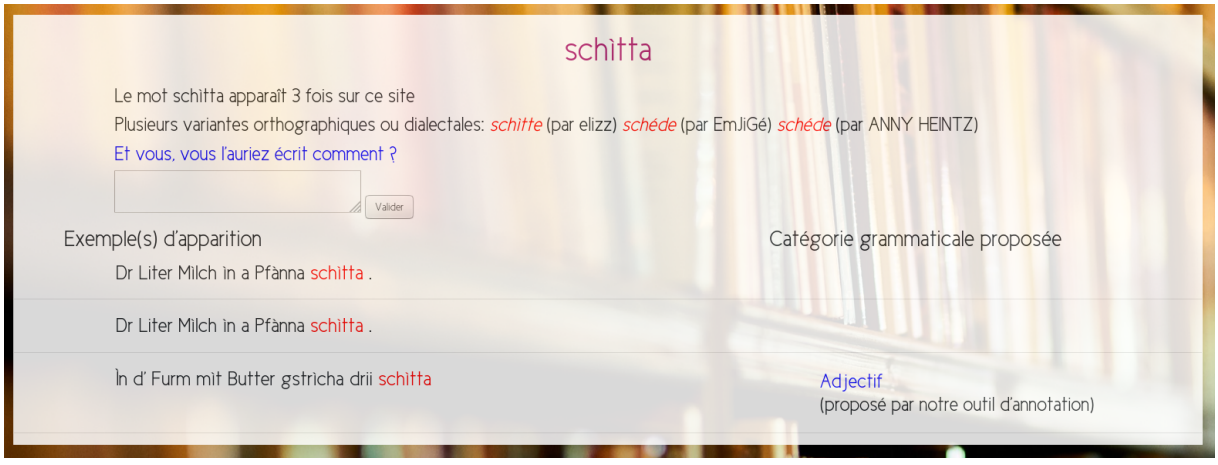


Figure 5.12 – Interface d'ajout de variantes pour le mot *schìtta*.

FIGURE 5.13 – Profil privé du participant.

Bien qu'assez simples, les tâches permettant de myriadiser ces ressources requièrent l'existence de ressources minimales préalables. Nous avons donné trois exemples de constructions de telles ressources. Le plus gros obstacle que nous n'avions pas anticipé et auquel nous nous sommes confrontée a été la trop faible quantité de corpus brut disponible.

Si ce manque de corpus pourrait laisser croire qu'il n'y a pas d'enjeu de traitement pour ces langues, ce serait oublier que nous avons choisi de travailler dans le cadre de cette recherche à la construction de corpus pérennes. Or, la majorité des contenus numériques produits par les locuteurs ne rentrent pas dans ce cadre. Les résultats des enquêtes menées, sans qu'elles nous permettent de généraliser sur les pratiques de tous les locuteurs, vont néanmoins dans ce sens : dans le cas de l'alsacien et du créole mauricien, la pratique de l'écrit en ligne concerne principalement le cadre conversationnel. Les corpus pouvant être composés de telles productions linguistiques posent des difficultés éthiques et légales quant à leur utilisation et redistribution.

Conclusion

Au cours de notre travail, les périodes de développement, de promotion des plateformes et de collecte des retours auprès des participants se sont succédé. Le dialogue ainsi amorcé avec les locuteurs nous a permis d'ajuster notre démarche au fur et à mesure que nous comprenions d'une part les difficultés que posait notre entreprise et d'autre part les attentes des différentes communautés linguistiques.

La première plateforme, P_Ann, a été conçue en supposant que le corpus brut ne constituait pas une ressource limitante. En réalité, nos expériences dans le cas de langues non standardisées ont montré que le manque de corpus brut mettait à mal l'efficacité de notre démarche. Ce manque est d'autant plus problématique que la production linguistique est variée et que les mécanismes de variation sont mal connus et ne permettent pas, par exemple, d'identifier les variantes automatiquement.

En effet, la présence de variation complique d'une part la myriadisation d'un point de vue linguistique : la variation observée à l'écrit provient d'une variation des pratiques avec lesquelles les locuteurs peuvent être plus ou moins à l'aise. D'autre part, comme il sera détaillé dans le chapitre 7 consacré à l'évaluation extrinsèque des ressources annotées myriadisées, la variation impacte le développement des outils.

La présence de variation nous conduit à diversifier les sources de corpus pour tâcher de remédier aux problématiques de représentativité qu'elle entraîne. Il nous est en effet apparu que pour atteindre l'objectif pratique de notre travail, à savoir le développement d'outils fonctionnels en conditions réelles, nous devions dépasser la perspective « opportuniste » (McEnery et Hardie, 2011) dans laquelle nous nous trouvions et organiser la collecte de corpus de manière à placer les locuteurs au cœur de celle-ci.

Ayant choisi de mener nos développements en conservant le maximum de liberté quant aux fonctionnalités, nous avons pu élargir nos objectifs en ouvrant nos plateformes à la production de ressources brutes. Le développement de P_Prod_Var a ainsi été l'occasion de répondre concrètement aux remarques et critiques formulées par les participants.

En conclusion, les expériences que nous avions entamées en considérant que la myriadisation permettrait à une somme de locuteurs de se substituer à un expert ont révélé que les locuteurs étaient détenteurs de connaissances linguistiques complémentaires et particulièrement précieuses dans un contexte linguistique varié. Il nous semble qu'une telle connaissance des pratiques linguistiques ne saurait en effet être détenue par un seul expert et qu'un tel contexte donne tout son sens à la myriadisation des ressources.

Troisième partie

Ressources myriadisées : évaluation et exploitation

Introduction

Nous consacrons la troisième partie de ce document à l'évaluation de la méthodologie de myriadisation de ressources pour le traitement automatique de langues non standardisées que nous avons introduite dans la partie II.

Outre la quantité de données produites ou le nombre de contributeurs impliqués, le succès d'une entreprise de myriadisation se traduit comme la faculté, ou non, de répondre au problème initial. Dans notre cas, évaluer notre méthodologie revient à répondre à la question suivante : « la myriadisation a-t-elle permis de produire des ressources linguistiques utiles pour le traitement automatique des langues étudiées ? ».

Comme présenté dans la partie II, les plateformes de myriadisation développées sont le fruit de nombreux choix d'implémentation. Leur succès repose notamment sur la capacité de ses auteurs à proposer des plateformes séduisantes aux yeux des participants potentiels et à fournir un effort de publicité suffisant. Les compétences de développement, de conception d'interfaces Web (*webdesign*) et de communication participent donc du succès ou de l'insuccès d'une telle entreprise. Les choix de développement ayant été présentés dans le chapitre 5, nous détaillons dans une première section (section 6.1) les stratégies de communication mises en place pour construire une communauté de participants. Nous présentons ensuite l'évaluation des ressources en tant que telle.

Les trois types de ressources que nous avons myriadisées sont : des corpus bruts (pour l'alsacien et le créole mauricien), des corpus annotés (pour l'alsacien et le créole guadeloupéen), et des lexiques de variantes graphiques (pour l'alsacien et le créole mauricien).

Plusieurs mesures d'évaluation peuvent être envisagées pour l'évaluation intrinsèque de ces ressources, comme leur taille, leur variété, leur représentativité, ou leur exactitude dans le cas des corpus annotés. Outre ces paramètres, nous devons nous assurer que les ressources produites permettent effectivement de progresser vers l'objectif fixé. En l'occurrence, la tâche linguistique à laquelle nous nous sommes attelée étant l'annotation en parties du discours, l'utilisation des ressources myriadisées pour l'entraînement de nouveaux outils d'annotation permet d'en fournir une évaluation extrinsèque. L'effort principal ayant porté sur la myriadisation de ressources pour l'alsacien, nous concentrons l'évaluation extrinsèque sur celles-ci.

Nous proposons par conséquent dans un premier chapitre (chapitre 6) une évaluation quantitative et qualitative intrinsèque lorsque celle-ci est possible. Concernant l'annotation en parties du discours, nous étant placée dans des contextes linguistiques dans lesquels les linguistes sont peu nombreux, nous n'avons pas pu valider manuellement l'ensemble des annotations produites par les participants. Nous présentons en revanche leurs performances sur des corpus de référence à partir desquelles nous inférons une estimation de la qualité des corpus myriadisés.

Les deux chapitres suivants (chapitres 7 et 8) sont consacrés à l'évaluation d'outils d'annotation en parties du discours entraînés pour l'alsacien. Ces expérimentations permettent de mesurer l'intérêt des ressources myriadisées pour le traitement automatique de l'alsacien. Nous les utilisons également pour mettre en lumière les difficultés inhérentes à la présence de phénomènes de variation.

Le chapitre 7 présente des expérimentations d'annotation supervisées classiques, et illustre en ce sens notre volonté première de reproduire sur des langues peu dotées ce qui fonctionne sur des langues mieux dotées et d'utiliser, en quelque sorte, la myriadisation pour « rattraper le retard » accusé par ces langues pâtissant du manque de ressources.

Le chapitre 8 représente quant à lui un pas de côté par rapport à l'utilisation habituelle de la myriadisation pour le TAL : nous y présentons une nouvelle nouvelle méthodologie tirant parti d'une ressource conçue pour compenser la variation observée dans les corpus sans requérir de ressource textuelle brute de grande taille.

Chapitre 6

Ressources myriadisées

Sommaire

Un élément clé de la réussite d'une campagne de myriadisation est la capacité d'une part à identifier la « bonne » communauté de locuteurs, et d'autre part à rentrer en contact avec elle, afin de la motiver à participer (voir (Cosquer *et al.*, 2012) pour plus de détails sur les communautés de participants aux sciences participatives). Nous présentons donc dans une première section les stratégies mises en place pour constituer une communauté de locuteurs participants grâce aux deux premières instances de P_Ann.

La seconde section de ce chapitre est consacrée à l'évaluation intrinsèque des corpus annotés en parties du discours myriadisés sur les instances Bisame et Krik !. Nous y présentons notamment comment nous avons évalué ces corpus à partir de l'évaluation des annotations produites par les participants sur un corpus de référence.

Les données myriadisées sur les instances de P_Prod_Var, Recette de grammaire et Ayo !, à savoir les corpus bruts et les graphies alternatives produites *via* la fonctionnalité « Moi, j'aurais dit ça comme ça ! », sont présentées dans une troisième section.

Les résultats concernant l'alsacien mentionnés dans ce chapitre ont été présentés dans un article publié à TALN 2017 (Millour *et al.*, 2017). Sauf mention contraire, nous avons mis à jour les chiffres de participation correspondant à la situation de juin 2020.

6.1 Construction d'une communauté de participants

6.1.1 La communauté alsacienne au rendez-vous

La première langue à laquelle nous nous sommes intéressée et celle sur laquelle nous avons le plus travaillé étant l'alsacien, c'est la communauté linguistique que nous sommes le mieux parvenue à mobiliser.

Nous fondions, en début de projet, beaucoup d'espoir sur le rayonnement de structures et de médias officiels tels que l'Office pour la Langue et la Culture d'Alsace (OLCA)[130] ou la radio régionale France Bleu Elsass[131] qui diffuse des contenus en alsacien.

Si l'une des principales contributrices de nos plateformes a eu connaissance de notre projet *via* une émission diffusée sur France Bleu Elsass, ces organismes se sont révélés globalement décevants en termes de rayonnement. Nos prises de contact nous ont en effet amenée à penser que les locuteurs de l'alsacien présents sur le Web sont répartis dans des micro-communautés qu'il n'est pas possible d'atteindre par un canal de communication unique. Nous nous sommes donc tournée vers différents organismes tels que le Fonds international pour la langue alsacienne (FILAL)[132] ou l'entreprise Marque Alsace[133]. Nous avons également mené une campagne de recrutement sur Facebook en identifiant les utilisateurs déclarant parler alsacien dans des groupes tels que le « Centre Culturel Alsacien / Elsässisches Kulturzentrum » ou « Alsace Bilingue ». Il est également probable que le cumul des diffusions ait contribué à intéresser certains des participants à nos plateformes.

En définitive, 248 personnes ont créé un compte sur une des deux plateformes consacrées à l'alsacien, 208 s'étant inscrits sur `Bisame`, 48 sur `Recettes de Grammaire` : le graphique 6.1 montre l'évolution des inscriptions mensuelles.

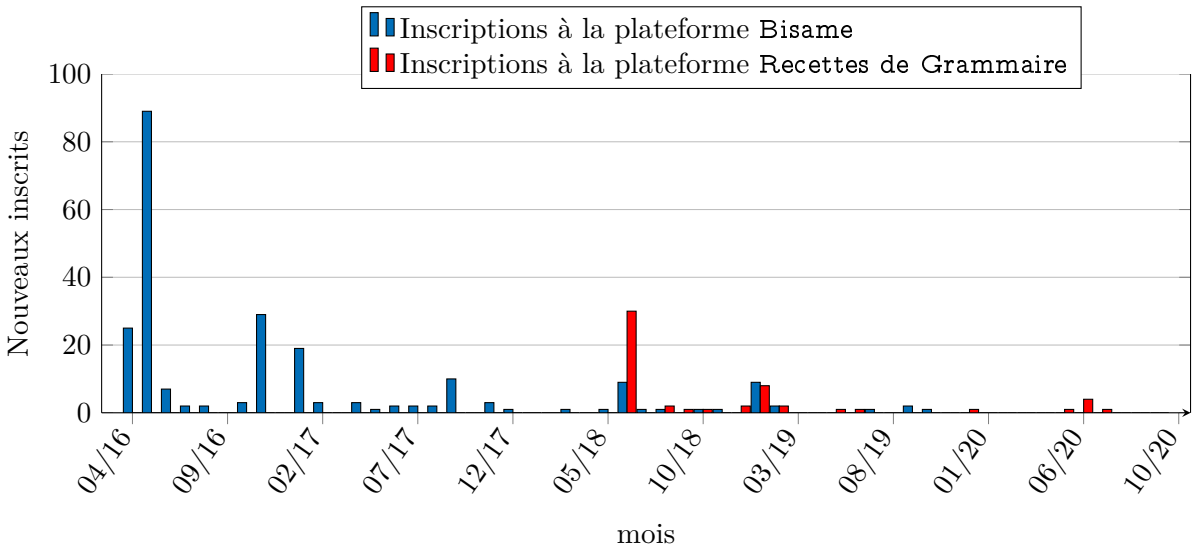


Figure 6.1 – Évolution des inscriptions aux plateformes instanciées pour l'alsacien.

130. Voir : https://www.olcalsace.org/, juin 2020.
131. Voir : https://www.francebleu.fr/elsass.
132. Voir : https://filalsace.net/.
133. Voir : http://www.marque-alsace.fr/.

Entre 2016 et 2018, 42 participants avaient contribué en produisant des annotations sur Bisame. Ils se répartissaient par intervalle de nombre d'annotations produites comme illustré dans le tableau 6.1 [134], seuls neuf d'entre eux ayant produit plus de 250 annotations. Ce premier constat confirme une tendance bien connue et déjà décrite, notamment par Chamberlain *et al.* (2013) : peu de personnes participent (et produisent) beaucoup. Notons par ailleurs qu'à deux exceptions près, les participants ne sont revenus sur la plateforme qu'après des relances par mail.

Nb. annotations	<50	[50-250]	[250-650]	>650
Nb. participants	13	20	5	4 (852, 1 178, 3 822 et 4 202)

Tableau 6.1 – Répartition des participants par intervalle de nombre d'annotations produites.

En moyenne, les dix participants ayant produit le plus grand nombre d'annotations se sont connectés quatre jours et ont annoté une quinzaine de séquences de quatre phrases depuis la mise en ligne de la plateforme. Afin de ne pas fausser ces moyennes, le participant le plus productif (96 séquences annotées pour neuf jours de connexion) a été exclu de ces observations.

Ces observations montrent que si nous pensons avoir identifié la bonne communauté de locuteurs, cette application n'est pas une plateforme autonome, c'est-à-dire ne nécessitant pas une publicité constante. Si les applications de myriadisation ne se passent pas de publicité, ou tout du moins de relances automatiques, nous pensons qu'améliorer l'attractivité et la dimension ludique de la plateforme permettrait d'influer positivement sur l'implication des participants.

Nous nous confrontons ici à la difficulté à susciter chez les participants la « volition » (Fenouillet *et al.*, 2009), c'est-à-dire l'envie de revenir nécessaire à leur « rétention » sur la plateforme et permettant de faire vivre la ressource.

La même observation a été faite dans le cadre extrême d'une mission humanitaire visant à traduire des SMS pour aider les rescapés du tremblement de terre à Haïti en 2010 (Munro, 2013) et pour laquelle la rétention des participants sur la plateforme de myriadisation n'a pas dépassé quelques semaines. Cela rejoint les conclusions de l'enquête concernant ZombiLingo, une plateforme de myriadisation ludique : si certains participent pour « aider les scientifiques », ils ne restent pas. Ceux qui reviennent et participent le plus le font pour le jeu (Fort *et al.*, 2017).

Or, la plateforme Bisame n'est pas un jeu et ne propose à l'origine qu'une seule fonctionnalité ludique : un classement par points égal au nombre d'annotations multiplié par le score de confiance du participant. Nous pensons que l'ajout de ce classement a favorisé l'investissement de certains participants et les éléments d'analyse ci-dessus nous ont amenée à pousser le développement d'éléments ludiques au cours de nos expérimentations afin de favoriser la rétention des participants.

6.1.2 Une participation décevante sur Krik !

La plateforme Krik !, instance développée pour le créole guadeloupéen, n'a pas connu le même succès, seuls 35 comptes ayant été créés. L'écart entre les deux plateformes s'explique à notre avis par la différence d'énergie déployée à communiquer sur chacune des instances : la plateforme Krik ! n'a pas bénéficié de l'effort de communication décrit dans le paragraphe ci-dessus, nos

134. Ces intervalles ont été choisis de manière empirique.

contacts étant moindres et l'expérimentation n'ayant pas été poursuivie du fait de la fin du mémoire de l'étudiante.

En effet, le corpus que nous avions à disposition pour le créole guadeloupéen rendait l'annotation trop difficile, voire impossible, notamment parce que le jeu d'étiquettes présenté dans la section 5.2.2.2 n'était pas adapté à l'annotation de l'oral. Comme détaillé dans la section 5.2.3, le corpus du créole guadeloupéen est en effet constitué en majorité de transcriptions et découpé en groupes de souffle ce qui produit des séquences très difficiles à annoter en l'état. Certaines des séquences étaient même rendues inintelligibles du fait de la présence de nombreux achoppements et de structures syntaxiques incomplètes.

N'ayant pas à notre disposition d'autres corpus libres de droits pour le créole guadeloupéen, et ayant amorcé le développement de P_PROD_VAR, incluant la myriadisation de corpus bruts (voir la section 5.4), nous avons donc rapidement mis cette instance en pause.

L'adaptation de `Bisame` a une autre langue a donc été possible d'un point de vue technique, mais la reproductibilité de la méthodologie n'a pas pu être validée, les ressources initiales requises, en l'occurrence la disponibilité d'un corpus brut, présentant déjà un obstacle dans le cas du créole guadeloupéen.

6.1.3 Conclusions

Nous avons observé, et cela est valable pour les deux plateformes d'annotation `Bisame` et `Krik !`, qu'environ 40 % des participants ne produisent aucune annotation après avoir finalisé la phase de formation. Nous faisons l'hypothèse que la durée de la formation, de huit minutes en moyenne d'après les statistiques extraites de la base de données des utilisateurs, ainsi que la nature de la tâche, pouvant être perçue comme difficile et rébarbative, sont la cause de cette démotivation.

Les 27 répondants de l'enquête *L'alsacien, Internet et vous* ayant auparavant contribué sur `Bisame` ont répondu à la question `Vous avez participé, et cela vous a paru…` à 41 % par `facile`, à 37 % `difficile`, à 19 % `amusante`, et à 15 % `ennuyeuse`. Ces résultats ne nous permettent pas de conclure sur l'hypothèse formulée.

En revanche, et bien que nos plateformes soient conçues pour être utilisables sur téléphone mobile, l'inconfort d'utilisation a été évoqué par plusieurs participants comme un facteur de découragement. Par ailleurs, certains participants se sont plaints de ne pas pouvoir contribuer sur une version de l'alsacien plus proche de leur habitude. C'est ce qui a conduit un des participants à nous proposer un de ses textes pour alimenter la plateforme (*E Hochzit in de 50er Johre*, Raymond W.). Ce texte, écrit dans la variante strasbourgeoise, nous a permis de mener les expériences d'analyse par variante de la section 7.1.4.

En amont des enquêtes introduites dans la section 4.1, nous avions réalisé une enquête auprès des participants de la plateforme `Bisame`[135] afin de recueillir un premier aperçu de leurs genres, âges, niveaux d'études, et langues maternelles.

Sur les 22 participants ayant répondu à l'enquête, 77 % sont des hommes, ce qui va à l'encontre des observations de Chamberlain *et al.* (2013), qui montrent que les femmes sont davantage enclines à participer à ce genre d'interface ludifiée. Par ailleurs, près de 30 % des répondants ont

135. Nous ne présentons pas les résultats de l'enquête menée pour la plateforme `Krik !`, celle-ci ayant reçu trop peu de réponses pour être exploitable.

pour langue maternelle le français et non l'alsacien et 36 % déclarent avoir au-delà de 60 ans, une majorité ayant entre 21 et 40 ans. Enfin, notons que les participants ont un niveau d'études élevé, 60 % d'entre eux ayant atteint au moins le niveau BAC + 4, ce qui participe sans doute à expliquer la bonne qualité des annotations obtenues présentée dans la section suivante.

6.2 Évaluer les participants pour estimer la qualité du corpus

La qualité des corpus produits a été évaluée *via* l'évaluation des participants : nous extrapolons la qualité des annotations produites sur le corpus brut grâce à l'évaluation des annotations produites sur des phrases du corpus de référence.

Le tableau 6.2 recense les statistiques de participation observées sur les deux plateformes de production d'annotation. La participation a été bien plus importante dans le cas de l'alsacien, avec 60 participants dans le cas de `Bisame` et 11 dans le cas de `Krik!`. Ce différentiel se traduit par un nombre d'annotations et une qualité inférieure dans le cas du créole guadeloupéen. Nous détaillons l'analyse de ces résultats dans les sections ci-dessous.

	`Bisame` (gsw)	`Krik!` (gcf)
Nombre d'inscrits	208	35
Participants ayant finalisé la phase d'entraînement	89	17
Participants ayant produit des annotations	60	11
Jours d'annotation	119	9
Nombre d'annotations produites	26 377	1 205
Taille du corpus annoté (mots)	9 611	933
Qualité des annotations produites (F-mesure)	0,93	0,87

TABLEAU 6.2 – Participation sur les deux plateformes.

6.2.1 Des annotations de qualité

Annotations produites sur `Bisame` :

La figure 6.2 montre l'évolution du nombre d'annotations mensuelles sur la plateforme `Bisame` entre 2016 et 2020.

Nous présentons ici l'analyse menée en 2019, lorsque 26 086 annotations avaient été produites [136].

Sur les 26 086 annotations, 7 750 ont été réalisées sur $CRef_{gsw}$ (qui comprend 1 468 mots morphosyntaxiques) afin de pouvoir proposer une évaluation dynamique des participants. Le nombre important de ces annotations s'explique par la redondance des phrases servant à l'évaluation :

136. Nous menons une analyse similaire à celle menée avec les données de 2017 dans l'article (Millour *et al.*, 2017) à ceci près qu'en 2017, nous travaillions avec le jeu d'étiquettes qui était celui utilisé par D. Bernhard à l'époque, c'est-à-dire le jeu d'étiquettes « principal » (« *core part-of-speech categories* ») défini par *Universal Dependencies*. Par rapport aux expériences que nous présentons dans ce chapitre, l'étiquette ADP+DET n'avait pas encore été ajoutée au jeu d'étiquettes.

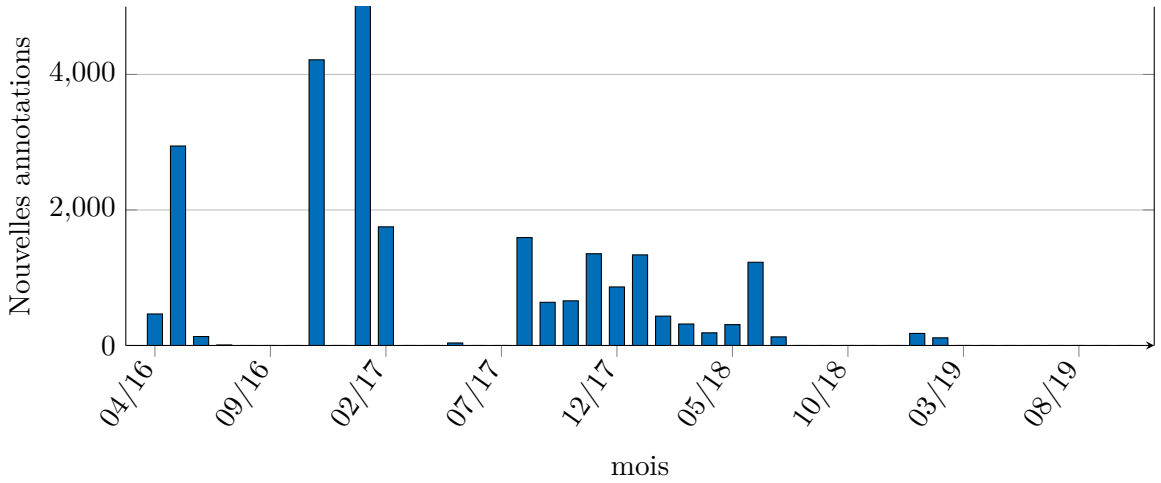


FIGURE 6.2 – Évolution du nombre d'annotations par mois sur la plateforme `Bisame`.

selon la méthodologie décrite dans la section 5.3.1, une phrase de $CRef_{gsw}$ étant annotée lors de chaque séquence de quatre phrases.

Les 8 096 annotations restantes ont permis d'annoter 9 547 mots, soit 440 phrases de $CBrut_{gsw}$. Ce corpus a été utilisé pour entraîner l'étiqueteur `MElt` (Denis et Sagot, 2012). Nous présentons les résultats de ces entraînements dans la section 7.1.

Nous avons évalué la qualité du corpus annoté en calculant, par catégorie, la F-mesure des annotations produites par les participants par rapport à la référence. Les résultats ainsi obtenus sont présentés en figure 6.3. La F-mesure moyenne arithmétique calculée en janvier est de $0,85$, et la moyenne pondérée par les effectifs des F-mesures par catégorie atteint $0,93$. En effet, les trois catégories PART, SYM et INTJ, pour lesquelles les résultats sont inférieurs à $0,5$, représentent chacune moins de 1 % du corpus et sont par conséquent peu annotées.

Par ailleurs, nous observons que la qualité de l'annotation augmente avec le nombre de participations, confirmant les résultats obtenus par Guillaume *et al.* (2016) : le nombre d'annotations ayant doublé, nous avons constaté entre juin 2016 et janvier 2017 un gain sur la F-mesure moyenne pondérée de plus de 40 %.

Concernant les erreurs produites par les participants, 25 % concernent la catégorie ADV, confondue dans un tiers des cas avec la catégorie ADJ. Les erreurs concernant la catégorie VERB (19 % du total) sont à 75 % dues à la confusion avec la catégorie AUX. Ces catégories sont des catégories identifiées comme difficiles, étant également source de confusion pour les expertes de l'annotation nous ayant fourni le corpus annoté de référence pour l'alsacien $CRef_{gsw}$ (voir 5.2.4). Par ailleurs, la catégorie X a entraîné une confusion entre le cas de l'alternance codique, et le cas des emprunts. Ce type d'erreur peut être corrigé en améliorant la documentation fournie aux participants.

En définitive, nous estimons que les annotations produites par les participants au cours de cette première expérience d'annotation en parties du discours sur une langue peu dotée et non standardisée sont de qualité satisfaisante.

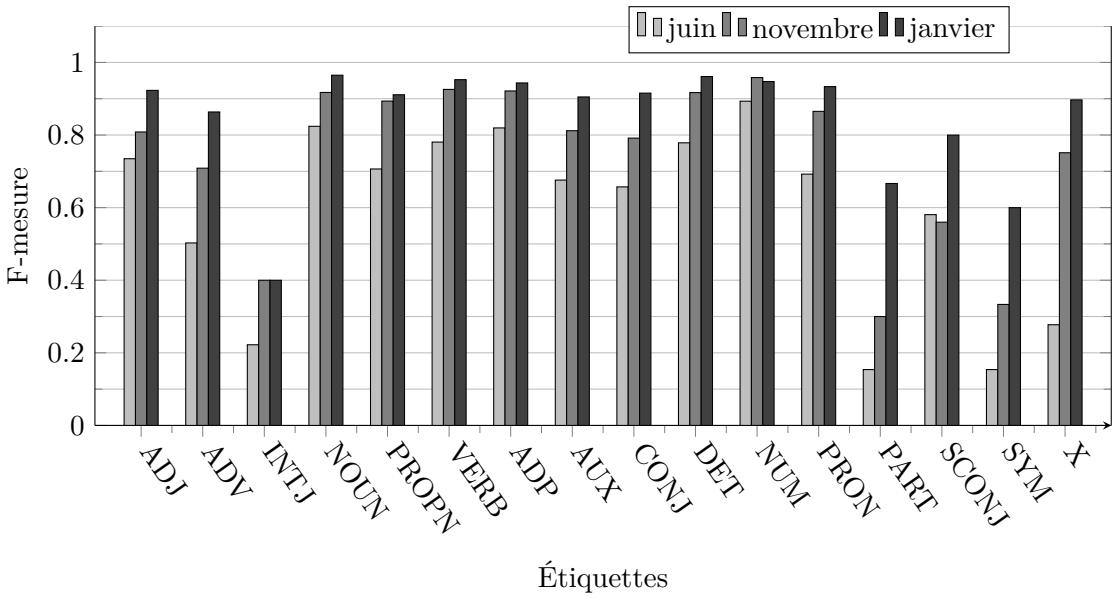


FIGURE 6.3 – Comparaison par étiquette de la F-mesure des annotations produites par les participants sur C_{Ref} en juin 2016 (3 436 annotations), novembre 2016 (5 888 annotations) et janvier 2017 (7 750 annotations).

Annotations produites sur `Krik!` :

Dans le cas de la plateforme `Krik!`, les difficultés liées à la nature du corpus se ressentent dans la qualité des annotations produites (voir le tableau 6.2) : celles-ci atteignent une exactitude de 87 %, bien en deçà de ce que nous observons sur `Bisame` (93 %). Notons que Hovy *et al.* (2014) obtiennent dans leur expérience d'annotation par travail parcellisé (de *tweets* en anglais) une exactitude d'environ 80 %, inférieure, donc, aux résultats que nous obtenons sur nos plateformes.

Pour comprendre la source des erreurs commises par les participants sur la plateforme `Krik!`, nous avons corrigé manuellement le corpus myriadisé. Outre les difficultés liées à la nature du corpus, cette analyse nous a permis de révéler des limitations de notre tokéniseur et de notre guide d'annotation dues à l'apparition dans le corpus d'habitudes scripturales qui n'étaient pas prises en compte. Par exemple, la forme séparée `anba la` (« en dessous ») génère deux *tokens*, qui, lorsqu'ils ne sont pas suivis d'un nom commun ne peuvent pas être annotés séparément. Nous n'avions pas documenté la bonne manière d'annoter ces *tokens*.

Nous avons par ailleurs analysé les erreurs commises par les participants de manière à en identifier les motifs récurrents, par exemple le cas de *té*, pouvant désigner le verbe « être » ou la particule désignant le passé en créole guadeloupéen.

Les analyses des erreurs dans les deux contextes linguistiques nous ont permis de mettre en évidence les cas les plus intrinsèquement ambigus. Ceux-ci requièrent une vigilance particulière : par la suite, nous les avons intégrés à la phase de formation, et ils ont fait l'objet d'explications spécifiques dans le guide d'annotation mis à disposition.

6.2.2 Des corpus annotés librement disponibles

Nous distribuons les deux corpus issus des expériences décrites précédemment sous licence CC BY-NC-SA. Dans le cas de l'alsacien, il s'agit du corpus myriadisé. Dans le cas du créole guadeloupéen, nous avons intégré le corpus myriadisé corrigé (933 mots) au corpus de référence (décrit dans la section 5.2.3, 1 024 mots) et nous distribuons l'ensemble.

- Corpus myriadisé de l'alsacien : 8 072 mots annotés par les participants de `Bisame`, distribué sur Ortolang [137] ;
- Corpus de référence du créole guadeloupéen : 1 957 mots, distribué sur Ortolang [138].

6.3 Myriadisation de ressources variées avec `Recettes de Grammaire` et `Ayo !`

Nous présentons dans cette section les ressources myriadisées grâce aux deux instances de P__Prod__Var : `Recettes de Grammaire` dans le cas de l'alsacien et `Ayo !` dans le cas du créole mauricien.

Le tableau 6.3 récapitule les ressources produites sur ces deux plateformes.

	`Recettes de Grammaire`	`Ayo !`
Langue	alsacien	créole mauricien
Participants	55 inscrits	17 inscrits
Textes saisis	• 10 recettes • 2 poèmes • 7 proverbes • 6 textes libres	• 11 recettes • 7 poèmes • 13 proverbes • 5 textes libres
Taille corpus	1 803 mots	1 903 mots
Annotations	199 annotations	1 050 annotations
Graphies alternatives	215 (148 mots)	46 (38 mots)

Tableau 6.3 – Ressources myriadisées sur les plateformes `Recettes de Grammaire` et `Ayo !`.

Sur `Recettes de Grammaire`, 7 inscrits ont ajouté un ensemble de 25 textes totalisant 1 803 mots, 11 inscrits ont produit 215 graphies alternatives (148 *via* les pages et 67 *via* le nuage de mots, pour un total de 106 mots) et 12 participants ont produit 200 annotations.

Sur `Ayo !`, 12 inscrits ont ajouté un total de 36 textes totalisant 1 903 mots, 3 inscrits ont produit 46 graphies alternatives (35 mots) et 7 participants ont produit 1 050 annotations.

6.3.1 Myriadisation de corpus bruts et annotés

La myriadisation de ressources textuelles brutes et annotées s'est révélée décevante sur `Recettes de Grammaire`.

137. Voir : https://repository.ortolang.fr/api/content/bisame_gsw/head/.
138. Voir : https://repository.ortolang.fr/api/content/krik_gcf/head/.

Le corpus brut collecté pour l'alsacien contient 1 803 mots. Dans un premier temps, la plateforme ne permettait de myriadiser que des recettes de cuisine. L'ajout de poèmes, proverbes ou phrases et textes libres, testé sur Ayo !, n'a été reporté sur Recettes de Grammaire qu'au printemps 2020. L'ajout de nouveaux textes sans qu'une communication ait été faite nous permet d'espérer pouvoir recueillir un corpus plus important grâce à ces nouvelles fonctionnalités.

Par ailleurs, le nombre d'annotations effectuées sur la plateforme (199) est bien en-deçà de ce qui a été produit sur Bisame (24 845). Cela s'explique peut-être par le fait que la fonctionnalité d'annotation apparaît de manière moins évidente que sur les instances de P_ANN. Nous n'avons pas recueilli de retours de participants à ce sujet.

Le corpus produit pour le créole mauricien sur Ayo est un peu plus important, avec 1 903 mots, de même que le nombre d'annotations produites (1 050).

Nous n'avons pas encore exploité ces corpus mais les distribuons sous licence CC BY-NC-SA. Les corpus bruts et annotés sont distribués sur Ortolang [139].

6.3.2 Myriadisation de graphies alternatives

La plateforme P_PROD_VAR permet aux participants de proposer une orthographe alternative pour un seul mot ou bien pour une séquence de mots. Cette deuxième option facilite la tâche des participants, mais conduit parfois à des séquences alternatives dont le nombre de *tokens* diffère de la version originale, et ne peut donc pas être immédiatement aligné. Dans de tels cas et lorsque cela était possible,les graphies alternatives ont été manuellement alignées sur la version originale.

Par exemple, la séquence de 14 *tokens* en alsacien :

> Ma kààt's eifàch a so assa, noch wàrm (zum Bispìl mìt Äpfelmüas) (V_2)

a été proposée comme alternative à la séquence de 13 *tokens* :

> Ma kààt's eifàch aso assa, noch wàrm (zum Bispìl mìt Äpfelmüas) (V_1) (« On peut servir comme ça, encore chaud (par exemple avec de la compote de pommes) »)

Cela produit l'alignement présenté dans le tableau 6.4 et la création de 4 paires de graphies alternatives.

Les tableaux 6.5 et 6.6 donnent des exemples de graphies collectées pour l'alsacien et le créole mauricien respectivement. Les listes complètes sont publiées dans les annexes C.1 et C.2.

139. Voir : https://repository.ortolang.fr/api/content/bisame_gsw/head/ pour l'alsacien et https://repository.ortolang.fr/api/content/ayo_mfe/head/ pour le créole mauricien.

Paire		1		2						3		4
V_1	Ma	kàt's	eifàch	a	so	assa,	noch	wàrm	zum	Beispiel	mìt	Äpfelmüeß
V_2	Ma	kààt's	eifàch	aso		assa,	noch	wàrm	zum	Bispìl	mìt	Äpfelmüas

TABLEAU 6.4 – Alignement de séquence alternative proposée par un participant pour la phrase correspondant au français « On peut servir comme ça, encore chaud (par exemple avec de la compote de pommes). »

Cas de l'alsacien

mot original	variante 1	variante 2	variante 3	variante 4	variante 5
'r	er				
Dr	D'r	De	Der		
Dreiha	Drahja	Dreihe	draje		
drüs	d'rüs				
e	a				
Galrìewle	Galerewle	Galerieble	Galriawla	Galerìewle	Galriawla
Griaß	Grees	Gress	Greß		
Griaßpflütta	Greespflüdde	Greßpflütte	Griesbap	Griespflüdde	GrussFlutta
güet	gùt	güat	guet		
kàt	kààt	kànt	kànn	kàt's	

TABLEAU 6.5 – Extrait des graphies alternatives myriadisées sur Recettes de Grammaire.

Lorsqu'ils ont renseigné ces informations dans leur profil, nous connaissons la ville ou le village d'origine des participants ainsi que les langues qu'ils parlent. Dans le cas de l'alsacien, quatre des zones dialectales principales sont représentées par les dix participants ayant ajouté un texte ou des graphies alternatives et ayant renseigné le lieu où ils ont appris l'alsacien (voir figure 6.4).

Au total, 215 paires de graphies alternatives ont été myriadisées sur Recettes de Grammaire. Les mots concernés sont de catégories variées, comme l'illustrent les cas de [Dr – D'r – De – Der] (déterminant « le »), de [Griaß – Grees – Gres – Greß] (nom commun « semoule »), de [güet – gùt – güat – guet] (adjectif ou adverbe « bien »), etc. On retrouve parmi ces paires l'alternance des voyelles -a et -e caractéristiques des variantes du sud et du nord de l'Alsace, mais aussi des alternances consonantiques telles que -dd- et -tt-, ou des motifs de variation plus complexes tels que l'alternance -eih-, -ahj-, -eih- , -aj-.

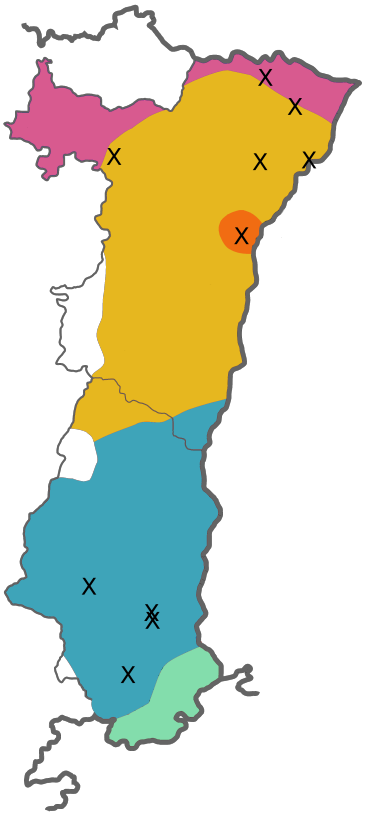


FIGURE 6.4 – Ville ou village de provenance de dix participants ayant produit des données variées.

Cas du créole mauricien

mot original	variante 1	variante 2
pandan	pendan	
kouyer	kuyer	couyere
karay	caraill	
tranpe	trampe	
lane-la	lannee la	
bizin	bisin	
di riz	diri	
conzelé	konzele	

TABLEAU 6.6 – Extrait des graphies alternatives myriadisées sur Ayo !.

Dans le cas du créole mauricien les trois participants ayant renseigné leurs villes d'origine proviennent de zones urbaines proches de la capitale Port Louis.

La taille de cette ressource ne nous permet naturellement pas de couvrir l'ensemble du vocabulaire. Néanmoins, nous pouvons utiliser ces ressources pour inférer les mécanismes de la variation à l'écrit. Nous donnons un exemple d'exploitation de la ressource produite pour l'alsacien dans le chapitre 8.

6.4 Conclusion

Les expériences de myriadisation que nous avons menées nous ont permis de valider une des hypothèses formulées dans l'introduction : « Il n'y a pas de raison que le succès d'une entreprise participative (en termes de qualité des ressources produites) dépende de la langue à laquelle elle est appliquée. ». En effet, les annotations produites par les participants dans le cas de l'alsacien, pour lequel le corpus brut était disponible en quantité suffisante, montrent l'intérêt de la démarche.

Les nombre de participants sur nos plateformes (une centaine de locuteurs ont contribué dans le cas de l'alsacien, et 17 dans le cas du créole guadeloupéen et du créole mauricien) est tout à fait respectable, notamment comparé aux chiffres rapportés par des expériences de myriadisation similaires portant sur des langues présentant un nombre bien plus important de locuteurs. D'après Chamberlain *et al.* (2013), la première version de `Phrase Detectives` (pour l'anglais) a par exemple attiré 2 000 joueurs en 32 mois, tandis que `Jeux de Mots` (pour le français) a recruté 2 700 joueurs en 56 mois, avec une moyenne de 48 joueurs par mois.

Néanmoins, nous n'avons pas pu valider la seconde hypothèse formulée : « Concernant la quantité de locuteurs à mobiliser, la motivation de ceux-ci quant à l'urgence de disposer de ressources et d'outils adaptés suffit à compenser un nombre de locuteurs moindre. ». En effet, la participation que nous avons suscitée suit la même tendance que celle observée dans les entreprises de myriadisation mentionnées ci-dessus : peu de participants produisent la majorité des données et le maintien d'une communauté active de participants demande un effort de communication permanent. C'est ce constat qui nous a d'une part poussée à développer les fonctionnalités ludiques de nos plateformes et nous a d'autre part encouragée à engager un dialogue plus soutenu pour favoriser un réel échange avec les communautés d'internautes.

L'enthousiasme des participants ayant contribué *via* la fonctionnalité « Moi, j'aurais dit ça comme ça ! » est un premier pas dans ce sens. Cette fonctionnalité permet de reconnaître la diversité des pratiques linguistiques et le (relatif) succès rencontré confirment que celle-ci constitue un enjeu réel pour le traitement automatique de ces langues.

En définitive, les résultats obtenus sont encourageants en termes de qualité, mais il existe une marge de progression quant à la quantité des participants et de ressources produites par ceux-ci. Dans la suite de ce travail, nous montrons comment nous avons évalué les ressources d'ores et déjà obtenues en les intégrant au développement de nouveaux outils de traitement.

Chapitre 7

Apprentissage supervisé sur le corpus myriadisé

Sommaire

Nous présentons dans ce chapitre nos expériences d'apprentissage supervisé menées pour l'alsacien. Nous nous sommes en particulier intéressée à deux stratégies permettant de tirer parti des ressources disponibles et de pallier le manque de données d'entraînement pour le développement d'un outil d'annotation en parties du discours.

La première stratégie consiste à utiliser, en complément d'un corpus d'entraînement de taille réduite, un lexique qui permet d'étendre le vocabulaire d'entraînement et ainsi d'améliorer les performances d'étiquetage sur les mots « hors vocabulaire », c'est-à-dire n'apparaissant pas dans le corpus d'entraînement. La seconde approche se propose d'exploiter les corpus bruts existants, même de petite taille, pour l'entraînement de plongements lexicaux intégrés à l'apprentissage supervisé.

Dans la première section de cette partie (section 7.1) nous présentons ainsi une série d'évaluations d'un système d'étiquetage séquentiel, `MElt` (Denis et Sagot, 2012), qui repose sur une architecture de type chaîne de Markov à maximum d'entropie (MEMM) (Ratnaparkhi, 1996). Cet étiqueteur est conçu pour bénéficier de l'exploitation d'un lexique complémentaire lors de son entraînement.

Afin de montrer la validité de notre démarche de myriadisation, nous avons utilisé le corpus annoté myriadisé de l'alsacien pour entraîner cet outil. Afin de pouvoir comparer les différents modèles entraînés nous les évaluons sur le seul corpus $CRef_{gsw}$, présenté dans la section 5.2.4 et également utilisé pour évaluer les annotations des participants. Ce corpus de 102 phrases contient plusieurs variantes de l'alsacien ; sa composition est détaillée dans le tableau 7.1. Les résultats que nous présentons pour les modèles entraînés correspondent toujours à la valeur d'exactitude moyenne accompagnée de l'écart-type sur 10 itérations (*runs*) de la phase d'entraînement, le corpus d'évaluation restant inchangé.

Langue	**Nom**	**Variante**	**Nb. phrases** (**Nb. *mots morpho.***)	**Source**
Alsacien	$CRef_{\text{Sv},1}$	Bas alémanique du sud	20 (372)	`Wikipédia`
	$CRef_{\text{Sv},2}$	Bas alémanique du sud	27 (503)	`Wikipédia`
	$CRef_{\text{Nv},1}$	Bas alémanique du nord	26 (362)	Pièce de théâtre
	$CRef_{\text{Nv},2}$	Bas alémanique du nord	29 (231)	Recettes

Tableau 7.1 – Composition du corpus de référence pour l'alsacien $CRef_{gsw}$.

La seconde section (section 7.2) reprend l'expérience de Magistry *et al.* (2018) qui montre que l'utilisation complémentaire de *MorphoSyntactic Embeddings* permet de dépasser les performances d'architectures de type supervisées classiques sur le picard, le malgache et l'alsacien. Nous y présentons notre expérience de réplicabilité de cette expérience.

7.1 Apprentissage supervisé sur le corpus myriadisé

Les premières expériences menées pour mesurer l'intérêt du corpus myriadisé consistent à l'utiliser pour entraîner un outil d'annotation supervisé, en l'occurrence `MElt` (Denis et Sagot, 2010). Nous avons choisi cet outil car, étant donnée la taille des corpus dont nous disposions, il était intéressant de pouvoir mesurer l'impact d'un lexique complémentaire. Par ailleurs, `MElt` ayant été entraîné pour de nombreuses langues et dans des contextes de ressources linguistiques de tailles variées (voir notamment (Sagot, 2016)), cela nous a permis de comparer les modèles obtenus pour l'alsacien aux modèles existants.

Le corpus utilisé est celui présenté et évalué dans la section 6.2.1 tel qu'il était en 2019, c'est-à-dire totalisant 9 282 mots morphosyntaxiques (la ressource étant dynamique, le corpus annoté en contient aujourd'hui 9 611).

Outre ce corpus d'entraînement, nous utilisons un lexique de noms propres, L_{licensed}, issu de la concaténation de deux ressources distribuées sous des licences claires :

- Le Lexicon of Place Names in the Alsatian Dialects (Bernhard, 2018a), distribué sous licence CC BY-SA [140], qui contient des noms de lieux en alsacien ;

140. Voir https://zenodo.org/record/1404873#.X0Y0yBngq-o, juin 2020.

- un extrait du Lefff (Sagot, 2010), distribué sous licence « Lesser General Public License For Linguistic Resources » (LGPL-LR) [141] se réduisant aux noms propres du lexique.

Ont par ailleurs été mis à notre disposition les deux lexiques présentés dans la section 4.4.1 : $L_{\text{MO\_gsw}}$ (contenant uniquement des « mots-outils ») et L_{gsw} (formé à partir de plusieurs sources). Ces ressources, dont la licence n'est pas connue, ont été développées par D. Bernhard.

Le lexique complet formé par la concaténation de L_{licensed} et de ces deux lexiques, $L_{\text{full\_gsw}}$, totalise 110 786 entrées. Il est, par construction, pour moitié composé de noms propres ; la répartition des entrées par catégorie est donnée dans le tableau 7.2.

ADJ	ADP	ADP +DET	ADV	CCONJ	DET	NOUN	PART	PRON	PROPN	PUNCT	SCONJ	VERB
3 878	97	9	539	14	39	28 842	12	126	59 505	31	4	17 690

Tableau 7.2 – Répartition des entrées du lexique par étiquette.

7.1.1 Apport du lexique

Nous avons mené plusieurs expériences permettant d'observer l'apport des différents lexiques.

Les lexiques $L_{\text{MO\_gsw}}$ et L_{gsw} ne sont pas distribués sous des licences claires. Si nous présentons ici les résultats de `MElt` lorsqu'ils sont utilisés comme lexique complémentaire à titre informatif, les modèles que nous distribuons sont ceux qui ont été développés à partir de ressources dont la licence est connue et qui sont distribuées de manière pérenne.

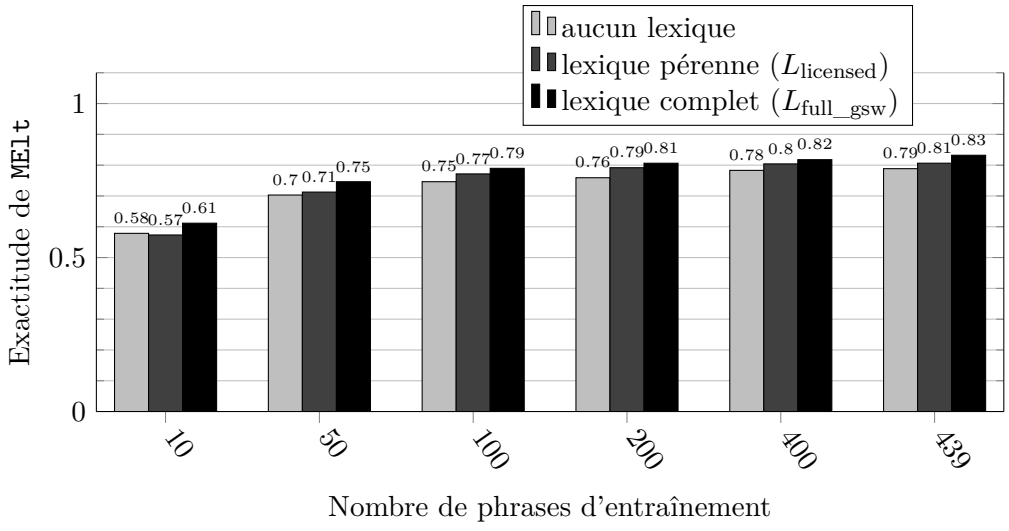


Figure 7.1 – Performances de `MElt` (exactitude) selon la taille en nombre de phrases du corpus d'entraînement et le lexique additionnel intégré.

Nous présentons dans la figure 7.1 les exactitudes moyennées des modèles `MElt` entraînés en faisant varier (i) la taille du corpus d'entraînement entre 10 et 439 phrases, et (ii) les lexiques complémentaires utilisés. Pour chaque lexique et chaque corpus de taille x phrases, 10 entraînements ont été effectués sur 10 tirages aléatoires de x phrases du corpus complet.

141. Voir http://pauillac.inria.fr/~sagot/index.html#lefff, juin 2020.

En entraînant MElt$_{Myriadise}$ avec le corpus myriadisé complet et le lexique $L_{\text{full\_gsw}}$, l'exactitude moyenne obtenue est de 83,22±0,002 %.

L'évolution des performances illustrées dans la figure 7.1 montrent l'efficacité de l'utilisation d'un lexique additionnel : par rapport à un entraînement sans lexique, l'ajout de L_{licensed} à l'entraînement de `MElt` apportant un gain moyen de près de 4 %, et l'ajout $L_{\text{full\_gsw}}$ de 5,5 % environ, d'un peu plus de quatre points.

Pour les modèles entraînés avec $L_{\text{full\_gsw}}$, le gain de près de six points lorsque le corpus d'entraînement passe de 50 à 100 phrases tombe à 2,1 points lorsque la taille est doublée une première fois puis à 1,4 points lorsque le corpus d'entraînement atteint 400 phrases.

7.1.2 Analyse par étiquette

Nous avons analysé les performances de MElt$_{Myriadise}$ lorsque entraîné avec L_{licensed} comme lexique complémentaire.

De manière similaire à ce que nous avons pu observer sur les annotations myriadisées produites par des annotateurs humains, notre analyse des F-mesure par étiquette des annotations automatiques montre que les performances les plus faibles (inférieures à 0,5) concernent les étiquettes les moins représentées (PART, SCONJ, SYM, ADP+DET), ainsi que la catégorie X qui représente 2 % du corpus d'évaluation et pour laquelle la F-mesure est de 0,3.

Les distributions des étiquettes sont similaires dans le corpus d'entraînement et d'évaluation pris dans leur ensemble. En revanche, comme montré par le tableau 7.3, les distributions ne sont pas équilibrées entre les corpus d'évaluation.

En particulier, la catégorie PROPN est 2,2 fois plus présente dans le corpus d'évaluation $CRef_{\text{Sv}}$ que dans le corpus d'apprentissage. Elle est confondue avec NOUN par l'outil dans près de 60 % des cas. L'ajout du lexique L_{licensed}, qui ne contient que des noms propres, permet d'augmenter les performances de `MElt` de 3,2 points sur le corpus $CRef_{\text{Sv}}$.

Une autre erreur fréquente est la confusion entre AUX et VERB (dans 30 % des cas) et entre VERB et ADJ (dans 25 % des cas). Ces confusions sont également classiques chez les annotateurs humains (voir section 5.2.4) et portent sur des catégories intrinsèquement difficiles.

7.1.3 Évaluation comparative de la méthode

Dans cette section, nous comparons d'une part notre méthode aux résultats rapportés par Bernhard et Ligozat (2013) pour l'annotation en parties du discours. D'autre part, nous montrons que les performances atteintes par le modèle entraîné sur le corpus myriadisé sont similaires à celles attendues par un modèle entraîné sur un corpus annoté manuellement par des linguistes.

Bernhard et Ligozat (2013) présentent les performances du `Stanford Tagger` une fois les « mots-outils » transposés en allemand sur deux textes de C_{Ref} : $CRef_{\text{Nv},1}$ et $CRef_{\text{Sv},2}$. Sur la pièce de théâtre, nous obtenons une exactitude moyenne de 72,18±0,004 % bien en deçà des 83 % obtenus par le `Stanford Tagger`, mais pour l'article *Elsassisch Museum (Stroßburri)* de la `Wikipédia` alémanique, MElt$_{Myriadise}$ atteint une exactitude moyenne de 86,59±0,011 %, plus proche des performances du `Stanford Tagger` (85 %).

	$C_{Myriadise}$	$CRef_{\text{Sv}}$ (`Wikipédia`)	$CRef_{\text{Nv,1}}$ (pièce de théâtre)	$CRef_{\text{Nv,2}}$ (recettes)
ADJ	6 %	3 %	3 %	6 %
ADP	12 %	10 %	7 %	9 %
ADP+DET	3 %	4 %	3 %	2 %
ADV	6 %	5 %	7 %	6 %
AUX	4 %	6 %	2 %	0 %
CCONJ	4 %	3 %	6 %	7 %
DET	11 %	13 %	8 %	12 %
INTJ	0 %	0 %	1 %	0 %
NOUN	17 %	14 %	9 %	21 %
NUM	3 %	4 %	0 %	3 %
PART	1 %	0 %	1 %	0 %
PRON	6 %	3 %	12 %	2 %
PROPN	**3** %	**8** %	4 %	0 %
PUNCT	13 %	10 %	19 %	13 %
SCONJ	1 %	0 %	2 %	1 %
SYM	1 %	0 %	0 %	0 %
VERB	10 %	8 %	11 %	18 %
X	1 %	9 %	4 %	0 %

TABLEAU 7.3 – Distribution des étiquettes dans le corpus d'entraînement $C_{Myriadise}$ et les corpus d'évaluation pour l'alsacien.

Bien que les faibles tailles de ces deux corpus (respectivement 230 et 396 mots) ne permettent pas de conclure sur les raisons de cette différence, nous formulons les deux hypothèses suivantes : le texte théâtral est le seul texte du corpus constitué de dialogues, registre absent de notre corpus d'entraînement. D'autre part, le fait que ce corpus soit rédigé en bas-rhinois peut avoir un impact sur les performances de `MElt`. Il semble donc que la méthodologie de Bernhard et Ligozat (2013) soit plus robuste à cette variation. Néanmoins, nous ne voyons pas de marge de progression pour les résultats avancés dans ce cadre, les performances maximales ayant été atteintes compte tenu de la méthodologie proposée.

D'autre part, Bernhard *et al.* (2018a) ont publié une nouvelle ressource en 2018 : le `Corpus annoté pour les dialectes alsaciens`, annoté par des linguistes experts avec le jeu de 18 étiquettes que nous utilisons ici, étendu avec les catégories EPE (épenthèse) et MOD (auxiliaire modal) (Bernhard, 2018b). Ce corpus contient 12 593 mots (708 phrases) et est disponible sous licence CC BY-SA [142]. Il est composé à 70 % d'articles issus de Wikipédia WKP et à 30 % de chroniques écrites dans un magazine d'information haut-rhinois (Bernhard *et al.*, 2018b). Nous en avons exclu l'extrait correspondant à C_{Ref} afin de former C_{Trad}, corpus contenant 11 087 mots (623 phrases) et avec lequel nous avons entraîné MElt_{Trad}. L'exactitude moyenne de ce modèle est de 84,16±0.003 %.

Nous avons modifié manuellement C_{Trad} pour qu'il corresponde au jeu d'étiquettes, utilisé pour

142. Voir : https://zenodo.org/record/2536041#.Xx1rcBHgq-o.

annoter $C_{Myriadise}$ [143].

Afin de comparer le corpus myriadisé au corpus annoté par des linguistes, nous avons procédé à 10 entraînements effectués sur 10 tirages aléatoires de 439 phrases de ce corpus pour former C_{Trad_comp}. L'exactitude moyenne obtenue de cette manière par MElt$_{Trad_comp}$ est de 83,48±0,005 %, un score très proche de ce que nous obtenons avec le corpus myriadisé.

Les meilleures performances sont obtenues en entraînant un modèle avec un corpus totalisant 20 369 mots formé par la concaténation du corpus myriadisé et du corpus traditionnel, $C_{Myriadise+Trad}$ soit une exactitude moyenne de 87,33±0,005 %.

Le tableau 7.4 récapitule les résultats obtenus pour les différents modèles `MElt`, chacun ayant été entraîné avec le lexique complémentaire $L_{\text{full\_gsw}}$.

	Corpus d'entraînement	**Nb. phrases (nb. mots morpho.)**	**Exactitude**
MElt$_{Myriadise}$	$C_{Myriadise}$	439 (9 282)	83,22±0,002 %
MElt$_{Trad_comp}$	C_{Trad_comp}	439 (8 988)	83,48±0,005 %
MElt$_{Trad}$	C_{Trad}	623 (11 087)	84,16±0,003 %
MElt$_{Myriadise+Trad}$	$C_{Myriadise+Trad}$	1 062 (20 369)	87,33±0,005 %

TABLEAU 7.4 – Exactitudes obtenues pour les différents modèles de `MElt` entraînés.

Ces résultats montrent l'intérêt du corpus myriadisé : le modèle entraîné sur le seul corpus $C_{Myriadise}$ est compétitif avec le modèle entraîné sur un corpus annoté par des linguistes. En outre, la combinaison de ces deux ressources permet de dépasser l'état de l'art pour l'annotation de l'alsacien.

On constate en revanche que toutes les performances obtenues sont en-deçà de ce que Sagot (2016) rapporte dans des situations similaires où `MElt` est entraîné avec des corpus de petite taille, notamment sur l'estonien (corpus d'entraînement de 7 687 mots), le roumain (9 291 mots), ou le tamoul (6 329 mots), langues pour lesquelles les performances sont de 89,64 %, 91,09 % et 89,14 % respectivement.

Les performances sur l'alsacien sont également en deçà des résultats rapportés par Fort et Sagot (2010), où `MElt` atteint 86,6 % d'exactitude avec un corpus d'entraînement de 100 phrases pour l'anglais, ou par Vergez-Couret *et al.* (2014) qui entraînent `Talismane` avec un corpus d'entraînement de 2 500 mots et un lexique de 225 000 entrées en occitan pour atteindre 89 % d'exactitude.

Nous avons fait l'hypothèse que la variation dialectale et scripturale présente dans nos corpus était à l'origine de ces moindres performances. C'est pour tester cette hypothèse que nous avons analysé les performances de nos outils sur les différents sous-corpus comme illustré par les expériences présentées dans la section suivante.

143. La seule épenthèse du corpus a été annotée avec l'étiquette X faute de mieux, et les étiquettes MOD ont été remplacées par AUX.

	$\text{MElt}^{Sans\_Weiss}_{Myriadise}$	$\text{MElt}_{Myriadise}$
$CRef_{\mathbf{Sv},1}$	**88,05±0,008** %	86,59±0,012 %
$CRef_{\mathbf{Sv},2}$	**83,24±0,003** %	82,39±0,006 %
$CRef_{\mathbf{Nv},1}$	62,48±0,008 %	**72,18±0,004** %
$CRef_{\mathbf{Nv},2}$	78,46±0,014 %	**87,96±0,006** %
$CRef_{gsw}$	78,82±0,010 %	**83,22±0,002** %

TABLEAU 7.5 – Comparaison des modèles entraînés sans et avec $C^{Weiss}_{Myriadise}$.

7.1.4 Analyse par variante

Afin de mettre en évidence les difficultés posées par le caractère multi-variant de nos corpus d'entraînement et d'évaluation, nous avons procédé à plusieurs évaluations. D'une part, nous avons regardé les performances des outils entraînés sur les quatre textes composant le corpus d'évaluation séparément. D'autre part, nous avons mesuré l'impact de l'ajout au corpus d'entraînement d'un corpus annoté d'une variante identifiée par son auteur : l'alsacien de Strasbourg. En effet, un des textes ayant été utilisés pour alimenter la plateforme d'annotation `Bisame` a été proposé par un des participants, R. Weissenburger. Ce texte, intitulé « *E Hochzit in de 50er Johre*[144] », a été rédigé par son auteur en « bas alémanique du nord, dans la forme dialectale du parler strasbourgeois (bas alémanique du nord teinté de francique) »[145].

Ce corpus, $C^{Weiss}_{Myriadise}$, comporte 169 phrases (4 173 mots morphosyntaxiques). Il constitue ainsi près de 45 % du corpus d'entraînement. Nous présentons dans le tableau 7.5 les performances par sous-corpus des modèles entraînés ainsi que l'impact de l'ajout de ce corpus à l'entraînement de MElt : le modèle $\text{MElt}^{Sans\_Weiss}_{Myriadise}$ est entraîné avec le corpus myriadisé de 271 phrases et 5 109 mots, le modèle $\text{MElt}_{Myriadise}$ est entraîné avec le corpus complet. Les performances des différents modèles varient selon les corpus. Globalement, l'impact de l'ajout de $C^{Weiss}_{Myriadise}$ au corpus d'entraînement est positif, avec un gain moyen d'environ 4,5 points. En revanche, les meilleures performances observées sur les corpus en variante du nord (+9,7 et +9,5 points) s'accompagnent d'une baisse des performances sur les corpus en variante du sud (-1,46 et -0,85).

L'augmentation des performances sur le sous-corpus du Nord est attendue, la variante strasbourgeoise étant une variante bas-rhinoise. En revanche, il est regrettable que cela se traduise également par une baisse des performances sur le sous-corpus du Sud. Il apparaît ici que la stratégie consistant à augmenter la taille du corpus d'entraînement sans prendre en considération les variantes n'est pas optimale.

7.1.5 Conclusion

Les expériences menées dans cette section ont permis de montrer l'intérêt du corpus myriadisé de l'alsacien. Nous avons pu entraîner grâce à lui un modèle `MElt` dont les performances sont équivalentes au même outil entraîné avec un corpus annoté par des linguistes. Ce résultat montre la validité de la myriadisation telle que nous l'avons implémentée pour produire un corpus annoté en parties du discours de qualité.

144. « Un mariage dans les années 50 ».
145. R. Weissenburger, communication personnelle du 27 septembre 2018.

Par ailleurs, nous avons observé que l'augmentation en taille du corpus d'entraînement peut se traduire par une baisse des performances de l'outil lorsque les variantes présentes dans le corpus d'entraînement diffèrent de celles du corpus d'évaluation. En doublant la taille du corpus d'entraînement de manière agnostique des variantes dialectales, on observe en effet une baisse de l'exactitude sur un des sous-corpus. Cette démarche pragmatique semble pour autant la plus réaliste : la collecte « opportuniste » de corpus d'entraînements et la nature *a priori* inconnue des corpus à annoter en situation réelle ne permettent pas de connaître précisément les variantes représentées dans chaque corpus.

Il nous semble que seul l'apprentissage supervisé d'outils propres à chacune des variantes permettrait d'éviter l'écueil observé. Cela requerrait de pouvoir d'une part les distinguer les unes des autres et d'autre part de recueillir des corpus annotés de taille suffisante pour chacune d'entre elles.

Étant donné les difficultés posées par ces deux tâches d'identification et d'annotation, nous nous sommes intéressée aux possibilités offertes par la myriadisation pour tirer parti de l'ensemble des ressources pour annoter chacune des variantes. Cela s'est traduit par la myriadisation de variantes scripturales alignées grâce auxquelles nous tâchons de faire le lien entre les différentes variantes présentes dans nos corpus. Nous présentons les résultats des expérimentations menées dans cette logique dans le chapitre 8.

7.2 Reproduction et extension d'une expérience d'annotation utilisant des plongements lexicaux

Cette section reprend un article écrit en commun avec Karën Fort et Pierre Magistry publié à l'atelier Ethique et TRaitemeNt Automatique des Langues (ETeRNAL) 2 de TALN 2020 (Millour *et al.*, 2020).

Nous y décrivons la réplication de l'expérience proposée par Magistry *et al.* (2018) concernant l'étiquetage en parties du discours de langues peu dotées par spécialisation des plongements lexicaux. Le travail de réplication a été réalisé en étroite collaboration avec un des auteurs de l'article d'origine et a mis au jour les éléments manquants dans la présentation de l'expérience, et a permis de les compléter, et d'étendre la recherche en proposant une étude préliminaire de la robustesse de cette méthode à la variation. Nous remercions ici A-L. Ligozat et S. Rosset (LIMSI-CNRS), co-autrices de l'article original, ainsi que D. Bernhard (LiLPa, Strasbourg) pour leur disponibilité, leurs conseils et l'aide qu'elles nous ont apportée.

L'architecture que nous avons reproduite lors de cette expérience de réplication est illustrée par la figure 7.2.

7.2.1 Reproduire ou répliquer ?

La terminologie utilisée mérite qu'on s'y attarde, tant elle rend compte de la complexité de l'acte, apparemment simple, de rejouer une expérience décrite dans un article de recherche. Nous reprenons ici de manière succincte les questions mises au jour et détaillées dans (Cohen *et al.*,

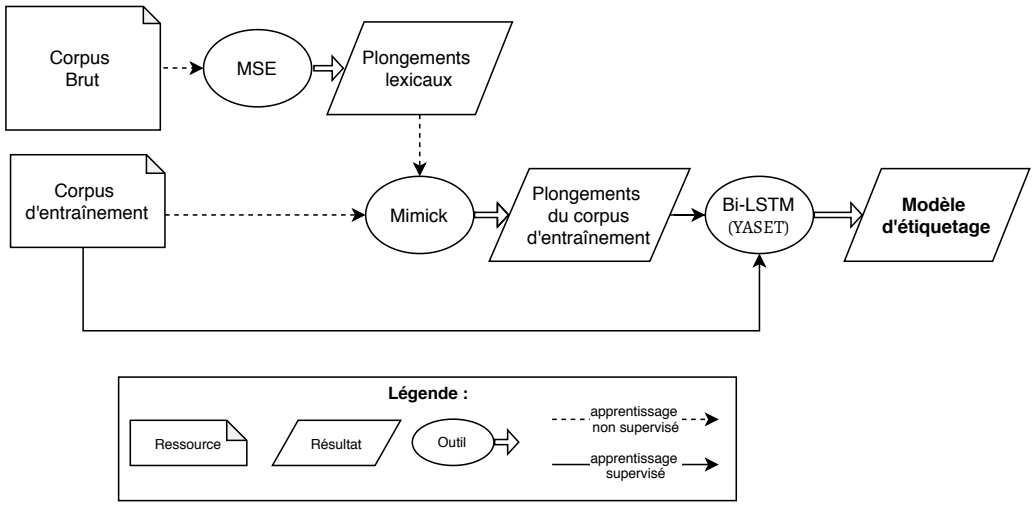


FIGURE 7.2 – Architecture pour la spécialisation des plongements lexicaux présentée dans (Magistry *et al.*, 2018).

2018). La réplicabilité est une propriété d'une expérience, celle d'être rejouée ou répétée [146], alors que la reproductibilité est une propriété des **résultats** de l'expérience menée : on peut obtenir les mêmes conclusions ou les mêmes valeurs [147]. Nous nous intéressons ici en priorité à répliquer l'expérience (pour mieux la comprendre), pour ensuite tenter d'en reproduire le résultat (pour être sûr de partir sur les mêmes bases) pour, enfin, étendre l'expérience.

De tels efforts sont de plus en plus valorisés dans le domaine du TAL. Après deux ateliers à LREC 2016 et 2018 (Branco *et al.*, 2016; Branco, 2018), une *shared task*, REPROLANG [148], a été organisée dans le cadre de LREC 2020. La possibilité de rejouer une expérience est même devenue un critère de sélection pour COLING 2018. Une étude menée parmi les chercheurs du domaine a montré que le sujet est perçu comme un problème important par la majorité des répondants (Mieskes *et al.*, 2019) et que, lorsque ceux-ci ont essayé de reproduire une expérience (et y sont parvenus), les résultats obtenus se sont très souvent révélés significativement différents de ceux publiés. Cela ne signifie pas pour autant que les auteurs originaux sont de mauvaise foi. Simplement, le manque de documentation des expérimentations empêche souvent de se replacer dans les conditions expérimentales de l'expérience initiale [149].

Parmi les éléments trop souvent mal documentés, les pré-traitements (dont la tokenisation) et les versions des logiciels et des ressources langagières utilisées sont des classiques (Fokkens *et al.*, 2013). L'expérience reproduite ici ne fait pas exception, malgré les efforts de ses auteurs.

146. « *Replicability or repeatability is a property of an experiment : the ability to repeat –or not– the experiment described in a study.* » (p. 3)
147. « *Reproducibility is a property of the outcomes of an experiment : arriving –or not– at the same conclusions, findings, or values.* » (p. 3) (Cohen *et al.*, 2018).
148. Voir : https://lrec2020.lrec-conf.org/en/reprolang2020/.
149. Nous ne formulons pas ici d'hypothèse quant aux raisons méthodologiques ou pratiques de ce manque de documentation dans le domaine du TAL. Il nous a été signalé par un des relecteurs que dans d'autres domaines souffrant tout autant de la précarité de leurs chercheurs, la reproductibilité systématique des expériences est assurée, notamment grâce à l'utilisation de « carnets de recherche » pour leur documentation.

7.2.2 Faire tourner le code

Plutôt que de ré-implémenter la solution proposée, nous avons essayé de retrouver les conditions initiales dans lesquelles l'expérience avait été menée, tant au niveau du logiciel que des ressources langagières. Dans cette section, nous présentons donc la méthodologie que nous avons souhaitée reproduire ainsi que les éléments relatifs (i) à la disponibilité du code source et (ii) à la mise en place des configurations logicielles nécessaires pour faire tourner ce code.

7.2.2.1 Méthodologie

La méthodologie proposée permet de spécialiser les plongements lexicaux à la tâche d'annotation en parties du discours en combinant l'analyse au niveau caractère et l'utilisation des propriétés morphosyntaxiques pour un mot cible et son contexte. Le système Mimick (Pinter *et al.*, 2017), qui se base sur la graphie des mots pour calculer les vecteurs, est utilisé pour établir les plongements des mots hors vocabulaire, nombreux dans le cas de langues peu dotées et non standardisées.

Cette méthodologie peut être découpée en trois étapes permettant d'obtenir des résultats intermédiaires : i) entraînement sur un corpus brut permettant de produire un fichier de plongements lexicaux dits morphosyntaxiques, ii) entraînement du modèle de *tagger* basé sur un Bi-LSTM en utilisant les plongements lexicaux et iii) évaluation du *tagger*. Il est à noter que les deux éléments intermédiaires (fichiers de plongements et modèle du Bi-LSTM entraînés pour l'alsacien) ne sont pas distribués.

7.2.2.2 Accès au code source

Le code tel qu'utilisé dans l'expérience originale n'a pas pu être retrouvé. Le co-auteur en charge des expériences ayant terminé le postdoctorat qu'il réalisait à l'époque de la publication de l'article, il n'a aujourd'hui plus accès aux machines sur lesquelles celui-ci était stocké. Nous avons néanmoins eu accès à deux versions ultérieures du code source, correspondant à deux implémentations de la méthodologie décrite dans l'article.

Le premier code source auquel nous avons eu accès, CS_1 [150] a été mis à disposition par le premier auteur de l'article. Le dépôt GitHub transmis contient une réécriture partielle du code original qui a été abandonnée avant son terme.

Un second dépôt à été identifié dans un second temps, CS_2 [151]. Il s'agit de la version simplifiée, documentée et transposée en python du code original, réalisée et distribuée par une postdoctorante ne faisant pas partie des auteurs initiaux de l'article.

Ces deux dépôts GitHub n'étant pas renseignés dans l'article, ni associés aux noms des auteurs, ils ne pouvaient pas être identifiés sans prise de contact avec ceux-ci. Ces derniers étant encore précaires et leurs affiliations changeant régulièrement, nous avons donc eu de la chance d'une part de parvenir à entrer en contact avec l'un d'entre eux malgré la désactivation de sa boîte mail, et d'autre part de pouvoir accéder à la deuxième version du code.

150. Accessible ici : https://github.com/a-tsioh/MSETagger.
151. Accessible ici : https://github.com/eknyazeva/MSETagger_py.

7.2.2.3 Accès à des modèles pré-entraînés

Les modèles pré-entraînés correspondant aux expériences décrites dans l'article ne sont pas diffusés. Le système qui y est décrit produit trois modèles :

1. les plongements lexicaux initiaux,
2. le modèle MIMICK qui permet de les compléter,
3. les poids du Bi-LSTM de l'étiqueteur final.

Si la distribution des modèles est une pratique de plus en plus courante en TAL, les auteurs initiaux ont fait le choix de diffuser le code permettant de reconstruire ces modèles, mais aucun des résultats intermédiaires. Chacune de ces trois étapes recourt à de l'apprentissage profond qui suppose une initialisation aléatoire de grandes matrices de poids. La stabilité de ces modèles n'est pas garantie. Elle a même d'autant plus de chances d'être problématique lorsque les corpus d'entraînement sont relativement petits, comme c'est le cas ici [152].

7.2.2.4 Configuration logicielle

Les deux dépôts GitHub sont accompagnés de `README` contenant la majorité des informations de configuration nécessaires à l'exécution du code, notamment une liste de dépendances quasi complète. Les versions de certaines librairies python sont absentes de la documentation, mais les versions compatibles entre elles des différentes librairies ont pu être déduites à tâtons.

De la même manière, l'architecture de Bi-LSTM sur laquelle s'appuie le travail des auteurs est l'implémentation YASET (Tourille *et al.*, 2017). La version de YASET utilisée n'était précisée que dans l'un des dépôts.

À ces difficultés s'est ajoutée la méconnaissance initiale des technologies employées par les auteurs (par exemple le langage `scala`, et le moteur de production `sbt`), qui constitue un frein important à la réplication de l'expérience. La mise en place de la configuration logicielle n'a pu être faite que grâce au soutien apporté par premier auteur de l'article d'origine.

Dans les deux cas, nous avons repris les paramètres spécifiés par les auteurs des codes initiaux. Lorsque les hyper-paramètres n'étaient pas précisés dans l'article, ils étaient donnés dans un fichier de configuration distribué avec le code source [153].

7.2.3 Données utilisées

7.2.3.1 Corpus bruts

Les corpus bruts utilisés pour entraîner les plongements lexicaux sont les corpus C_{Brut}_56k et C_{Brut}_200k. C_{Brut}_56k, communiqué sur demande par l'un des auteurs l'ayant lui-même obtenu de D. Bernhard, est constitué d'un ensemble de 103 pages Wikipédia totalisant 56 965 tokens. Ce corpus, libre de droit, peut être reconstruit à partir de la liste des pages fournies avec

152. Un article plus long, décrivant le système plus en détails et détaillant ce problème était en cours de rédaction suite à l'article de TALN 2018, mais il n'a pas pu être terminé avant la fin du projet ANR.

153. Nous avons, comme les auteurs de l'article initial, fixé le paramètre `patience` correspondant au nombre minimal d'itérations à réaliser indépendamment de l'amélioration des performances à 75 pour toutes les expériences réalisées avec le code C_2

le corpus. C_{Brut}_200k a été obtenu ultérieurement auprès de D. Bernhard. C'est un ensemble de documents contenant des pages de la Wikipédia alémanique rédigées en alsacien, ainsi que des documents dont les licences ne sont pas claires. La proportion de ce corpus qui est effectivement libre de droit n'a pas été déterminée.

7.2.3.2 Corpus annoté

Le corpus annoté de l'alsacien utilisé pour entraîner et évaluer le *tagger*, est le corpus distribué sous licence CC BY-SA [154] C_{Trad}. C'est le même corpus que celui que nous avons utilisé dans la section 7.1.3 pour comparer les performances de `MElt` entraîné sur un corpus myriadisé ou produit par des linguistes. Comme précédemment, nous en avons exclu l'extrait correspondant à C_{Ref} afin de former C_{Trad}, corpus contenant 11 087 mots (623 phrases)

7.2.4 Résultats obtenus

Le protocole d'évaluation mené dans l'article initial est très peu détaillé, en particulier quant à l'origine des corpus utilisés pour l'entraînement des plongements lexicaux et pour l'entraînement du Bi-LSTM.

Les résultats que nous obtenons sont globalement inférieurs à ceux annoncés. Outre la difficulté à reconstituer les corpus et la méthodologie initiale, il est possible que cette différence doive être attribuée à l'instabilité des plongements lexicaux entraînés sur de petits corpus (voir Section 7.2.2.3). Ceci pose la question de l'importance de la diffusion de modèles pré-entraînés. Une telle pratique favorise la reproductibilité des résultats mais dans le même temps, elle masque des propriétés importantes de la chaîne de traitement complète.

7.2.4.1 Premières expériences, réalisées avec la réécriture partielle du code (CS_1)

La première tentative de reproduction des résultats a été réalisée à partir du code CS_1 en utilisant C_{Brut}_56k pour entraîner les plongements, 80 % de C_{Trad} pour entraîner le modèle, et 20 % C_{Trad} pour l'évaluer.

Cette expérience nous a permis d'attester que nos conditions logicielles étaient les mêmes que celles de l'auteur initial (à ce jour) : nous avons en effet mené cette expérience en parallèle et obtenu le même résultat (une exactitude du *tagger* de 78 %). Il n'y a donc pas d'élément de configuration implicite n'ayant pas été communiqué par l'auteur. En revanche, la taille du corpus C_{Brut}_56k transmis par l'auteur ne correspondant pas aux données présentées dans l'article initial, nous avons poussé nos recherches pour finalement obtenir l'accès au corpus C_{Brut}_200k.

Cette expérience a également permis de mettre au jour que soit le corpus C_{Trad} disponible à ce jour en ligne n'est pas dans l'état dans lequel les expériences initiales ont été menées, soit la réécriture du code utilisé à l'époque est incomplète et ne gère plus certains cas particuliers propres à l'alsacien et pris en charge avant la réécriture. Nous n'avons en effet pas pu retrouver plusieurs éléments utilisés à l'époque, tels qu'un filtre sur les corpus bruts permettant d'éliminer les entrées de dictionnaire, et une opération visant à uniformiser les jeux d'étiquettes.

154. Voir : https://zenodo.org/record/2536041.

Concernant le jeu d'étiquettes et la tokenisation, l'article initial ne mentionne pas les choix qui ont été faits à ce sujet. Les corpus C_{Brut} et C_{Trad} sont aujourd'hui disponibles tokénisés de deux manières différentes, et le tokéniseur distribué pour l'alsacien [155] ne gère pas les cas divergents, en l'occurrence le découpage – ou non – des contractions de prépositions (ADP) et déterminants (DET), par exemple : « *zum*/ADP+DET », découpé en « *zu*/ADP *dem*/DET ».

Nous avons réalisé une seconde expérience en utilisant C_{Brut}_200k pour entraîner les plongements, et en utilisant les mêmes corpus que précédemment après uniformisation du jeu d'étiquettes. Cette nouvelle configuration nous a permis d'obtenir un *tagger* d'une exactitude de 81 %. Un score de 87 % a été obtenu plus tard par l'auteur de l'article après activation d'une option non spécifiée dans la documentation.

Ces diverses expériences ont donc montré que la réplicabilité du travail en question ne pouvait se faire sans que l'auteur ne complète le code mis à disposition. Par ailleurs, certaines ressources langagières, non librement disponibles, n'ont pu être retrouvées que par relations inter-personnelles. Enfin, certains traitements (en particulier la tokenisation) n'étaient pas suffisamment documentés et n'ont pas pu être reconstitués, ce qui, comme nous l'avons précisé en section 7.2.1, est un oubli classique.

7.2.4.2 Un pas plus loin : tester la robustesse à la variation avec le code CS_2

La première a été réalisée en utilisant C_{Brut}_200k pour l'entraînement des plongements. Nous avons utilisé C_{Trad} et $C_{Myriadisé}$ pour entraîner deux modèles dont les exactitudes moyennes sur 10 entraînements sont respectivement de 88,19±0,005 % et de 87,03±0,01 %. Il est possible que la qualité imparfaite de l'annotation sur $C_{Myriadisé}$ soit à l'origine de cette différence.

Ce code implémente selon nous de manière fiable la méthodologie présentée par les auteurs de l'article d'origine. Nous l'avons donc utilisé pour mener des expériences additionnelles, afin d'en tester la robustesse à la variation.

Pour ce faire, nous avons séparé le corpus annoté en deux sous-ensembles, en nous basant sur une caractéristique linguistique identifiée : la prédominance de la terminaison des noms et adjectifs en « -e » dans les variantes du nord, et en « -a » dans les variantes du sud (Brunner, 2001). Chaque sous-ensemble n'est pas uniforme et contient lui-même plusieurs variantes, par exemple la variante strasbourgeoise parmi les variantes du nord.

Nous avons fait l'hypothèse qu'un article Wikipédia ne contenait qu'une seule variante. Néanmoins, lorsque le calcul des fréquences relatives des terminaisons propres à chaque variante ne nous permettait pas de décider, nous avons examiné le fichier à la main. Dans tous les cas, nous avons pu identifier le biais à l'origine de l'équilibre des terminaisons, comme la fréquence élevée d'un élément de vocabulaire (par exemple le déterminant « ***de*** »), ou la présence de mots en français (par exemple, « Stad**e** d**e** l'Ill »). Nous avons ainsi pu attribuer à chaque article la variante lui correspondant. Les fréquences relatives des terminaisons en « -e » et en « -a » sont en moyenne d'un facteur 30. Nous avons ainsi pu déterminer que 40 % du corpus annoté contenait des variantes du sud ($C_{Trad-Nord}$, 4 998 tokens) et 60 % des variantes du nord ($C_{Trad-Sud}$, 7 646 tokens).

Les résultats présentés dans le tableau 7.6 ont été obtenus en découpant les deux corpus obtenus

155. Distribué par l'équipe du projet RESTAURE, voir https://zenodo.org/record/2454993.

en 3 sous-corpus : corpus d'entraînement ($C80_{Trad-X}$, 80 %), de développement (10 %), et d'évaluation ($C10_{Trad-X}$).

	$C10_{Trad-Nord}$	$C10_{Trad-Sud}$
$C80_{Trad-Nord}$	75 %	74 %
$C80_{Trad-Sud}$	74 %	79 %

TABLEAU 7.6 – Résultats de l'entraînement sur des corpus plus uniformes quant aux variantes présentes dans les corpus d'entraînement et d'évaluation.

En première analyse, il semble que la méthodologie proposée soit sensible aux variantes présentes dans les corpus : les performances les meilleures sont obtenues lorsque le corpus d'entraînement et d'évaluation contiennent les mêmes variantes. Notamment, les performances du *tagger* entraîné sur le corpus $C80_{Trad-Sud}$ diminuent de 4 points sur le corpus d'évaluation $C10_{Trad-Nord}$. Il serait intéressant de prolonger cette étude en mesurant à tailles de corpus égales pour les deux sous-variantes, l'impact de la présence - ou non - de celles-ci dans le corpus utilisé pour entraîner les plongements.

7.3 Conclusion

Nous avons proposé dans la première section de ce chapitre une évaluation extrinsèque des ressources myriadisées grâce à diverses expériences d'annotation en tirant parti. Cette évaluation montre l'intérêt d'utiliser des ressources myriadisées pour entraîner des outils lorsque les mêmes ressources annotées par des linguistes ne peuvent être obtenues. Les performances des outils entraînés sur le corpus myriadisé sont équivalentes à celles obtenues sur le corpus annoté par des linguistes, ce qui montre la qualité des ressources produites et la validité de la démarche générale.

L'étude des performances des outils développés nous a par ailleurs permis de mettre au jour leur manque de robustesse face à la variation. Il est apparu dans ce contexte qu'il n'est pas possible de tirer parti de toutes les ressources disponibles, en particulier multi-variantes, pour entraîner un outil performant : la présence de plusieurs variantes dans un corpus d'entraînement peut faire baisser les performances d'un outil sur un corpus mono-variante.

Nous avons choisi de distribuer uniquement les modèles développés à partir de ressources disponibles sous des licences claires, garantissant la reproductibilité de nos expériences [156]. En particulier, nous ne distribuons pas les modèles ayant bénéficié de l'utilisation de lexiques qui nous ont été fournis, mais qui ne sont pas distribués.

La seconde série d'expériences présentée s'inscrit dans une démarche qui vise à répliquer l'état de l'art en termes de performances d'annotation de l'alsacien publié à TALN 2018 (Magistry *et al.*, 2018). Si cette expérience ne s'inscrit pas directement dans l'évaluation des ressources produites, elle nous a permis de questionner plus avant la définition d'un résultat « état de l'art ».

Le logiciel conçu par les auteurs de l'article initial n'est plus disponible en tant que tel et nous n'avons pas réussi à le reconstruire à partir des pièces accessibles au premier auteur de l'article. Nous avons appris entre temps qu'il a été repris par une autre personne (précaire également), a

156. Voir https://alicemillour.github.io/pages/models.html.

fait l'objet d'améliorations et est désormais disponible sur un autre dépôt GitHub [157]. Nous avons répliqué l'expérience sur cette nouvelle base (qui n'est pas non plus celle de l'article initial), mais ne sommes pas parvenus à en reproduire les résultats (nous obtenons 87 % à 88 % d'exactitude en fonction du code source utilisé, *vs* 91 % dans l'article initial).

Comme évoqué en section 7.2.2.3, la question de la distribution des résultats intermédiaires (fichiers de vecteurs, modèles de *tagger*) se pose dans le cas général. Cependant, dans le contexte de langues peu dotées, la distribution de modèles instables ne paraît pas indiquée. L'effort doit selon nous être dirigé en priorité vers l'accessibilité aux ressources. L'accès à ces dernières pose de nombreux problèmes : le corpus C_{Brut} à partir duquel sont entraînés les plongements dans l'expérience initiale n'est par exemple pas librement disponible.

L'analyse des performances des outils développés par la méthode répliquée ont montré qu'ils étaient également sensibles aux variantes présentes dans les corpus d'entraînement et que l'utilisation conjointe de plongements lexicaux n'était pas suffisante pour compenser l'impact sur les performances de l'outil supervisé généré par la variabilité des corpus.

157. Voir : https://github.com/eknyazeva/MSETagger_py.

Chapitre 8

Exploitation des variantes graphiques myriadisées

Sommaire

Comme nous l'avons montré dans les deux précédents chapitres, la présence de variation dialectale en contexte non standardisé conduit à l'apparition à l'écrit de graphies caractéristiques de chacune des variantes. Les outils que nous développons ne sont pas nativement robustes à cette diversité de graphies. Dans ce contexte, utiliser toutes les ressources à disposition pour une langue donnée peut conduire à une baisse des performances sur un corpus correspondant à une variante spécifique (voir section 7.1.4).

Pour illustrer l'intérêt de la myriadisation de variantes graphiques présentée dans la section 6.3.2, nous avons évalué son impact sur la tâche d'annotation en parties du discours. En particulier, nous proposons une méthodologie visant à transposer le corpus d'application, c'est-à-dire le corpus destiné à être annoté par un outil d'annotation, pour le rapprocher des données d'entraînement.

Nos expériences d'apprentissage supervisé ont montré que l'hétérogénéité des variantes entre le corpus d'entraînement et d'évaluation peut être à l'origine de mauvaises performances sur les sections ne correspondant pas aux variantes présentes dans le corpus d'entraînement (voir section 7.1.4.

Une des raisons des faibles performances observées est la forte proportion de mots « hors vocabulaire » dans le corpus d'application. Nous référons ici par ce terme aux mots du corpus

d'application qui ne sont pas « connus » de l'outil entraîné de façon supervisée, c'est-à-dire dans notre cas qui ne sont présents ni dans le corpus d'entraînement, ni dans le lexique externe intégré à l'entraînement. Étant donné la variation graphique observée dans les langues que nous avons étudiées, nous faisons l'hypothèse que certains des mots hors vocabulaire sont en réalité des variantes graphiques de mots « connus » de l'outil.

La méthodologie que nous présentons ici a pour but de faire diminuer la proportion de mots hors vocabulaire en identifiant parmi eux ceux qui peuvent être remplacés par une de leurs variantes connues de l'outil. Il est important de noter que les opérations que nous présentons sont réalisées *après* l'entraînement de l'outil d'annotation supervisée : il ne s'agit pas ici de développer un outil mieux adapté au corpus d'application, mais bien de modifier ce dernier pour le « rapprocher » des données d'entraînement.

Le contenu de ce chapitre a fait l'objet d'une publication lors de la conférence Recent Advances in Natural Language Processing conference (RANLP 2019) (Millour et Fort, 2019).

8.1 Transposition du corpus d'application

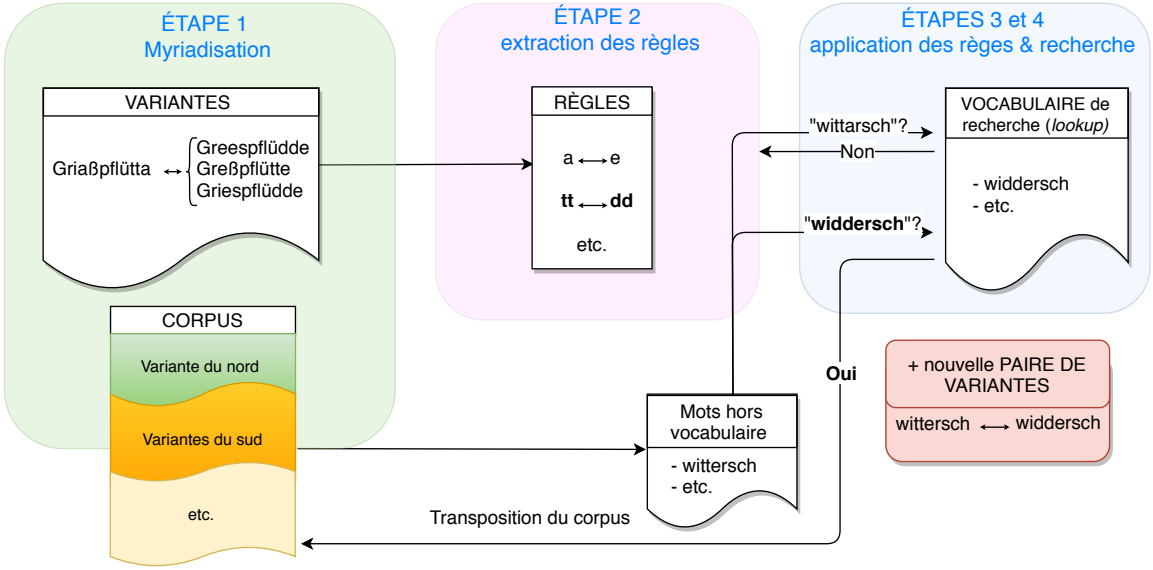


Figure 8.1 – Exploitation des variantes myriadisées pour la transposition du corpus d'application.

La méthodologie que nous proposons est illustrée dans la figure 8.1. Elle comporte quatre étapes :

1. La myriadisation de variantes, présentée dans la section 5.5 ;
2. L'extraction de règles de substitution : les variantes myriadisées sont alignées afin que puissent être identifiés les motifs de la variation que nous extrayons sous forme de règles de substitution (section 8.3) ;
3. L'application des règles aux mots hors vocabulaire pour générer des variantes candidates (voir section 8.4) ;

4. Le filtrage des variantes candidates par le vocabulaire connu (voir section 8.4).

Lorsqu'une variante candidate est présente dans le vocabulaire connu, le mot cible est remplacé par celle-ci dans le corpus d'application. À l'application de l'outil d'annotation sur le corpus transposé, le mot cible à annoter fera donc partie du vocabulaire « connu ».

Par exemple, comme illustré dans la figure 8.1, les variantes graphiques proposées pour le mot `Griaßpflütta` (`Greespflüdde`, `Greßpflütte`, et `Griespflüdde`) ont permis d'extraire, entre autres, les règles de substitution r1 : « a ↔ e » et r2 : « tt ↔ dd ».

D'autre part, dans cet exemple, un outil d'annotation a été entraîné sur le corpus myriadisé. Lorsque cet outil est appliqué sur un nouveau corpus, les mots hors vocabulaires sont identifiés. Dans l'exemple proposé, un des mots cible hors vocabulaire est *wittersch* (« plus loin »). Les règles extraites pouvant s'appliquer sur ce mot sont utilisées. L'application de la règle r1 produit le *token wittarsch*. Ce mot ne faisant pas partie du vocabulaire connu, la règle suivante est appliquée. Celle-ci permet de produire le *token widdersch* qui est lui présent dans le vocabulaire. Dans ce cas, le corpus d'application est transposé : les occurrences de *wittersch* sont remplacées par *widdersch*. En outre, la paire de graphies alternatives [`wittersch` – `widdersch`] est créée.

Outre la transposition temporaire du corpus d'application pour l'annotation, la méthodologie proposée à en effet comme effet de bord l'augmentation de la ressource initiale grâce à l'appariement progressif de nouvelles variantes graphiques.

8.2 Annoter automatiquement la variation

Le traitement automatisé de langues peu dotées et non standardisées nous amène aux frontières du TAL : dans le cas que nous avons traité, il n'existe pas de norme à laquelle nous rattacher, ni suffisamment de ressources linguistiques pour compenser les mécanismes de variation à l'œuvre. Il n'existe pas non plus de documentation sur les pratiques graphiques récentes permettant d'appréhender cette variation théoriquement. Nous avons trouvé peu de travaux impliquant de travailler avec ces deux contraintes, et, à notre connaissance, la solution que nous proposons n'avait jamais été utilisée auparavant.

Il est toutefois possible de la rapprocher d'autres expériences qui impliquent au moins une orthographe standard et sont parfois réalisées sous la supervision d'un ou plusieurs professionnels de la question linguistique. Un de ces exemples concerne le développement de `VARD 2`, un outil qui permet de standardiser manuellement et automatiquement l'« anglais moderne naissant » (*Early Modern English*) (Baron et Rayson, 2008, 2009). Un autre concerne la langue basque (Etxeberria Uztarroz *et al.*, 2014) et propose une solution pour rapporter les variantes rencontrées à leur forme standard en utilisant un analyseur morphologique existant et un corpus parallèle. Il existe par ailleurs de nombreuses publications plus ou moins récentes concernant la conception et l'utilisation de règles morphophonologiques, en particulier sous forme de transducteurs finis (*Finite State Transducteurs* (FST)), mais la plupart nécessitent l'intervention d'un linguiste au cours de leur développement. Parmi ces publications, le travail de Kimmo Koskenniemi sur la modélisation des correspondances régulières entre le finnois et l'estonien est particulièrement inspirant (Koskenniemi, 2013)Il en va de même pour le type de travail décrit par Theron et Cloete (1997), dans lequel les règles sont automatiquement extraites mais dans le but d'atteindre une cible morphologique connue. Étant donné les contextes linguistiques dépourvus de

norme consensuelle dans lesquels nous nous trouvons, ces méthodes sont inapplicables dans notre cas.

Le travail le plus proche du nôtre est celui décrit par Barteld (2017), qui concerne la détection de variantes orthographiques en moyen bas allemand, indépendamment de l'existence d'une norme. Pour autant, la méthode décrite nécessite le développement d'un classifieur permettant de filtrer les paires de cognats identifiées. Ce classifieur est basé sur une ressource qui contient 1 834 paires de variantes orthographiques, ressource qui n'était pas disponible dans nos cas mais que nous avons cherché à construire.

8.3 Génération automatique de variantes graphiques

8.3.1 Extraction des règles de substitution

Suivant les recommandations de (Prokić *et al.*, 2009) qui l'utilise avec des transcriptions phonétiques de dialectes bulgares, nous avons utilisé l'outil `ALPHAMALIG` [158] pour effectuer l'alignement des variantes myriadisées.

Pour utiliser cet outil, nous avons dû transformer les variantes collectées pour les rendre compatibles au format FASTA utilisé par `ALPHAMALIG` [159]. En particulier, ce format ne reconnaît pas les caractères accentués et certains éléments de ponctuation. Les correspondances utilisées sont présentées dans le tableau 8.1.

'à' : '0'	'è' : '1'	'ì' : '2'	'ò' : '3'	'ù' : '4'
'á' : '5'	'é' : '6'	'í' : '7'	'ó' : '8'	'ú' : '9'
'ä' : '!'	'ë' : '"'	'ï' : '#'	'ö' : '%'	'ü' : '='
'â' : '['	'ê' : '('	'î' : ')'	'ô' : '*'	'û' : '+'
'ß' : '_'	'\' : '~'	'-' : ']'		

TABLEAU 8.1 – Correspondance des caractères compatibles avec le format FASTA.

L'alphabet est pondéré avec des scores de correspondance (*match*), de discordance (*mismatch*), d'insertion et de suppression pour chaque caractère. Ces pondérations servent à guider l'algorithme lorsque la distribution des probabilités d'alignement de caractères est connue au préalable. Ne disposant pas de telles connaissances *a priori*, la seule hypothèse que nous avons émise est que les voyelles sont plus susceptibles d'être alignées entre-elles et qu'il en est de même pour les consonnes. Tous les caractères reçoivent la même pondération pour l'insertion et la délétion.

Le tableau 8.2 donne un exemple de résultat fourni par `ALPHAMALIG` pour l'alignement des quatre variantes myriadisées correspondant au français « gâteau de carottes » : `Galrìewleküeche`, `Galerieblekü ecba`, `Galerewlekùche`, `Galrìawlaküacha`.

À partir des alignements produits, nous avons extrait des motifs de substitution contraints par différents degrés de rigidité, c'est-à-dire forçant ou non l'alignement des contextes gauche et droit.

158. Le code source d'`ALPHAMALIG` est disponible à cette url : http://alggen.lsi.upc.es/recerca/align/alphamalig/intro-alphamalig.html.
159. Le format FASTA est utilisé « pour stocker des séquences biologiques de nature nucléique ou protéique », voir : https://fr.wikipedia.org/wiki/FASTA_(format_de_fichier).

^	G	A	L	-	R	**Ì**	**E**	**W**	L	**E**	K	**Ü**	**E**	C	H	**E**	$	(1)
^	G	A	L	**E**	R	**I**	**E**	**B**	L	**E**	K	**Ü**	**E**	C	H	**A**	$	(2)
^	G	A	L	**E**	R	-	**E**	**W**	L	**E**	K	**Ù**	-	C	H	**E**	$	(3)
^	G	A	L	-	R	**Ì**	**A**	**W**	L	**A**	K	**Ü**	**A**	C	H	**A**	$	(4)

Tableau 8.2 – Alignement de quatre variantes du mot alsacien signifiant « gâteau de carottes ».

Nous extrayons ainsi trois ensembles de règles forçant soit la correspondance des contextes gauches (L), soit celle des contextes droits (R), ou soit celle des deux (L + R). Nous n'avons considéré qu'une fenêtre d'un seul caractère autour du motif substitué.

Les caractères ^ et $, représentent respectivement le début et la fin d'un mot, et sont interprétés comme des éléments de contexte. Les règles sont extraites pour chaque paire de variantes alignées. Par exemple, à partir de la paire formée par les variantes (1) et (2) présentées dans la figure 8.2, quatre règles L + R sont extraites :

- LR ↔ LER ;
- RÌE ↔ RIE ;
- EWL ↔ EBL ;
- HE$ ↔ HA$.

Ces quatre règles permettent de déduire huit règles forçant uniquement les contextes gauches et droits (par exemple, à gauche : L ↔ LE ; RÌ ↔ RI ; etc.).

Il est à noter que ne cherchons pas ici à trouver des règles de normalisation, chaque règle de substitution peut donc être utilisée dans les deux sens qui sont considérés comme également fréquents.

Des 215 paires de variantes myriadisées pour l'alsacien, nous avons ainsi extrait 227 règles forçant le contexte gauche seulement, 186 le droit, et 213 règles forçant les contextes gauche et droit (voir l'annexe D).

8.3.2 Identification des variantes et filtrage

Étant donné un vocabulaire de mots connus V_{lookup}, l'identification des variantes potentielles d'un mot hors vocabulaire consiste en trois étapes :

1. un filtrage préliminaire facultatif : si le mot inconnu est identifié comme un nom propre connu du lexique, il est ignoré ;
2. l'application de règles correspondantes : pour chaque ensemble de règles, L + R, L et R, utilisés dans cet ordre, le sous-ensemble de règles s'appliquant au mot hors vocabulaire original est identifié puis classé par ordre décroissant de fréquence d'utilisation des règles. A partir de ce sous-ensemble, nous appliquons sur le mot cible chaque *combinaison* possible de règles, ce qui signifie que si trois règles A, B, C s'appliquent, les séquences de règles {A}, {B}, {C}, {A ; B}, {A ; C}, {B ; C} et {A ; B ; C} sont appliquées, générant des variantes candidates ;
3. au fur et à mesure que les variantes candidates sont générées, nous vérifions si elles sont présentes dans le vocabulaire de recherche (V_{lookup}). Si c'est le cas, la recherche s'arrête et la paire formée par le mot hors vocabulaire et la variante identifiée est considérée comme une nouvelle paire de variantes.

Bien que cette méthode systématique génère une grande quantité de bruit et de graphies fantaisistes, le filtrage opéré grâce à V_{lookup} conduit le mot cible à n'être apparié qu'avec des variantes candidates réelles.

Étant donné qu'une partie des mécanismes de variation dialectale et graphique sont similaires à certaines des règles de flexion morphosyntaxique (comme la variation selon le genre, le nombre, la conjugaison ou la déclinaison), la légitimité de la variante identifiée doit être vérifiée en contexte. Nous illustrons ces cas lors de l'analyse des paires générés pour l'alsacien dans la section 8.4.3.

8.4 Application des règles de transposition

8.4.1 Ressources utilisées

Lors des expérimentations menées pour l'article de RANLP 2019, nous avons utilisé les deux corpus suivants :

- Le `Corpus myriadisé`, $C_{Myriadise}$, annoté par des participants contenant alors 9 282 tokens (439 phrases), annoté avec le jeu d'étiquettes (Petrov *et al.*, 2012), enrichi de la catégorie ADP+DET (contraction d'un déterminant et d'une préposition). Il est disponible sous licence CC BY-NC-SA.
- Le `Corpus annoté pour les dialectes alsaciens` (Bernhard *et al.*, 2018a), C_{Trad}, annoté « traditionnellement » par des linguistes avec le jeu d'étiquettes décrit ci-dessus, étendu avec les catégories EPE (épenthèse) et MOD (auxiliaire modal) (Bernhard, 2018b). Le corpus contient 12 570 mots (533 phrases) et est disponible sous licence CC BY-SA [160]. Il a été annoté manuellement par des linguistes experts.

Le corpus résultant de la concaténation des deux corpus, C_{Concat}, a été utilisé pour les expériences suivantes. Nous avons effectué une validation croisée sur quatre subdivisions (80 % utilisé pour l'entraînement, $C_{Concat}80$, 20 % pour l'évaluation, $C_{Concat}20$).

Les lexiques utilisés sont :

- Les lexiques fournis par D. Bernhard : L_{MO} (lexique de mots-outils (Bernhard et Ligozat, 2013)) et L_{gsw} qui est formé par diverses entrées issues de (i) l'Office pour Lexiques bilingues de la langue et de la culture alsaciennes (OLCA), (ii) du dictionnaire établi par l'Association Culture et Patrimoine d'Alsace (ACPA), et (iii) d'un dictionnaire multilingue français-allemand-alsacien (Adolf, 2006). Le lexique L_{gsw} contient plusieurs variantes graphiques pour certaines entrées. ;
- le LEXICON OF PLACE NAMES IN THE ALSATIAN DIALECTS (Bernhard, 2018a), distribué sous licence CC BY-SA [161], qui contient des noms de lieux en alsacien.

8.4.2 Évaluation de la méthodologie

L'identification des paires de variantes potentielles dépend des conditions initiales de l'expérience, c'est-à-dire du corpus, et éventuellement du lexique utilisé pour entraîner le modèle au préalable, nous présentons deux expériences dans lesquelles ces paramètres varient.

160. Voir : https://zenodo.org/record/2536041#.Xx1rcBHgq-o.
161. Voir https://zenodo.org/record/1404873#.X0Y0yBngq-o, juin 2020.

Pour chaque expérience, nous extrayons du corpus d'apprentissage le vocabulaire *VC_lookup* et du lexique externe, le vocabulaire *VL_lookup*. L'ensemble de ces deux vocabulaires forme *V_lookup*.

Nous utilisons l'ensemble de règles présenté dans la section 8.3.1.

Si la longueur du mot cible est inférieure ou égale à quatre caractères (^ et $ exclus), seules les règles L + R sont appliquées : il a été observé dans les tests préliminaires que les mots plus courts étaient plus susceptible de générer des appariements erronées comme *das* (déterminant)/*dass* (conjonction de subordination) ou *dien* (auxiliaire)/*dene* (déterminant). De plus, nous forçons les variantes candidates à avoir la même casse que le mot cible.

Une fois que les paires de variantes ont été identifiées comme telles, les mots hors vocabulaire sont remplacés par la variante connue et le modèle pré-entraîné est appliqué sur le corpus transposé. Une fois le corpus étiqueté, les mots transposés sont remplacés par leur forme originale.

Conditions initiales non contraintes

Par « non contraintes », nous entendons ici que nous nous plaçons dans des conditions réelles, les corpus d'entraînement et d'évaluation étant tous deux extraits du corpus concaténé C_{Concat} contenant plusieurs variantes.

Notre premier modèle est donc construit à partir de $C_{Concat}80$ (17 136 mots) et évalué sur $C_{Concat}20$ (4 374 mots) avant et après réalisation de la transposition. Après l'application des trois ensembles de règles, en utilisant à la fois les vocabulaires extraits de $C_{Concat}80$ et des lexiques L_{MO} et L_{gsw} pour la recherche, 56 nouvelles paires de variantes ont été découvertes et le même nombre de mots a été transposé.

	Avant transposition	**Après transposition**
Exactitude	85,91±0,004 %	86,42±0,007 %
% mots hors vocabulaire	24 %	22 %

Tableau 8.3 – Exactitude du modèle entraîné sur un corpus multi-variant et évalué sur un corpus multi-variant avant et après transposition.

La proportion de mots hors vocabulaire a diminué d'environ 2 points, ce qui se traduit par une amélioration des performances d'étiquetage de 0,5 point (voir tableau 8.3). Ce faible impact n'est pas étonnant, étant donné que les performances sur les mots du vocabulaire sont environ 10 points plus élevées que sur les mots hors vocabulaire dans cette configuration. Compte tenu de la taille de nos corpus, on pourrait attendre d'une réduction du nombre de mots hors vocabulaire de 100 qu'elle améliore les résultats d'environ 0,2 point.

Conditions initiales contraintes

Par « contraintes », nous entendons que les corpus utilisés pour l'entraînement et l'évaluation ont été soigneusement choisis de manière à contenir chacune des variantes spécifiques de l'alsacien. Dans ce qui suit, nous comparons des configurations homogènes et hétérogènes, dans lesquelles les corpus d'apprentissage et d'évaluation contiennent dans un cas des variantes identiques et dans l'autre cas des variantes distinctes de l'alsacien.

Pour mettre en évidence l'effet de notre méthodologie dans un contexte hétérogène, nous avons scindé manuellement C_{Concat} en deux sous-corpus C_{Nord} (4 880 mots) et C_{Sud} (7 690 mots) en nous basant sur les fréquences des terminaisons -e et -a des noms communs, qui sont respectivement spécifiques des variantes nord et sud. Le résultat de ces expériences est présenté dans le tableau 8.4.

	$C_{Nord}20$		$C_{Sud}20$	
$C_{Nord}80$			**Avant transposition**	**Après transposition**
Exactitude	85,31±0,004 %		71,44±0,006 %	**75,20±0,007 %**
mots hors vocabulaire	40 %		54 %	52 %
$C_{Sud}80$	**Avant transposition**	**Après transposition**		
Exactitude	78,80±0,005 %	**80,92±0,004 %**	86,42±0,007 %	
mots hors vocabulaire	51 %	48 %	29 %	

TABLEAU 8.4 – Exactitude du modèle entraîné sur un corpus mono-variante et évalué sur un corpus mono-variante avant et après transposition.

À nouveau, les meilleurs résultats sont obtenus lorsque les corpus d'entraînement et d'évaluation sont de la même variante avec une exactitude d'étiquetage de 85,31±0,004 % sur le corpus du bas alémanique du nord $C_{Nord}20$ et 86,42±0,007 % sur le corpus du bas alémanique du sud $C_{Sud}20$.

Ce que ces expériences mettent en revanche en valeur, c'est que dans cette configuration mono-variante, la transposition a un effet plus important : les performances augmentent de 2,1 points lorsque l'outil est entraîné avec le corpus $C_{Sud}80$ et évalué sur le corpus $C_{Nord}20$ après transposition, et de 3,8 points lorsque l'outil est entraîné avec le corpus $C_{Nord}80$ et évalué sur le corpus $C_{Sud}20$ après transposition.

L'impact observé de la méthode proposée dépend de plusieurs facteurs : la taille du corpus d'entraînement et par extension du vocabulaire *V_lookup* utilisé pour chercher des correspondances potentielles. Il dépend également du différentiel de variation entre le corpus d'entraînement et le corpus d'évaluation.

Cette expérience montre que les performances d'un outil entraîné sur un corpus donné peuvent être améliorées en modifiant temporairement le corpus sur lequel il est appliqué pour le faire correspondre au vocabulaire avec lequel l'outil a été entraîné.

8.4.3 Évaluation manuelle de la ressource

Les diverses expériences décrites ci-dessus ont abouti à la création d'une nouvelle ressource contenant 876 paires de variantes. 400 d'entre elles ont été identifiées grâce à l'appariement d'un mot hors vocabulaire avec un mot du corpus de formation et 476 avec un mot du lexique complémentaire utilisé.

La taille de la ressource ainsi générée dépend de la taille des corpus et du lexique de recherche, et du nombre de règles. Appliquer l'outil d'annotation avec la méthodologie décrite à tout nouveau texte inédit peut augmenter le nombre de paires de variantes identifiées.

Un sous-ensemble de 60 de ces paires de variantes générées automatiquement a été soumis à un enseignant alsacien, familier des variantes dialectales et graphiques. Les paires lui ont été présentées en contexte. Nous évaluons ici la précision de la méthode permettant d'identifier des paires de graphies parmi deux vocabulaires, pas son rappel.

Parmi les 60 paires examinées :

- 30 paires correspondent effectivement à des variantes graphiques de mêmes entités lexicales,
- 13 paires sont des paires de flexions, par exemple :
  - [*ìhm* (pronom datif) – *ìrhem* (pronom génitif)],
  - [*kált* (adjectif féminin) – *kálte* (adjectif masculin)],
  - [*wùrd* (auxiliaire au futur) – *wärd* (auxiliaire au conditionnel)], etc.

  Dans ces cas de correspondance erronées, les variantes présentent la même catégorie morphosyntaxique étant donné le jeu d'annotation choisi.
- 10 paires sont le fruit d'une correspondance erronée que nous avons réussi à corriger en effectuant les ajustements décrits dans la section 8.4, c'est-à-dire (i) en forçant la correspondance de la casse entre variantes potentielles, (ii) en ne considérant comme candidats que les mots dont la longueur est égale ou supérieure à quatre caractères.
- 7 paires sont le fruit de correspondances erronées que nous ne sommes pas encore parvenue à filtrer, par exemple :
  - [*kräfti* (« fortement », adverbe) – *kräftiger* (« plus fort », adjectif)],
  - [*mien* (« mon », pronom) – *meine* (« croire », verbe)], etc.

  Dans ces cas de correspondances erronées, les variantes sont de catégorie morphosyntaxique différentes, et leur transposition porte préjudice aux performances de l'outil d'annotation.

Ces résultats montrent qu'il est pour l'instant nécessaire de valider manuellement la légitimité des variantes appariées. Cette tâche peut elle-même être myriadisée à condition d'avoir accès au contexte d'apparition des deux éléments, ce qui ne pose pas de problème dans notre cas, les variantes étant ajoutées sur des segments en contexte.

Par construction, les paires nouvellement générées ne permettent pas d'obtenir de nouvelles règles de génération. Elles fournissent néanmoins des informations sur la fréquence des motifs de substitution.

Par ailleurs, les paires de variantes erronées filtrées manuellement peuvent par la suite être utilisées comme contre-exemples pour entraîner un classifieur de variantes (comme, par exemple, décrit dans (Barteld, 2017)).

8.5 Conclusion

Nous avons montré à travers diverses expériences l'intérêt que présente la ressource myriadisée que constituent les paires de variantes graphiques. Cette ressource, dont la production ne nécessite aucune formation ni expertise, est une ressource facile à myriadiser.

Nous avons montré à travers diverses expériences le double intérêt que présente cette ressource :

1. Grâce aux règles de substitutions extraites automatiquement, il est possible d'identifier des variantes graphiques et de faire ainsi baisser la proportion de mots hors vocabulaire d'un corpus et d'améliorer les performances d'un outil d'annotation sur celui-ci. Dans le cadre de langues non standardisées dans lequel nous nous trouvons cela permet de compenser dans une certaine mesure la dispersion des graphies possible pour un mot donné. L'effet

de la méthodologie présentée est d'autant plus marqué que les corpus d'entraînement et d'application diffèrent, par exemple lorsqu'un outil entraîné sur le bas-rhinois est évalué sur un corpus du haut-rhinois et inversement.

2. L'application des règles permet de générer de nouvelles paires de variantes qui viennent s'ajouter à la ressource myriadisée. Comme nous l'avons montré, la génération automatique de variantes telle que nous l'avons conçue conduit à des appariements erronés. Comme la tâche de production de variantes, nous pensons néanmoins que la validation de paires générées automatiquement est une tâche qui pourrait être aisément myriadisée, ce qui permettrait d'affiner l'algorithme génératif.

L'originalité de la méthode que nous proposons est qu'elle s'auto-alimente : chaque nouveau texte examiné est une source d'appariements de variantes potentielles. Cela est particulièrement utile dans les contextes dans lesquels nous nous trouvons où les ressources sont peu nombreuses mais proviennent de sources variées. La variation est ici prise en charge uniquement au moment du traitement automatisé : l'application de règles issues de la myriadisation permet d'amenuiser la variabilité entre corpus sans forcer une quelconque standardisation, et sans qu'aucune connaissance préalable sur le texte à annoter ne soit nécessaire.

Conclusion

L'évaluation présentée dans le chapitre 6 met en avant d'une part l'importance d'établir un lien durable avec les communautés linguistiques participantes pour les motiver à contribuer, et d'autre part l'intérêt de la démarche de myriadisation une fois ce lien établi, les ressources produites étant satisfaisantes en termes de qualité.

A contrario, nous expliquons le relatif insuccès des plateformes développées pour les créoles guadeloupéen et mauricien, outre la problématique de la disponibilité des ressources brutes que nous avons déjà commentée, par le manque de temps que nous leur avons consacré. Ces expériences ont été menées avec le soutien d'étudiantes dont l'implication n'a pas dépassé les six mois, et il nous apparaît que la promotion de ces plateformes requiert un investissement au long cours. Alors que dans le cas de l'alsacien, nous sommes parvenue dans une certaine mesure à nous intégrer à la communauté des internautes locuteurs, c'est le lien avec une personne ressource qui nous a fait défaut dans les deux autre cas.

La myriadisation, dans les situations où nous l'avons présentée, permet de produire des ressources de qualité en limitant la dépendance au travail de linguistes. En contrepartie, un effort de communication, de pédagogie et de dialogue doit être mis en place.

Dans le chapitre 7, nous avons tiré parti des ressources myriadisées et mises à notre disposition pour entraîner différents modèles d'annotation en parties du discours pour l'alsacien. Dans un premier temps, nous avons entraîné `MElt` (Denis et Sagot, 2012) afin de tirer parti de l'existence de lexiques additionnels. Dans un second temps, nous avons reproduit une expérience au cours de laquelle des plongements lexicaux (Magistry *et al.*, 2018) sont entraînés afin de compenser la petite taille des corpus d'entraînement.

Les expériences que nous avons menées ne nous permettent pas réellement de conclure à une meilleure efficacité d'aucune des deux méthodes. Dans les deux cas, les meilleurs résultats exactitude sur C_{Ref} ($87,33$ % et $88,19$ % d'exactitude pour les deux méthodes respectivement) ont été obtenus grâce à des ressources additionnelles qui ne sont pas distribuées sous une licence claire. Cela est regrettable car nous ne pouvons pas garantir la reproductibilité de cette recherche.

Le dialogue avec les participants, le manque de ressources représentatives des variantes en usage et l'analyse des performances des différents outils par variante nous ont poussée à dépasser le cadre conventionnel de la myriadisation : nous avons fait produire aux participants des ressources brutes, imaginées non pas pour documenter les pratiques, mais bien pour être intégrées à des chaînes de traitement de TAL. Nous avons présenté dans le chapitre 8 une façon de tirer parti de cette ressource.

De nombreuses communautés linguistiques étant concernées par les phénomènes de variation qui nous ont occupée, nous espérons que cette recherche saura encourager les chercheurs et

chercheuses de notre discipline à impliquer les locuteurs dans la production de ressources. Il nous semble en effet que sans cela, le développement d'outils sera voué à se limiter aux pratiques les plus standardisées, excluant *de facto* une partie des internautes.

Conclusion

Les locuteurs au cœur des ressources langagières

Au cours de ce travail, nous avons développé des plateformes de myriadisation de ressources linguistiques pour des langues non standardisées et avons montré l'intérêt que ces ressources présentent pour le développement d'outils de traitement automatique.

Au-delà de la dimension *pratique* de l'implication des locuteurs pour la construction de ressources linguistiques, nos travaux nous ont convaincue de son caractère *nécessaire*. Pour les langues dont la pratique écrite est récente, il n'est pas envisageable de développer des outils correspondant à la pratique réelle des locuteurs, collectivement détenteurs de l'information linguistique, sans les placer au cœur de la construction de ressources.

Notre première ambition a en effet été d'utiliser la myriadisation pour pallier le manque de moyens humains et financiers nécessaires pour « doter » une langue : la myriadisation se présentait alors à nos yeux comme une solution économique au retard technologique accusé par certaines langues.

Les premières expériences menées, concernant la myriadisation d'annotations en parties du discours pour l'alsacien, ont permis de valider l'hypothèse formulée dans l'introduction : « il n'y a pas de raison que le succès d'une entreprise participative (en termes de qualité des ressources produites) dépende de la langue à laquelle elle est appliquée ». La quantité et la qualité des annotations produites sur `Bisame` sont en effet satisfaisantes et nous ont permis de mener diverses expériences d'apprentissage supervisé par la suite.

Le dialogue qui s'est engagé avec les participants lors de cette expérience a en revanche fait émerger des difficultés mais aussi de nouvelles opportunités de myriadisation liées au caractère non standardisé de l'alsacien. La présence de variation rend malaisées et parfois impossibles la description, la construction et l'exploitation des ressources telles que nous avons l'habitude de les mener.

Les obstacles principaux identifiés quant à la dimension participative du projet incluent la faible couverture des ressources disponibles et la difficulté à préparer ces ressources pour l'annotation en prenant en compte toutes les pratiques scripturales en usage et dans un contexte où la « normalisation » n'a pas de sens. Par ailleurs, l'inconfort à contribuer sur une variante dialectale ou graphique qui n'est pas la leur a découragé certains participants, et conduit l'un d'entre eux à nous envoyer un de ses textes pour alimenter la plateforme.

En parallèle de ce dialogue, nous nous confrontions à la difficulté d'adapter à des langues non standardisées les méthodes de traitement automatique imaginées pour des productions langagières canoniques : la variation observée dans nos corpus conduit à une dégradation des performances des outils entraînés. En cause, une proportion importante de mots hors vocabulaire dans les corpus à annoter, et une couverture insuffisante des lexiques complémentaires à disposition.

Ces difficultés nous ont poussée à envisager la myriadisation différemment, en ne considérant plus les locuteurs comme un ensemble uniforme de contributeurs dont les efforts cumulés permettent de produire un travail habituellement effectué par un linguiste, mais comme un ensemble de détenteurs de connaissances *complémentaires*. Dans ce second paradigme, chaque participation a une valeur propre et c'est la variété des profils des participants qui confère un intérêt à la démarche participative.

C'est dans cette perspective que nous avons entrepris la myriadisation de corpus textuels et de

variantes graphiques. Cette dernière ressource ne peut être obtenue sans l'investissement des locuteurs, et présente l'avantage notable d'être une ressource de complexité faible, au sens où elle ne requiert pas de compétence linguistique particulière.

Afin d'assurer la réplicabilité de nos expérimentations, nous avons instancié les plateformes développées pour deux langues autres que l'alsacien : le créole guadeloupéen et le créole mauricien. L'adaptation n'a pas posé de problème technique, au sens où les choix linguistiques propres à notre méthodologie ont pu être transposés. Les résultats moindres obtenus pour les deux langues créoles sont le fait des contextes formels des Master 1 dans lesquels les adaptations ont été faites : au terme des encadrements de mémoire, nous n'avons pas pu poursuivre les efforts de communication nécessaires, faute de contacts suffisants avec les communautés linguistiques concernées.

Nous espérons néanmoins que la présentation de trois contextes linguistiques bien distincts a permis de montrer que l'enjeu de l'intégration des productions linguistiques variées aux chaînes de traitement de TAL est réel et mérite qu'on s'y intéresse. Notre travail gagnerait à être poursuivi en tâchant de tirer le meilleur des deux démarches de myriadisation menées. L'annotation en séquence implémentée sur la première plateforme P_Ann a par exemple été mieux accueillie par les participants que l'annotation par étiquette implémentée sur la plateforme P_Prod_Var. En revanche, la seconde plateforme a permis de toucher une nouvelle partie de la communauté linguistique.

De plus, les ressources que nous produisons sont dynamiques, et l'interruption de cette thèse survient alors que la collecte est en cours et que nous n'avons pas encore exploité l'ensemble des ressources myriadisées. Nous pensons notamment que l'intégration des variantes graphiques que nous avons présentée n'est qu'une des multiples manières de tirer parti de cette ressource riche.

Vers un réel *échange* entre locuteurs et chercheurs

Les enquêtes menées sur les communautés linguistiques alsacienne et mauricienne comprenaient une section sur les facteurs de motivation pour les participants que nous n'avons pas présentée dans le manuscrit, mais qui constitue une partie des perspectives de ce travail. La difficulté que nous avons eue à motiver les locuteurs nous portent à penser que nous conservons aux yeux des participants un statut de demandeurs, alors que nous pourrions profiter du dialogue établi pour créer un véritable échange entre locuteurs et chercheurs.

Parmi les réponses apportées par les participants à la question « Vous aimeriez que votre participation à la création de ressources en ligne vous permette... » ressortent deux réponses principales : « d'apprendre des choses en général » (51 % des réponses sur les deux langues) et « d'améliorer mon [alsacien|créole mauricien] » (50 % des réponses sur les deux langues). Il nous semble que la première réponse mérite d'être examinée et son efficacité d'être testée, et il est certain que les directions possibles de poursuite de cette recherche incluent en effet l'utilisation de données myriadisées pour produire du matériel pédagogique pour ces langues.

Une des autres réponses qui nous a été suggérée par une trentaine de répondants de l'enquête sur l'alsacien concerne l'envie de participer à la valorisation de la langue et à sa transmission. La survie d'une langue dépendant avant tout de sa pratique quotidienne, nous avons amorcé un travail visant à combiner l'encouragement à la transmission et la production de ressources linguistiques.

En parallèle des travaux que nous avons présentés dans ce manuscrit, nous avons en effet eu la possibilité au cours de deux *hackathons* organisés dans le cadre de l'action COST EnetCollect [162] de développer un jeu destiné à collecter des ressources linguistique. L'envie de développer un jeu est née d'une réflexion autour du caractère artificiel et parfois insatisfaisant des fonctionnalités ludiques des plateformes de myriadisation. Le prototype de jeu issu du premier *hackathon* a fait l'objet d'une publication à LTC 2019 (Millour *et al.*, 2019).

FIGURE 1 – Captures du jeu *Katana, the Game of the Lost Words*. Si le mot *crann* (« arbre », en irlandais) est renseigné (écran de gauche), les arbres apparaissent dans le jeu (écran de droite)

Le jeu développé, *Katana, the Game of the Lost Words* (Katana, le jeu des mots perdus), est un jeu de rôles (*Role Play Game* (RPG)) classique, mais où la progression du joueur est soumise à des épreuves linguistiques intégrées, comme la myriadisation d'entrées lexicales. Le jeu a été pensé pour être joué en collaboration, un apprenant la langue cible pouvant demander conseil à un de ses proches, lui-même locuteur de la langue, pour la résolution de ces épreuves. Pour l'instant, le jeu a été développé en anglais et la première langue cible choisie pour les données myriadisées est l'irlandais, comme visible sur les captures d'écran présentées en figure 1 et 2.

FIGURE 2 – Myriadisation de lexique dans *Katana, the Game of the Lost Words* : une des épreuves consiste à produire du vocabulaire dans la langue cible.

La traduction du jeu et le test dans des conditions réelles ont été interrompus en raison de la pandémie survenue en 2020, mais la poursuite de ce travail nous paraît tout à fait prometteuse.

162. Voir : https://www.enetcollect.net, juillet 2020.

Enfin, nous pensons qu'une réflexion reste à mener au sujet de la distribution des ressources myriadisées. En particulier, et comme défendu par Prys (2019), apposer sur les ressources produites des licences empêchant les industriels d'en faire un usage commercial n'est pas un service rendu à la communauté linguistique. De manière générale, les conditions permettant de valoriser au mieux les efforts déployés par les chercheurs et par les locuteurs peuvent sans aucun doute être affinées : nous espérons que le travail amorcé dans cette thèse pourra être poursuivi et étendu dans ce sens.

Bibliographie

Andrea Abel et Christian M Meyer : The dynamics outside the paper : user contributions to online dictionaries. *In Proceedings of the 3rd eLex conference Electronic lexicography in the 21st century : thinking outside the paper*, pages 179–194, Tallinn, Estonia, octobre 2013.

Paul Adolf : *Dictionnaire comparatif multilingue : français-allemand-alsacien-anglais*. Midgard, Strasbourg, France, 2006. ISBN 2-84512-038-9.

Željko Agić, Anders Johannsen, Barbara Plank, Héctor Alonso Martínez, Natalie Schluter et Anders Søgaard : Multilingual projection for parsing truly low-resource languages. *Transactions of the Association for Computational Linguistics*, 4:301–312, 2016.

Željko Agić et Ivan Vulić : JW300 : A wide-coverage parallel corpus for low-resource languages. *In Proceedings of the the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3204–3210, Florence, Italie, juillet 2019. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/P19-1310.

Iñaki Alegria, Unai Cabezon, Unai Fernández de Betoño, Gorka Labaka, Aingeru Mayor, Kepa Sarasola et Arkaitz Zubiaga : *Reciprocal Enrichment Between Basque Wikipedia and Machine Translation*, pages 101–118. Springer Berlin Heidelberg, 02 2013. ISBN 9783642350849.

Marc Alexander et Mark Davies : The Hansard Corpus 1803-2005, 2015. URL http://eprints.gla.ac.uk/115023/.

Mathieu Andro et Imad Saleh : Digital Libraries and Crowdsourcing : A Review. *In* Samuel Szoniecky et Nasreddine Bouhaï, éditeurs : *Collective Intelligence and Digital Archives : Towards Knowledge Ecosystems*, pages 135–162. ISTE, 2017. URL https://hal.archives-ouvertes.fr/hal-01436766.

Jannis K Androutsopoulos : Grammaticalization in young people's language : The case of German. *Belgian journal of linguistics*, 13(1):155–176, 1999.

Ron Artstein et Massimo Poesio : Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596, 2008.

Mathieu Avanzi, Cécile Barbet, Julie Glikman et Julie Peuvergne : Présentation d'une enquête pour l'étude des régionalismes du français. *In Proceedings of the the SHS Web of Conferences*, volume 27, page 03001, Tours, France, juillet 2016. EDP Sciences.

Mathieu Avanzi et Elisabeth Stark : A crowdsourcing approach to the description of regional variation in French object clitic clusters. *Belgian Journal of Linguistics*, 31(1):76–103, 2017.

Adrien Barbaresi : Challenges in web corpus construction for low-resource languages in a post-BootCaT world. *In Proceedings of the 6th Language & Technology Conference : Human Language Technologies as a Challenge for Computer Science and Linguistics (LTC 2013)*, pages 69–73, Poznań, Pologne, décembre 2013. URL https://halshs.archives-ouvertes.fr/halshs-00919410.

Adrien Barbaresi et Kay-Michael Würzner : For a fistful of blogs : Discovery and comparative benchmarking of republishable German content. *In Proceedings of the NLP4CMC workshopn KONVENS 2014*, KONVENS 2014, NLP4CMC workshop proceedings, pages 2–10, Hildesheim, Germany, octobre 2014. Hildesheim University Press. URL https://hal.archives-ouvertes.fr/hal-01083750.

Alistair Baron et Paul Rayson : Vard 2 : A tool for dealing with spelling variation in historical corpora. *In* Aston University, éditeur : *Proceedings of the 2008 Postgraduate Conference in Corpus Linguistics*, Birmingham, Royaume-Uni, mai 2008.

Alistair Baron et Paul Rayson : Automatic standardisation of texts containing spelling variation : How much training data do you need ? *In Proceedings of the 2009 Corpus Linguistics Conference*, Liverpool, Royaume-Uni, 2009.

Marco Baroni, Silvia Bernardini, Adriano Ferraresi et Eros Zanchetta : The WaCky wide web : a collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation*, 43(3):209–226, septembre 2009. ISSN 1574-0218. URL https://doi.org/10.1007/s10579-009-9081-4.

Marco Baroni et Motoko Ueyama : Building general-and special-purpose corpora by web crawling. *In Proceedings of the 13th NIJL international symposium, language corpora : Their compilation and application*, pages 31–40, Tokyo, Japon, mars 2006.

Corinne Barre et Mélanie Vanderschelden : *L'enquête "étude de l'histoire familiale" de 1999 - Résultats détaillés*. INSEE, Paris, France, 2004. ISBN 978-2-11-068285-7.

Fabian Barteld : Detecting spelling variants in non-standard texts. *In Proceedings of the Student Research Workshop, 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2017)*, Valence, Espagne, mai 2017.

Harmonie Begue : Développement de ressources langagières et d'outils de TAL pour le créole mauricien. Mémoire de Master 1 à Sorbonne Université, 2019.

Yassine Benajiba, Mona Diab et Paolo Rosso : Using Language Independent and Language Specific Features to Enhance Arabic Named Entity Recognition. *International Arab Journal of Information Technology (IAJIT)*, 6(5):463–471, 2009.

Martin Benjamin : Hard Numbers : Language Exclusion in Computational Linguistics and Natural Language Processing. *In Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japon, mai 2018. URL http://lrec-conf.org/workshops/lrec2018/W26/pdf/23{_}W26.pdf.

Vincent Berment : *Méthodes pour informatiser les langues et les groupes de langues "peu dotées"*. Thèse, Université Joseph-Fourier - Grenoble I, mai 2004. URL https://tel.archives-ouvertes.fr/tel-00006313.

Jean Bernabé : *La graphie créole*. Guides du CAPES de Créole, Ibis Rouge édition, 2001.

Delphine Bernhard : Lexicon of place names in the alsatian dialects, août 2018. URL https://doi.org/10.5281/zenodo.1404873.

Delphine Bernhard : Tokeniser for the Alsatian Dialects, décembre 2018. URL https://doi.org/10.5281/zenodo.2454993.

Delphine Bernhard, Pascale Erhart, Dominique Huck et Lucie Steiblé : Guide d'annotation morphosyntaxique pour les dialectes alsaciens. Guide d'annotation, LiLPa, Université de Strasbourg, 2016.

Delphine Bernhard, Pascale Erhart, Dominique Huck et Lucie Steiblé : Annotated corpus for the alsatian dialects. Guide d'annotation, LiLPa, Université de Strasbourg, 2018.

Delphine Bernhard et Anne-Laure Ligozat : Es esch fàscht wie Ditsch, oder net ? Étiquetage morphosyntaxique de l'alsacien en passant par l'allemand. *In Actes de TALARE 2013 : Traitement Automatique des Langues Régionales de France et d'Europe (TALN 2013)*, pages 209–220, Les Sables d'Olonne, France, juin 2013.

Delphine Bernhard, Anne-Laure Ligozat, Fanny MARTIN, Myriam Bras, Pierre Magistry, Marianne Vergez-Couret, Lucie Steible, Pascale Erhart, Nabil Hathout, Dominique Huck, Christophe Rey, Philippe Reynés, Sophie Rosset, Jean Sibille et Thomas Lavergne : Corpora with Part-of-Speech Annotations for Three Regional Languages of France : Alsatian, Occitan and Picard. *In 11th edition of the Language Resources and Evaluation Conference*, Miyazaki, Japan, mai 2018. URL https://hal.archives-ouvertes.fr/hal-01704806.

Delphine Bernhard et Lucie Steiblé : Quand l'oral se fait entendre à l'écrit : alignement de lexiques en l'absence de normalisation graphique. *In Actes de TALARE 2015 : Traitement Automatique des Langues Régionales de France et d'Europe (TALN 2015)*, Caen, France, juin 2015. URL https://hal.archives-ouvertes.fr/hal-01158489.

Mat Bettinson et Steven Bird : Developing a suite of mobile applications for collaborative language documentation. *In Proceedings of the 2nd Workshop on Computational Methods for Endangered Languages*, pages 156–164, Honolulu, Hawaï, mars 2017.

David Blachon, Elodie Gauthier, Laurent Besacier, Guy-Noël Kouarata, Martine Adda-Decker et Annie Rialland : Parallel Speech Collection for Under-resourced Language Studies Using the Lig-Aikuma Mobile Device App. *Procedia Computer Science*, 81:61–66, 2016. ISSN 1877-0509. URL http://www.sciencedirect.com/science/article/pii/S1877050916300448.

Alena Böhmová, Jan Hajič, Eva Hajičová et Barbora Hladká : The Prague Dependency Treebank : Three-Level Annotation Scenario. *In* Anne Abeillé, éditeur : *Treebanks : Building and Using Syntactically Annotated Corpora*. Kluwer Academic Publishers, 2001. URL http://ufal.mff.cuni.cz/pdt/Corpora/PDT_1.0/References/Czech_PDT.pdf.

Danushka Bollegala, Ryuichi Kiryo, Kosuke Tsujino et Haruki Yukawa : Language-Independent Tokenisation Rivals Language-Specific Tokenisation for Word Similarity Prediction. *In Proceedings of the 13th International Conference on Language Resources and Evaluation (LREC 2020)*, Marseille, France, mai 2020.

Olivier Bonami et Fabiola Henri : How complex is creole inflectional morphology ? *In Budapest : International Meeting of Morphology*, Budapest, Hongrie, mai 2010.

Miriam Bouzouita, Mónica Castillo et Enrique Pato : Dialectos del español. una nueva aplicación para conocer la variación actual y el cambio en las variedades del español. *Dialectologia : revista electrònica*, 1(20):61–83, 2018.

António Branco : We Are Depleting Our Research Subject as We Are Investigating It : In Language Technology, more Replication and Diversity Are Needed. *In Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japon, mai 2018. URL http://www.lrec-conf.org/proceedings/lrec2018/pdf/397.pdf.

António Branco, Nicoletta Calzolari et Khalid Choukri, éditeurs. *Proceedings of the Workshop on Research Results Reproducibility and Resources Citation in Science and Technology Proceedings*, 2016.

Jean-Jacques Brunner : *L'alsacien sans peine*. Assimil, 2001.

Thierry Burger-Helmchen et Julien Pénin : Crowdsourcing : définition, enjeux, typologie. *Management & Avenir*, 41(1):254–269, 2011.

Chris Callison-Burch et Mark Dredze : Creating speech and language data with Amazon's Mechanical Turk. *In Proceedings of the Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk (CSLDAMT 2010), Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology (NAACL 2010)*, pages 1–12, Los Angeles, CA, États-Unis, juin 2010. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/W10-0701.

Arnaud Carpooran : *Lortograf Kreol Morisien*. Phoenix, Mauritius : Akademi Kreol Morisien, 2011.

Paola Carrión Gonzalez et Emmanuel Cartier : Technological tools for dictionary and corpora building for minority languages : example of the French-based Creoles. *In Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012) (Workshop on Language Technology for Normalisation of Less-Resourced Languages (SALTMIL8/AfLaT2012))*, pages 47–53, Istanbul, Turquie, mai 2012.

Jon Chamberlain, Karën Fort, Udo Kruschwitz, Mathieu Lafourcade et Massimo Poesio : Using Games to Create Language Resources : Successes and Limitations of the Approach. *In* Iryna Gurevych et Jungi Kim, éditeurs : *The People's Web Meets NLP*, Theory and Applications of Natural Language Processing, pages 3–44. Springer Berlin Heidelberg, 2013. ISBN 978-3-642-35084-9. URL http://dx.doi.org/10.1007/978-3-642-35085-6_1.

Thierry Chanier, Céline Poudat, Benoît Sagot, Georges Antoniadis, Ciara R. Wigham, Linda Hriba, Julien Longhi et Djamé Seddah : The CoMeRe corpus for French : structuring and annotating heterogeneous CMC genres. *Journal for language technology and computational linguistics*, 29(2):1–30, 2014. URL https://halshs.archives-ouvertes.fr/halshs-00953507. Final version to Special Issue of JLCL (Journal of Language Technology and Computational Linguistics (JLCL, http ://jlcl.org/) : BUILDING AND ANNOTATING CORPORA OF COMPUTER-MEDIATED DISCOURSE : Issues and Challenges at the Interface of Corpus and Computational Linguistics (ed. by Michael Beißwenger, Nelleke Oostdijk, Angelika Storrer & Henk van den Heuvel).

Jacob Cohen : A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20(1):37–46, 1960.

K Bretonnel Cohen, Jingbo Xia, Christophe Roeder et Lawrence E Hunter : Reproducibility in natural language processing : a case study of two R libraries for mining PubMed/MEDLINE. *In Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*, page 6, Portorož, Slovénie , mai 2016. NIH Public Access.

K. Bretonnel Cohen, Jingbo Xia, Pierre Zweigenbaum, Tiffany Callahan, Orin Hargraves, Foster Goss, Nancy Ide, Aurélie Névéol, Cyril Grouin et Lawrence E. Hunter : Three Dimensions of Reproducibility in Natural Language Processing. *In Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japon, mai 2018. ISBN 979-10-95546-00-9.

Serge Colot et Ralph Ludwig : Guadeloupean and Martinican Creole. *In* Susanne Maria Michaelis, Philippe Maurer, Martin Haspelmath et Magnus Huber, éditeurs : *The survey of pidgin and creole languages.*, volume 2. Oxford University Press, 2013. URL http://apics-online.info/surveys/50.

Sarah Cooper, Dewi Bryn Jones et Delyth Prys : Crowdsourcing the Paldaruo Speech Corpus of Welsh for Speech Technology. *Information*, 10(8):247, 2019.

Eugenio Coseriu : *Probleme der strukturellen Semantik : Vorlesung gehalten im Wintersemester 1965/66*, volume 40. Narr, 1973.

Alix Cosquer, Richard Raymond et Anne-Caroline Prevot-Julliard : Observations of everyday biodiversity : A new perspective for conservation ? *Ecology and Society*, 17(4):1–15, 2012. URL http://hdl.handle.net/10535/8668.

Christopher Cox : Probabilistic tagging of minority language data : a case study using Qtag. *Language & Computers*, 71(1):213–231, 2010. ISSN 09215034. URL https://sites.ualberta.ca/{~}cdcox/PDF/Cox2010ProbabilisticTaggingMinorityLanguageData.pdfhttp://search.ebscohost.com/login.aspx?direct=true{&}db=afh{&}AN=51613244{&}lang=pt-br{&}site=ehost-live.

Danielle Crévenat-Werner et Edgar Zeidler : *Orthographe alsacienne - Bien écrire l'alsacien de Wissembourg à Ferrette*. Jérôme Do Bentzinger, 2008.

Danielle Crévenat-Werner et Edgar Zeidler : *Le système ORTHAL 2016 – Orthographe alsacienne*. Jérôme Do Bentzinger, 2016.

Raj Dabre, Aneerav Sukhoo et Pushpak Bhattacharyya : Anou Tradir : Experiences In Building Statistical Machine Translation Systems For Mauritian Languages – Creole, English, French. *In Proceedings of the 11th International Conference on Natural Language Processing*, pages 82–88, Goa, Inde, décembre 2014. NLP Association of India. URL https://www.aclweb.org/anthology/W14-5113.

Sandipan Dandapat, Priyanka Biswas, Monojit Choudhury et Kalika Bali : Complex Linguistic Annotation — No Easy Way out ! : A Case from Bangla and Hindi POS Labeling Tasks. *In Proceedings of the 3rd Linguistic Annotation Workshop (LAW III)*, ACL-IJCNLP '09, pages 10–18, Stroudsburg, PA, États-Unis, août 2009. ISBN 978-1-932432-52-7. URL http://dl.acm.org/citation.cfm?id=1698381.1698383.

Fabrice Delumeau : *Une description linguistique du créole guadeloupéen dans la perspective de la génération automatique d'énoncés*. Thèse de doctorat, Université de Nanterre - Paris X, 2006. URL https://halshs.archives-ouvertes.fr/tel-00169457/document.

Pascal Denis et Benoît Sagot : Exploitation d'une ressource lexicale pour la construction d'un étiqueteur morphosyntaxique état-de-l'art du français. *In Actes de Traitement Automatique des Langues Naturelles (TALN 2010)*, Montréal, Canada, juillet 2010. URL http://hal.inria.fr/inria-00521231.

Pascal Denis et Benoît Sagot : Coupling an annotated corpus and a lexicon for state-of-the-art pos tagging. *Lang. Resour. Eval.*, 46(4):721–736, décembre 2012. ISSN 1574-020X. URL http://dx.doi.org/10.1007/s10579-012-9193-0.

Marcel Diki-Kidiri : Comment assurer la présence d'une langue dans le cyberespace. *UNESCO. Retrieved December*, 31:2007, 2007.

Miguel Domingo, Mercedes Garcia-Martinez, Alexandre Helle, Francisco Casacuberta et Manuel Herranz : How Much Does Tokenization Affect Neural Machine Translation ? arXiv preprint arXiv :1812.08621, 2018.

Pieter Duijff, Frits van der Kuip, Hindrik Sijens et Willem Visser : User Contributions in the Online Dutch-Frisian Dictionary. *In Proceedings of the European Network of e-Lexicography (Enel) COST Action (WG1 meeting)*, Barcelone, Espagne, mars 2016. URL https://www.elexicography.eu/wp-content/uploads/2016/03/Kuip_User-Contributions-in-the-Online-Dutch-Frisian-Dictionary.pdf.

Christa Dürscheid et Elisabeth Stark : Sms4science : An international corpus-based texting project and the specific challenges for multilingual switzerland. *Digital Discourse : language in the new media*, pages 299–320, 2011.

Dagmara Dziedzic : Use of the Free to Play model in games with a purpose : the RoboCorp game case study. *Bio-Algorithms and Med-Systems*, 12(4):187 – 197, 2016. URL https://www.degruyter.com/view/journals/bams/12/4/article-p187.xml.

David M. Eberhard, Gary F. Simons et Fennig Charles D. : Ethnologue : Languages of the world. Twenty-third edition, 2019. URL http://www.ethnologue.com.

Iris Eshkol-Taravella : *Specification of linguistic annotations according to corpora : from newspaper to spoken corpora.* Habilitation à diriger des recherches, Université d'Orléans, octobre 2015. URL https://hal.archives-ouvertes.fr/tel-01250650.

Izaskun Etxeberria Uztarroz, Iñaki Alegría Loinaz, Mans Hulden et Larraitz Uria Garin : Learning to map variation-standard forms in basque using a limited parallel corpus and the standard morphology. *Procesamiento del Lenguaje Natural*, 52:13–20, 2014. ISSN 1135-5948.

Caterina Falbo : La transcription : une tache paradoxale, 2005.

Andri Imam Fauzi et Dwi Puspitorini : Dialect and Identity : A Case Study of Javanese Use in WhatsApp and line. *IOP Conference Series : Earth and Environmental Science*, 175:012111, jul 2018. URL https://doi.org/10.1088%2F1755-1315%2F175%2F1%2F012111.

Fabien Fenouillet, Jonathan Kaplan et Nora Yennek : Serious games et motivation. *In Actes de 4ème Conférence francophone sur les Environnements Informatiques pour l'Apprentissage Humain (EIAH'09), vol. Actes de l'Atelier "Jeux Sérieux : conception et usages"*, pages 41–52, Le Mans, France, juin 2009.

Darja Fišer, Aleš Tavčar et Tomaž Erjavec : sloWCrowd : a Crowdsourcing Tool for Lexicographic Tasks. *In Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 2014)*, Reykjavik, Islande, mai 2014. ISBN 978-2-9517408-8-4.

Antske Fokkens, Marieke van Erp, Marten Postma, Ted Pedersen, Piek Vossen et Nuno Freire : Offspring from Reproduction Problems : What Replication Failure Teaches Us. *In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL 2013)*, pages 1691–1701, Sofia, Bulgarie, août 2013. URL http://www.aclweb.org/anthology/P13-1166.

∀, Iroro Orife, Julia Kreutzer, Blessing Sibanda, Daniel Whitenack, Kathleen Siminyu, Laura Martinus, Jamiil Ali, Jade Abbott, Vukosi Marivate, Salomon Kabongo Kabenamualu, Musie Meressa Berhe, Espoir Murhabazi, Orevaoghene Ahia, Elan Biljon, Arshath Ramkilowan, Adewale Akinfaderin, Alp Oktem, Wole Akin, Ghollah Kioko et Abdallah Bashir : Masakhane – machine translation for africa. *In Proceedings of the "AfricaNLP" Workshop at the 8th International Conference on Learning Representations*, 03 2020.

Karën Fort : *Les ressources annotées, un enjeu pour l'analyse de contenu : vers une méthodologie de l'annotation manuelle de corpus.* Thèse de doctorat, Université Paris-Nord - Paris XIII, décembre 2012. URL https://tel.archives-ouvertes.fr/tel-00797760.

Karën Fort : *Collaborative Annotation for Reliable Natural Language Processing.* Focus series. ISTE Wiley, 2016.

Karën Fort : Experts ou (foule de) non-experts ? la question de l'expertise des annotateurs vue de la myriadisation (crowdsourcing). *Corela. Cognition, représentation, langage*, HS-21:12 p., 2017. URL http://journals.openedition.org/corela/4835. consulté le 23 juin 2020.

Karën Fort, Gilles Adda et Kevin Bretonnel Cohen : Amazon Mechanical Turk : Gold mine or coal mine ? *Computational Linguistics (editorial)*, 37(2):413–420, juin 2011. URL http://aclweb.org/anthology-new/J/J11/J11-2010.pdf.

Karën Fort, Bruno Guillaume et Nicolas Lefèbvre : Who wants to play Zombie ? A survey of the players on ZOMBILINGO. *In Proceedings of the 2017 Games4NLP Workshop - Using Games and Gamification for Natural Language Processing*, page 2, Valence, Espagne, avril 2017. URL https://hal.inria.fr/hal-01494043.

Karën Fort, Adeline Nazarenko et Sophie Rosset : Modeling the complexity of manual annotation tasks : a grid of analysis. *In Proceedings of the 24th International Conference on Computational Linguistics (COLING 2012)*, pages 895–910, Mumbai, Inde, décembre 2012.

Karën Fort et Benoît Sagot : Influence of Pre-annotation on POS-tagged corpus development. *In Proceedings of the 4th Linguistic Annotation Workshop (LAW IV)*, pages 56–63, Uppsala, Suède, juillet 2010. URL http://aclweb.org/anthology-new/W/W10/W10-1807.pdf.

Françoise Gadet : L'oralité ordinaire à l'épreuve de la mise en écrit : ce que montre la proximité. *Langages*, 208(4):113–129, 2017.

Marcos Garcia, Pablo Gamallo, Iria Gayo et Miguel A.Pousada Cruz : PoS-tagging the web in portuguese. National varieties, text typologies and spelling systems. *Procesamiento de Lenguaje Natural*, 53:95–101, 2014. ISSN 19897553. URL http://www.taln.upf.edu/pages/sepln2014/full{_}papers/edited{_}paper{_}21.pdf.

Dan Garrette, Jason Mielens et Jason Baldridge : Real-World Semi-Supervised Learning of POS-Taggers for Low-Resource Languages. *In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL-2013)*, ACL 2013, pages 583–592, Sofia, Bulgarie, août 2013.

Maristella Gatto : *Web as corpus : Theory and practice*. A&C Black, 2014.

David Geiger, Stefan Seedorf, Thimo Schulze, Robert C. Nickerson et Martin Schader : Managing the Crowd : Towards a Taxonomy of Crowdsourcing Processes. *In Proceedings of the 17th Americas Conference on Information Systems (AMCIS 2011)*, Detroit, MI, États-Unis, août 2011. URL http://aisel.aisnet.org/amcis2011_submissions/430.

Dirk Goldhahn, Thomas Eckart et Uwe Quasthoff : Building Large Monolingual Dictionaries at the Leipzig Corpora Collection : From 100 to 200 Languages. *In Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012)*, pages 759—-765, Istanbul, Turquie, mai 2012. European Language Resources Association (ELRA).

Mark Graham, Scott A Hale et Devin Gaffney : Where in the world are you ? Geolocation and language identification in Twitter. *The Professional Geographer*, 66(4):568–578, 2014.

Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin et Tomas Mikolov : Learning Word Vectors for 157 Languages. *In Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japon, mai 2018. European Language Resources Association (ELRA).

Adriana Guevara-Rukoz, Isin Demirsahin, Fei He, Shan-Hui Cathy Chu, Supheakmungkol Sarin, Knot Pipatsrisawat, Alexander Gutkin, Alena Butryna et Oddur Kjartansson : Crowdsourcing Latin American Spanish for low-resource text-to-speech. *In Proceedings of the the 12th Language Resources and Evaluation Conference*, pages 6504–6513, Marseille, France, mai 2020. European Language Resources Association. ISBN 979-10-95546-34-4. URL https://www.aclweb.org/anthology/2020.lrec-1.801.

Bruno Guillaume, Karën Fort et Nicolas Lefebvre : Crowdsourcing Complex Language Resources : Playing to Annotate Dependency Syntax. *In Proceedings of the 26th International Conference on Computational Linguistics (COLING 2016)*, Osaka, Japon, décembre 2016.

Lennart Guldbrandsson : Swedish Wikipedia surpasses 1 million articles with aid of article creation bot. Wikimedia blog, juin 2013. URL https://diff.wikimedia.org/2013/06/17/swedish-wikipedia-1-million-articles/. Consulté le 4 septembre 2020.

Antton Gurrutxaga, Igor Leturia, Eli Pociello, Iñaki San Vicente et Xabier Saralegi : Exploiting the Internet to build language resources for less resourced languages. *In Proceedings of the 7th SaLTMiL Workshop on Creation and use of basic lexical resources for less-resourced languages*, La Valette, Malte, mai 2010.

Ivan Habernal, Omnia Zayed et Iryna Gurevych : C4Corpus : Multilingual Web-size corpus with free license. *In Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*, pages 914–922, Portorož, Slovénie , mai 2016.

Benoît Habert : Portrait de linguiste(s) à l'instrument. *In* Guillot, Céline ;Heiden, Serge ;Prévost et Sophie ;, éditeurs : *À la quête du sens : études littéraires, historiques et linguistiques en hommage à Christiane Marchello-Nizia*, pages 124–132. ENS Éditions, 2006. URL https://halshs.archives-ouvertes.fr/halshs-00355997.

Benoit Habert : Construire ensemble des mémoires numériques durables : l'archivage numérique pérenne. *In* Manuel Zacklad et Khaldoun Zreik Madjid Ihadjadene, éditeur : *Actes de 13e Congrès International sur le Document Électronique (CIDE)*, pages 5–24, Paris, France, décembre 2010. Europia Productions. URL https://halshs.archives-ouvertes.fr/halshs-00991508.

Benoit Habert, Gilles Adda, Martine Adda-Decker, P Boula de Marëuil, Serge Ferrari, Olivier Ferret, Gabriel Illouz et Patrick Paroubek : Towards tokenization evaluation. *In Proceedings of the 1st International Conference on Language Resources and Evaluation (LREC 1998)*, volume 98, pages 427–431, Grenade, Espagne, mai 1998.

Claude Hagège : *Halte à la mort des langues*. Odile Jacob, 2000.

Xiaochuang Han et Jacob Eisenstein : Unsupervised domain adaptation of contextualized embeddings for sequence labeling. *In Proceedings of the the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4238–4248, Hong Kong, Chine, novembre 2019. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/D19-1433.

Jiri Hana, Anna Feldman et Chris Brew : A Resource-light Approach to Russian Morphology : Tagging Russian using Czech resources. *In Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP'04)*, pages 222–229, Barcelone, Espagne, juillet 2004. ACL.

Jirka Hana, Anna Feldman et Katsiaryna Aharodnik : A low-budget tagger for old czech. *In Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pages 10–18, Portland, OR, États-Unis, juin 2011.

Marie-Christine Hazaël-Massieux : *Ecrire en créole : Oralité et écriture aux Antilles*. L'Harmattan, 2000.

Nora Hollenstein et Noëmi Aepli : Compilation of a Swiss German Dialect Corpus and its Application to PoS Tagging. *In Proceedings of the 1st Workshop on Applying NLP Tools to Similar Languages, Varieties and Dialects, VarDial@COLING 2014*, pages 85–94, Dublin, Irlande, août 2014. URL http://www.aclweb.org/anthology/W14-5310.

Vinesh Y Hookoomsing : *A harmonized writing systemforthe Mauritian Creole Language Grafi-larmoni*. Ministry of Education and Scientific Research, 2004.

Dirk Hovy, Taylor Berg-Kirkpatrick, Ashish Vaswani et Eduard Hovy : Learning Whom to Trust with MACE. *In Actes de 2013 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, pages 1120–1130, Atlanta, GA, États-Unis, juin 2013. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/N13-1132.

Dirk Hovy, Barbara Plank et Anders Søgaard : Experiments with crowdsourced re-annotation of a POS tagging data set. *In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2 : Short Papers)*, pages 377–382, Baltimore, MD, États-Unis, juin 2014. URL http://www.aclweb.org/anthology/P14-2062.

Jeff Howe : The rise of crowdsourcing. *Wired magazine*, 14(6):1–4, 2006.

Dominique Huck, Arlette Bothorel-Witz et Anemone Geiger-Jaillet : L'Alsace et ses Langues. Eléments de description d'une situation sociolinguistique en zone frontalière. Language Bridges a Sub-theme Working Group of the Interreg IIIC project, 2007.

Inspection générale de l'éducation du sport et de la recherche IGÉSR : Maîtrise de la langue française en contexte plurilingue, Guadeloupe, septembre 2018. URL https://cache.media.eduscol.education.fr/file/Outre-Mer/13/4/Fiche_Guadeloupe_1180134.pdf.

Anupam Jamatia et Amitava Das : Part-of-Speech Tagging System for Indian Social Media Text on Twitter. *In Proceedings of the 1st Workshop on Language Technologies For Indian Social Media (SOCIAL-INDIA)*, pages 21–28, Goa, Inde, novembre 2014.

Béatrice Jeannot-Fourcaud : Créole, contact de langues et variabilité graphique dans les sms en guadeloupe. *Études Créoles*, 35(1):2–27, 2017.

David Jurgens et Roberto Navigli : It's all fun and games until someone annotates : Video games with a purpose for linguistic annotation. *Transactions of the Association for Computational Linguistics*, 2:449–464, 2014. URL https://www.aclweb.org/anthology/Q14-1035.

Elke Karan : Standardization : What's the hurry. *In Developing Orthographies for Unwritten Languages*, pages 107–138. SIL International Dallas, 2014.

Doruk Kicikoglu, Richard Bartle, Jon Chamberlain et Massimo Poesio : Wormingo : a'true gamification'approach to anaphoric annotation. *In Proceedings of the 14th International Conference on the Foundations of Digital Games*, pages 1–7, San Luis Obispo, CA, États-Unis, août 2019.

Adam Kilgarriff et Gregory Grefenstette : Web as corpus. *In Proceedings of the Corpus Linguistics 2001 conference*, volume 2001, pages 342–344, Lancaster, Royaume-Uni, mars 2001.

Adam Kilgarriff et Gregory Grefenstette : Introduction to the special issue on the web as corpus. *Computational linguistics*, 29(3):333–347, 2003.

Philipp Koehn : Europarl : A Parallel Corpus for Statistical Machine Translation. *In Proceedings of the 10th Machine Translation Summit (MT Summit X)*, pages 79–86, Phuket, Thaïlande, 2005. AAMT, AAMT. URL http://mt-archive.info/MTS-2005-Koehn.pdf.

BalaKrishna Kolluru, Lezan Hawizy, Peter Murray-Rust, Junichi Tsujii et Sophia Ananiadou : Using workflows to explore and optimise named entity recognition for chemistry. *PloS one*, 6(5):e20181, 2011.

Kimmo Koskenniemi : Finite-state relations between two historically closely related languages. *In* Northern European Association for Language Technology, éditeur : *Proceedings of the workshop on computational historical linguistics at NODALIDA 2013*, volume 18, pages 43–53, Oslo, Norvège , May 2013.

John Krumm, Nigel Davies et Chandra Narayanaswami : User-Generated Content. *IEEE Pervasive Computing*, 7(4):10–11, octobre 2008. ISSN 1536-1268. URL http://dx.doi.org/10.1109/MPRV.2008.85.

Mathieu Lafourcade et Alain Joubert : JeuxDeMots : un prototype ludique pour l'émergence de relations entre termes. *In Actes de Journées internationales d'Analyse statistique des Données Textuelles (JADT)*, Lyon, France, mars 2008. URL http://www.cavi.univ-paris3.fr/lexicometrica/jadt/jadt2008/pdf/lafourcade-joubert.pdf.

Mathieu Lafourcade, Nathalie Lebrun et Alain Joubert : *Jeux et intelligence collective : résolution de problèmes et acquisition de données sur le web*. Collection science cognitive et management des connaissances. ISTE, 2015. ISBN 9781784050528. URL https://books.google.fr/books?id=1UFqrgEACAAJ.

Guillaume Lample, Alexis Conneau, Ludovic Denoyer et Marc'Aurelio Ranzato : Unsupervised Machine Translation Using Monolingual Corpora Only, 2017.

Jacques Leclerc : Vitalité et mort des langues. *Les langues du monde*, Québec, CEFAN, Université Laval, 2020. URL http://www.axl.cefan.ulaval.ca/Langues/2vital.htm.

Tak Yeon Lee, Casey Dugan, Werner Geyer, Tristan Ratchford, Jamie Rasmussen, N Sadat Shami et Stela Lupushor : Experiments on motivational feedback for crowdsourced workers. *In Proceedings of the 7th International AAAI Conference on Weblogs and Social Media (ICWSM 2013)*, Cambridge, MA, États-Unis, juillet 2013.

Geoffrey Leech : Introducing corpus annotation. *In Corpus annotation*, pages 11–28. Routledge, 1997.

Adrian Leemann, Marie-José Kolly, Jean-Philippe Goldman, Volker Dellwo, Ingrid Hove, Ibrahim Almajai, Sarah Grimm, Sylvain Robert et Daniel Wanitsch : Voice Äpp : a mobile app for crowdsourcing swiss german dialect data. *In Proceedings of the 16th Annual Conference of the International Speech Communication Association (INTERSPEECH 2015)*, Dresden, Allemagne, septembre 2015.

Adrian Leemann, Marie-José Kolly, Ross Purves, David Britain et Elvira Glaser : Crowdsourcing language change with smartphone applications. *PloS one*, 11(1):1–25, 2016.

William D Lewis et Phong Yang : Building MT for a severely under-resourced language : White Hmong. *In Proceedings of the 10th Biennial Conference of the Association for Machine Translation in the Americas (AMTA 2012)*, San Diego, CA, États-Unis, octobre 2012.

Shen Li, João V. Graça et Ben Taskar : Wiki-ly Supervised Part-of-speech Tagging. *In Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP 2012)*, pages 1389–1398, Jeju, Corée du Sud, juillet 2012. URL http://dl.acm.org/citation.cfm?id=2390948.2391106.

Brook Danielle Lillehaugen : *Why write in a language that (almost) no one can read ? Twitter and the development of written literature*, volume 10, pages 356–393. University of Hawaii Press, 2016.

Nikola Ljubešic, Katja Zupan, Darja Fišer et Tomaz Erjavec : Normalising Slovene data : historical texts vs. user-generated content. *In Proceedings of the 13th Conference on Natural Language Processing (KONVENS 2016)*, pages 146–155, Bochum, Allemagne, septembre 2016.

Ralph Ludwig, Danièle Montbrand, Hector Poullet et Sylviane Telchid : Abrégé de grammaire du créole guadeloupéen. *In Dictionnaire créole français (Guadeloupe), avec un abrégé de grammaire créole et un lexique français-créole*, pages 17–38. SERVEDIT, 1990.

Ralph Ludwig, Danièle Montbrand, Hector Poullet et Sylviane Telchid : Dictionnaire créole français (guadeloupe). nouvelle édition, 2002.

Marco Lui et Timothy Baldwin : Accurate language identification of twitter messages. *In Proceedings of the 5th workshop on language analysis for social media (LASM)*, pages 17–25, Göteborg, Suède, avril 2014.

Verena Lyding, Egon Stemle, Claudia Borghetti, Marco Brunello, Sara Castagnoli, Felice Dell'Orletta, Henrik Dittmann, Alessandro Lenci et Vito Pirrelli : The paisa'corpus of italian web texts. *In Proceedings of the 9th Web as Corpus Workshop (WaC-9), 14th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2014)*, pages 36–43, Göteborg, Suède, avril 2014. EACL (European chapter of the Association for Computational Linguistics).

Friederike Lüpke : Orthography development. *In Handbook of endangered languages*, pages 312–336. Cambridge : Cambridge University Press, Austin, Peter and Sallabank, Julia édition, 2011.

Maaya : *Net.lang : Towards the Multilingual Cyberspace*. C & F Editions, 2012. ISBN 9782915825244. URL https://books.google.fr/books?id=ZF4snQAACAAJ.

Chris Madge, Richard Bartle, Jon Chamberlain, Udo Kruschwitz et Massimo Poesio : Making text annotation fun with a clicker game. *In Proceedings of the the 14th International Conference on the Foundations of Digital Games*, FDG '19, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450372176. URL https://doi.org/10.1145/3337722.3341869.

Chris Madge, Juntao Yu, Jon Chamberlain, Udo Kruschwitz, Silviu Paun et Massimo Poesio : Crowdsourcing and aggregating nested markable annotations. *In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 797–807, Florence, Italie, janvier 2019.

Pierre Magistry, Anne-Laure Ligozat et Sophie Rosset : Étiquetage en parties du discours de langues peu dotées par spécialisation des plongements lexicaux. *In Actes de Conférence sur le Traitement Automatique des Langues Naturelles (TALN 2018)*, Rennes, France, mai 2018. URL https://hal.archives-ouvertes.fr/hal-01793092.

Michel Malherbe : *Les langages de l'humanité (une encyclopédie des 3000 langues parlées dans le monde)*. Collection Bouquins. Laffont, 1983.

Meï-Lan Mamode : À l'encontre des présupposés linguistiques : Morphologie flexionnelle et dérivationnelle du créole mauricien. *In Proceedings of the Bilingual Workshop on Theoretical Linguistics*, Waterloo, Canada, décembre 2013.

Mitchell P. Marcus, Mary Ann Marcinkiewicz et Beatrice Santorini : Building a Large Annotated Corpus of English : The Penn Treebank. *Computational Linguistics*, 19(2):313–330, juin 1993. ISSN 0891-2017.

Héctor Martínez Alonso, Djamé Seddah et Benoît Sagot : From noisy questions to Minecraft texts : Annotation challenges in extreme syntax scenario. *In Proceedings of the 2nd Workshop on Noisy User-generated Text (WNUT)*, pages 13–23, Osaka, Japon, décembre 2016. The COLING 2016 Organizing Committee. URL https://www.aclweb.org/anthology/W16-3905.

Yann Mathet, Antoine Widlöcher, Karen Fort, Claire François, Olivier Galibert, Cyril Grouin, Juliette Kahn, Sophie Rosset et Pierre Zweigenbaum : Manual Corpus Annotation : Giving Meaning to the Evaluation Metrics. *In International Conference on Computational Linguistics*, pages 809–818, Mumbaï, Inde, décembre 2012. URL https://hal.archives-ouvertes.fr/hal-00769639.

Tony McEnery et Andrew Hardie : *Corpus Linguistics : Method, Theory and Practice*. Cambridge Textbooks in Linguistics. Cambridge University Press, 2011. ISBN 9781139502443. URL https://books.google.fr/books?id=3j3Wn_ZT1qwC.

Maite Melero, Marta R. Costa-Jussà, Judith Domingo, Montse Marquina et Martí Quixal : Holaaa!! writin like u talk is kewl but kinda hard 4 NLP. *In Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012)*, Istanbul, Turquie, mai 2012. ISBN 978-2-9517408-7-7.

Sabrina J. Mielke : Language diversity in ACL 2004 - 2016, Dec 2016. URL https://sjmielke.com/acl-language-diversity.htm.

Margot Mieskes, Karën Fort, Aurélie Névéol, Cyril Grouin et Kevin B Cohen : NLP Community Perspectives on Replicability. *In Proceedings of the Recent Advances in Natural Language Processing (RANLP 2019)*, Varna, Bulgarie, septembre 2019. URL https://hal.archives-ouvertes.fr/hal-02282794.

Allison Miller : *Kreol in Mauritian Schools : Mother Tongue Language Education and Public*. Thèse de doctorat, Yale University, New Haven, CT, États-Unis, 2015.

Alice Millour : Getting to Know the Speakers : a Survey of a Non-Standardized Language Digital Use. *In Proceedings of the 9th Language & Technology Conference : Human Language Technologies as a Challenge for Computer Science and Linguistics (LTC 2019)*, Poznań, Pologne, mai 2019. URL https://hal.archives-ouvertes.fr/hal-02137280.

Alice Millour : Production participative et ressources linguistiques, quels enjeux pour la diversité linguistique en TAL ? *In Séminaire Cognition & Langage de Maxime Amblard et Manuel Rebuschi*, IDMC (Institut des Sciences du Digital - Management & Cognition), Nancy, février 2019.

Alice Millour et Karën Fort : Unsupervised Data Augmentation for Less-Resourced Languages with no Standardized Spelling. *In Proceedings of the Recent Advances in Natural Language Processing conference (RANLP 2019)*, pages 776 – 784, Varna, Bulgarie, septembre 2019. URL https://hal.archives-ouvertes.fr/hal-02280002.

Alice Millour et Karën Fort : Text Corpora and the Challenge of Newly Written Languages. *In Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-resourced languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL), LREC 2020*, pages 111–120, Marseille, France, mai 2020. European Language Resources association.

Alice Millour, Karën Fort, Delphine Bernhard et Lucie Steible : Vers une solution légère de production de données pour le TAL : création d'un tagger de l'alsacien par crowdsourcing bénévole. *In Actes de Traitement Automatique des Langues Naturelles (TALN 2017)*, Orléans, France, juin 2017.

Alice Millour, Karën Fort et Pierre Magistry : Répliquer et étendre pour l'alsacien "étiquetage en parties du discours de langues peu dotées par spécialisation des plongements lexicaux". *In Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 31e édition), Traitement Automatique des Langues Naturelles (TALN, 27e édition), Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition). 2e atelier Éthique et TRaitemeNt Automatique des Langues (ETeRNAL)*, pages 29–37, Nancy, France, juin 2020. ATALA. en ligne .

Alice Millour, Marianne Grace Araneta, Ivana Lazić Konjik, Annalisa Raffone, Yann-Alan Pilatte et Karën Fort : Katana and Grand Guru : a Game of the Lost Words (DEMO). *In Proceedings of the 9th Language & Technology Conference : Human Language Technologies as a Challenge for Computer Science and Linguistics (LTC 2019)*, Poznań, Pologne, mai 2019. URL https://hal.archives-ouvertes.fr/hal-02106757.

Alberto M Mioni : Italiano tendenziale : osservazioni su alcuni aspetti della standardizzazione. *Benincà, Paola et al. (ed.) Scritti linguistici in onore di Giovan Battista Pellegrini*, 1:495–517, 1983.

Christian Monson, Ariadna Font Llitjós, Roberto Aranovich, Lori M. Levin, R. Brown, Eric Peterson, Jaime G. Carbonell et Alon Lavie : Building NLP Systems for Two Resource-Scarce Indigenous Languages : Mapudungun and Quechua. *In Proceedings of the 5th SALTMIL Workshop on Minority Languages, 5th International Conference on Language Resources and Evaluation (LREC 2006)*, pages 15–24, Gênes, Italie, mai 2006.

Marie-Louise Moreau : *Sociolinguistique : les concepts de base*, volume 218. Editions Mardaga, 1997.

Benedikt Morschheuser, Juho Hamari et Alexander Maedche : Cooperation or competition – When do people contribute more ? A field experiment on gamification of crowdsourcing. *International Journal of Human-Computer Studies*, 127:7 – 24, 2019. ISSN 1071-5819. URL http://www.sciencedirect.com/science/article/pii/S1071581918305822.

Christopher (ed.) Moseley : Atlas des langues en danger dans le monde, 3ème edn. Editions UNESCO, 2010. URL http://www.unesco.org/culture/languages-atlas/fr/atlasmap.html.

Robert Munro : Crowdsourcing and the Crisis-Affected Community : lessons learned and looking forward from Mission 4636. *Journal of Information Retrieval*, 16(2):210–266, 2013. URL http://www.robertmunro.com/research/Mission_4636_Haiti_2010_SMS.pdf.

Robert Munro : Languages at ACL this year, juillet 2015. URL http://www.junglelightspeed.com/languages-at-acl-this-year/. Consulté le 4 septembre 2020.

Wilhelmina Nekoto, Vukosi Marivate, Tshinondiwa Matsila, Timi Fasubaa, Taiwo Fagbohungbe, Solomon Oluwole Akinola, Shamsuddeen Muhammad, Salomon Kabongo Kabenamualu, Salomey Osei, Freshia Sackey, Rubungo Andre Niyongabo, Ricky Macharm, Perez Ogayo, Orevaoghene Ahia, Musie Meressa Berhe, Mofetoluwa Adeyemi, Masabata Mokgesi-Selinga, Lawrence Okegbemi, Laura Martinus, Kolawole Tajudeen, Kevin Degila, Kelechi Ogueji, Kathleen Siminyu, Julia Kreutzer, Jason Webster, Jamiil Toure Ali, Jade Abbott, Iroro Orife, Ignatius Ezeani, Idris Abdulkadir Dangana, Herman Kamper, Hady Elsahar, Goodness Duru, Ghollah Kioko, Murhabazi Espoir, Elan van Biljon, Daniel Whitenack, Christopher Onyefuluchi, Chris Chinenye Emezue, Bonaventure F. P. Dossou, Blessing Sibanda, Blessing Bassey, Ayodele Olabiyi, Arshath Ramkilowan, Alp Öktem, Adewale Akinfaderin et Abdallah Bashir : Participatory research for low-resourced machine translation : A case study in African languages. *In Findings of the Association for Computational Linguistics : EMNLP 2020*, pages 2144–2160, en ligne , novembre 2020. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/2020.findings-emnlp.195.

Vlad Niculae et Cristian Danescu-Niculescu-Mizil : Conversational markers of constructive discussions. *In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology (NAACL 2016)*, San Diego, CA, États-Unis, juin 2016.

Natacha Niemants : Des enregistrements aux corpus : transcription et extraction de données d'interprétation en milieu médical. *Meta*, 63(3):665–694, 2018.

Umera-Okeke Nneka et Mercy Okitikpi : Age variation in the use of Nigerian Pidgin (NP) : A case of Sapele, Delta State, Nigeria. *International Journal of English and Literature*, 8 (2):16–25, 2017.

Peter Norvig : On Chomsky and the two cultures of statistical learning. *In Berechenbarkeit der Welt ?*, pages 61–83. Springer, 2017.

Observatoire des pratiques linguistiques : *Les créoles à base française*, volume 5. DGLFLF, 2005.

Javier Olivas Alguacil : Lexicography, a socio-political battleground. Two cases : Reunion and Mauritius. *Carnets de Recherches de l'océan Indien*, 3:1–10, juin 2019. URL https://hal.archives-ouvertes.fr/hal-02167886.

Pedro Javier Ortiz Suárez, Benoît Sagot et Laurent Romary : Asynchronous Pipeline for Processing Huge Corpora on Medium to Low Resource Infrastructures. *In Proceedings of the 7th Workshop on the Challenges in the Management of Large Corpora (CMLC-7)*, Cardiff, United Kingdom, juillet 2019. URL https://hal.inria.fr/hal-02148693.

Michaël Oustinoff : English Won't be the Internet's Lingua Franca. *In Towards the Multilingual Cyberspace*, pages 171–178. Vannini, Laurent and Le Crosnier, Hervé, c&f édition, 2012.

Sean Packham et Hussein Suleman : Crowdsourcing a Text Corpus is not a Game. *In Proceedings of the 2015 International Conference on Asian Digital Libraries (ICADL 2015)*, pages 225–234, Seoul, Corée du Sud, juillet 2015. Springer.

Xiaoman Pan, Boliang Zhang, Jonathan May, Joel Nothman, Kevin Knight et Heng Ji : Cross-lingual name tagging and linking for 282 languages. *In Proceedings of the the 55th Annual Meeting of the Association for Computational Linguistics*, volume 1, pages 1946–1958, Vancouver, Canada, juillet 2017. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/P17-1178.

John C Paolillo et Anupam Das : Evaluating language statistics : The ethnologue and beyond. UNESCO Institute for Statistics, 2006.

Douglas B Paul et Janet M Baker : The design for the Wall Street Journal-based CSR corpus. *In Proceedings of the 5th DARPA Speech and Natural Language workshop*, pages 357–362, Harriman, NY, États-Unis, février 1992. Association for Computational Linguistics.

Silviu Paun, Bob Carpenter, Jon Chamberlain, Dirk Hovy, Udo Kruschwitz et Massimo Poesio : Comparing Bayesian Models of Annotation. *Transactions of the Association for Computational Linguistics*, 6:571–585, 2018.

Slav Petrov, Dipanjan Das et Ryan McDonald : A Universal Part-of-Speech Tagset. *In Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012)*, Istanbul, Turquie, mai 2012.

Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych et Sebastian Ruder : MAD-X : An Adapter-Based Framework for Multi-Task Cross-Lingual Transfer. *In Proceedings of the the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7654–7673, en ligne , novembre 2020. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/2020.emnlp-main.617.

Daniel Pimienta et Daniel Prado : *Étude sur la place des langues de France sur l'Internet.* DGLFLF, 2014.

Sriharini Pingali, David Mortensen, Patrick Littell et Lori Levin : Phonetically-Aware Approximate Search for Low-Resource Languages. Rapport technique, Carnegie Mellon University, Pittsburgh, PA, États-Unis, 2017.

Yuval Pinter, Robert Guthrie et Jacob Eisenstein : Mimicking word embeddings using subword RNNs. *In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP 2017)*, pages 102–112, Copenhague, Danemark, septembre 2017. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/D17-1010.

Barbara Plank : What to do about non-standard (or non-canonical) language in NLP. *In Proceedings of the 13th Conference on Natural Language Processing (KONVENS 2016)*, pages 13–20, Bochum, Allemagne, août 2016. URL http://arxiv.org/abs/1608.07836.

Massimo Poesio, Jon Chamberlain, Udo Kruschwitz, Livio Robaldo et Luca Ducceschi : Phrase Detectives : Utilizing Collective Intelligence for Internet-scale Language Resource Creation. *ACM Trans. Interact. Intell. Syst.*, 3(1):3 :1–3 :44, avril 2013. ISSN 2160-6455. URL http://doi.acm.org/10.1145/2448116.2448119.

Jelena Prokić, Martijn Wieling et John Nerbonne : Multiple sequence alignments in linguistics. *In Proceedings of the EACL 2009 Workshop on Language Technology and Resources for Cultural Heritage, Social Sciences, Humanities, and Education*, pages 18–25, Stroudsburg, PA, États-Unis, 2009. Association for Computational Linguistics. URL http://dl.acm.org/citation.cfm?id=1642049.1642052.

Delyth Prys : Developing language technologies for less-resourced languages. Invited talk at the 9th Language & Technology Conference : Human Language Technologies as a Challenge for Computer Science and Linguistics (LTC 2019), mai 2019.

Delyth Prys, Gruffudd Prys et Dewi Bryn Jones : Cysill Ar-lein : A corpus of written contemporary Welsh compiled from an on-line spelling and grammar checker. *In Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*, pages 3261–3264, Portorož, Slovénie , mai 2016. European Language Resources Association (ELRA).

Alexander J Quinn et Benjamin B Bederson : Human Computation :A Survey and Taxonomy of a Growing Field. *In Proceedings of the 2011 SIGCHI conference on human factors in computing systems (CHI 2011)*, pages 1403–1412, Vancouver, Canada, mai 2011. ACM.

Adwait Ratnaparkhi : A maximum entropy model for part-of-speech tagging. *In Proceedings of the 1996 Conference on Empirical Methods in Natural Language Processing (EMNLP'1996)*, Philadeplhie, PA, États-Unis, 1996.

Republic Of Mauritius Ministry Of Finance And Economic Development Statistics Mauritius : Housing and Population census. Republic of Mauritius, 2011. Vol. II : DEMOGRAPHIC AND FERTILITY CHARACTERISTICS.

Vassili Rivron : L'usage de Facebook chez les Éton du Cameroun. *In Net.lang Réussir le cyberespace multilingue*, pages 171–178. Vannini, Laurent and Le Crosnier, Hervé, c&f édition, 2012.

Pablo Ruiz Fabo, Delphine Bernhard, Pascale Erhart, Dominique Huck et Carole Werner : MeThAL : Vers une macroanalyse du théâtre en alsacien, mai 2020. URL https://doi.org/10.5281/zenodo.3788020.

Marta Sabou, Kalina Bontcheva, Leon Derczynski et Arno Scharl : Corpus Annotation through Crowdsourcing : Towards Best Practice Guidelines. *In Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 2014)*, pages 859–866, Reykjavik, Islande, mai 2014. European Language Resources Association (ELRA). URL http://www.lrec-conf.org/proceedings/lrec2014/pdf/497_Paper.pdf.

Benoît Sagot : The Lefff, a freely available and large-coverage morphological and syntactic lexicon for French. *In 7th international conference on Language Resources and Evaluation (LREC 2010)*, La Valette, Malte, mai 2010. URL https://hal.inria.fr/inria-00521242.

Benoît Sagot : Étiquetage multilingue en parties du discours avec melt (multilingual part-of-speech tagging with melt)[in french]. *In Actes de la conférence conjointe JEP-TALN-RECITAL 2016. volume 2 : TALN (Posters)*, pages 435–442, Paris, France, juillet 2016.

Benoît Sagot : *Computerising the lexicon*. Habilitation à diriger des recherches, Sorbonne Université, juin 2018. URL https://hal.inria.fr/tel-01895229.

Benoît Sagot : Développement d'un lexique morphologique et syntaxique de l'ancien français. *In Actes de26ème Conférence sur le Traitement Automatique des Langues Naturelles (TALN)*, Toulouse, France, juillet 2019. URL https://hal.inria.fr/hal-02148701.

Benoît Sagot, Karën Fort, Gilles Adda, Joseph Mariani et Bernard Lang : Un turc mécanique pour les ressources linguistiques : critique de la myriadisation du travail parcellisé. *In Actes de Traitement Automatique des Langues Naturelles (TALN)*, Montpellier, France, juin 2011. URL http://hal.inria.fr/inria-00617067/.

Tanja Samardzic, Yves Scherrer et Elvira Glaser : Normalising orthographic and dialectal variants for the automatic processing of Swiss German. *In Proceedings of the 7th Language & Technology Conference : Human Language Technologies as a Challenge for Computer Science and Linguistics (LTC 2015)*, Poznań, Pologne, novembre 2015.

Kevin P Scannell : The Crúbadán project : Corpus building for under-resourced languages. *In Proceedings of the 3rd Web as Corpus Workshop : Building and Exploring Web Corpora*, volume 4, pages 5–15, Louvain-la-Neuve, Belgique, septembre 2007.

Roland Schäfer et Felix Bildhauer : Web corpus construction. *Synthesis Lectures on Human Language Technologies*, 6(4):1–145, 2013.

Emmanuel Schang : Extended Projections in a Guadeloupean TAG Grammar. *In Proceedings of the 2013 ESSLLI (HMGE workshop)*, pages 55–67, Düsseldorf, Germany, June 2013.

Emmanuel Schang, Jean-Yves Antoine et Anaïs Lefeuvre-Halftermeyer : Les chaînes coréférentielles en créole de la Guadeloupe. *In Actes de Traitement Automatique des Langues Naturelles (TALN 2017) (DILITAL workshop)*, pages 54–61, Orléans, France, juin 2017. URL https://hal.archives-ouvertes.fr/hal-01627260.

Yves Scherrer et Benoît Sagot : Lexicon induction and part-of-speech tagging of non-resourced languages without any bilingual resources. *In Proceedings of the Workshop on Adaptation of language resources and tools for closely related languages and language variants, 9th Recent Advances in Natural Language Processing conference (RANLP 2013)*, Hissar, Bulgarie, septembre 2013. URL https://hal.inria.fr/hal-00862693.

Anne Schiller, Simone Teufel et Christine Thielen : Guidelines für das Tagging deutscher Textcorpora mit STTS. Rapport technique, Universitäten Stuttgart und Tübingen, 1995.

Helmut Schmid : Probabilistic part-of-speech tagging using decision trees. *In New Methods in Language Processing, Studies in Computational Linguistics*, pages 154–164. UCL Press, 1997. URL http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/.

Holger Schwenk, Guillaume Wenzek, Sergey Edunov, Edouard Grave et Armand Joulin : Ccmatrix : Mining billions of high-quality parallel sentences on the web, 2020.

William A Scott : Reliability of content analysis : The case of nominal scale coding. *The Public Opinion Quarterly*, 19(3):321–325, 1955.

Djamé Seddah, Farah Essaidi, Amal Fethi, Matthieu Futeral, Benjamin Muller, Pedro Javier Ortiz Suárez, Benoît Sagot et Abhishek Srivastava : Building a user-generated content North-African Arabizi treebank : Tackling hell. *In Proceedings of the the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1139–1150, en ligne , juillet 2020. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/2020.acl-main.107.

Djamé Seddah, Benoît Sagot, Marie Candito, Virginie Mouilleron et Vanessa Combet : Building a treebank of noisy user-generated content : The French Social Media Bank. *In Proceedings of the TLT 11 - The 11th International Workshop on Treebanks and Linguistic Theories*, Lisbonne, Portugal, novembre 2012. URL https://hal.inria.fr/hal-00780898. Cet article constitue une version réduite de l'article "The French Social Media Bank : a Treebank of Noisy User Generated Content" (mêmes auteurs).

Serge Sharoff : Creating General-Purpose Corpora Using Automated Search Engine Queries. *In WaCky ! Working papers on the Web as Corpus*. Gedit, 2006.

Shashi Shekhar, Dilip Kumar Sharma et MM Sufyan Beg : Hindi Roman Linguistic Framework for Retrieving Transliteration Variants using Bootstrapping. *Procedia Computer Science*, 125:59–67, 2018.

J Sinclair et J Ball : Preliminary Recommendations on Text Typology. EAGLES Document EAG-TCWG-TTYP/P, 1996. http ://www.ilc.cnr.it/EAGLES96/texttyp/texttyp.html.

John Sinclair : Corpus and Text-Basic Principles. *In Developing Linguistic Corpora : a Guide to Good Practice*, chapitre 1, pages 1–16. Oxford : Oxbow Books, m. wynne édition, 2005.

Elena Soare, Anne Zribi-Hertz, Sarra El Ayari, Maxime Deglas M., Tomer Rosenberg et Coralie Vincent : Histoire de l'âne en créole guadeloupéen. Editeur : Structures formelles du langage. Collection : "Langues et Grammaires en Ile-de-France (LGIDF)", 2016. Récupéré sur la plateforme COCOON, <http ://purl.org/poi/crdo.vjf.cnrs.fr/cocoon-20b2e6fa-045f-3394-8c4e-9f2bd06e3863>. (Consulté le 16 juin 2020).

Claudia Soria, Valeria Quochi et Irene Russo : The DLDP Survey on Digital Use and Usability of EU Regional and Minority Languages. *In Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japon, mai 2018. European Language Resources Association (ELRA). URL https://www.aclweb.org/anthology/L18-1656.

Lucie Steible et Delphine Bernhard : Pronunciation Dictionaries for the Alsatian Dialects to Analyze Spelling and Phonetic Variation. *In Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japon, mai 2018. URL https://hal.archives-ouvertes.fr/hal-01704814.

Dominique Stich, Xavier Gouvert et Alain Favre : *Dictionnaire des mots de base du francoprovençal : orthographe ORB supradialectale standardisée*. Le Carré, 2003.

Oscar Täckström, Dipanjan Das, Slav Petrov, Ryan McDonald et Joakim Nivre : Token and type constraints for cross-lingual part-of-speech tagging. *Transactions of the Association for Computational Linguistics*, 1:1–12, 2013.

The Digital Language Diversity Project : Sardinian — a digital language ? *In Reports on Digital Language Diversity in Europe*. Editors : Claudia Soria, Irene Russo, Valeria Quoch, 2017.

Pieter Theron et Ian Cloete : Automatic Acquisition of Two-level Morphological Rules. *In Proceedings of the 5th Conference on Applied Natural Language Processing*, ANLC 1997, pages 103–110, Stroudsburg, PA, États-Unis, 1997. Association for Computational Linguistics. URL https://doi.org/10.3115/974557.974573.

André Thibault : Le sort des consonnes finales en français, en galloroman et en créole : le cas de'moins'. *Revue de linguistique romane*, 81(321/322):5–41, 2017.

Anne-Marie Thomson : Language contact and codification : Mauritian creole. Unity in Diversity : the prospect of a standardised Creole, as a symbol of unity and identity in Mauritius : A case study., avril 2006. http ://languagecontact.humanities.manchester.ac.uk/McrLC/casestudies/AMT.html.

Jörg Tiedemann : Parallel Data, Tools and Interfaces in OPUS. *In* Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk et Stelios Piperidis, éditeurs : *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012)*, Istanbul, Turquie, mai 2012. European Language Resources Association (ELRA). ISBN 978-2-9517408-7-7.

Taha Tobaili, Miriam Fernandez, Harith Alani, Sanaa Sharafeddine, Hazem Hajj et Goran Glavas : SenZi : A Sentiment Analysis Lexicon for the Latinised Arabic (Arabizi). In : International Conference Recent Advances. *In Proceedings of the Recent Advances in Natural Language Processing conference (RANLP 2019)*, Varna, Bulgarie, septembre 2019.

Julien Tourille, Olivier Ferret, Aurélie Névéol et Xavier Tannier : Neural architecture for temporal relation extraction : A bi-LSTM approach for detecting narrative containers. *In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2 : Short Papers)*, pages 224–230, Vancouver, Canada, juillet 2017. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/P17-2035.

Nicolas Tournadre : The Tibetic languages and their classification. *In Trans-Himalayan linguistics : Historical and descriptive linguistics of the Himalayan area.* Owen-Smith, Thomas / Hill, Nathan, 2014.

Kristina Toutanova, Dan Klein, Christopher D. Manning et Yoram Singer : Feature-rich Part-of-speech Tagging with a Cyclic Dependency Network. *In Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology (NAACL 2003)*, pages 173–180, Stroudsburg, PA, États-Unis, mai 2003. URL http://dx.doi.org/10.3115/1073445.1073478.

Huihsin Tseng, Daniel Jurafsky et Christopher Manning : Morphological features help POS tagging of unknown words across language varieties. *In Proceedings of the 4th SIGHAN Workshop on Chinese Language Processing*, pages 32–39, Jeju, Corée du Sud, octobre 2005. URL https://web.stanford.edu/{~}jurafsky/sighan{_}pos.pdf.

Kathleen Tuite : GWAPs : Games with a Problem. *In Proceedings of the 9th International Conference on the Foundations of Digital Games*, Liberty of the Seas, Caraïbes, avril 2014.

Assaf Urieli : *Robust French syntax analysis : reconciling statistical methods and linguistic knowledge in the Talismane toolkit.* Thèse de doctorat, Toulouse le Mirail-Toulouse II University, 2013.

David C Uthus et David W Aha : The ubuntu chat corpus for multiparticipant chat analysis. *In Proceedings of the 2013 AAAI Spring Symposium Series*, Stanford, CA, États-Unis, mars 2013.

Andrius Vabalas, Emma Gowen, Ellen Poliakoff et Alexander J Casson : Machine learning algorithm validation with a limited sample size. *PloS one*, 14(11):e0224365, 2019.

Daan van Esch, Elnaz Sarbar, Tamar Lucassen, Jeremy O'Brien, Theresa Breiner, Manasa Prasad, Evan Elizabeth Crew, Chieu Nguyen et Francoise Beaufays : Writing Across the World's Languages : Deep Internationalization for Gboard, the Google Keyboard. Rapport technique, Google, 2019. URL https://arxiv.org/abs/1912.01218.

Marianne Vergez-Couret, Assaf Urieli et France Foix : Pos-tagging different varieties of Occitan with single-dialect resources. *In Proceedings of the 1st Workshop on Applying NLP Tools to Similar Languages, Varieties and Dialects, VarDial@COLING 2014*, pages 21–29, Dublin, Irlande, août 2014.

Luis von Ahn : *Human computation (PhD thesis)*. Thèse de doctorat, Carnegie Mellon University, 2005.

Holger Voormann et Ulrike Gut : Agile corpus creation. *Corpus Linguistics and Linguistic Theory*, 4(2):235–251, 2008.

Aobo Wang, Cong Duy Vu Hoang et Min-Yen Kan : Perspectives on crowdsourcing annotations for natural language processing. *Language resources and evaluation*, 47(1):9–31, 2013.

Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco GuzmÃ¡n, Armand Joulin et Edouard Grave : Ccnet : Extracting high quality monolingual datasets from web crawl data. *In Proceedings of the the 12th Language Resources and Evaluation Conference*, pages 4003–4012, Marseille, France, mai 2020. European Language Resources Association. URL https://www.aclweb.org/anthology/2020.lrec-1.494.

Wikimedia : Language proposal policy, 2020. URL https://meta.wikimedia.org/wiki/Language_proposal_policy.

Mark D Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E Bourne *et al.* : The FAIR Guiding Principles for scientific data management and stewardship. *Scientific data*, 3(1):1–10, 2016.

Jennifer Williams et Charlie Dagli : Twitter language identification of similar languages and dialects without ground truth. *In Proceedings of the 4th Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial), co-located with EACL 2017*, pages 73–83, Valence, Espagne, avril 2017.

M Wynne, éditeur. *Developing Linguistic Corpora : a Guide to Good Practice*. Oxford : Oxbow Books., 2005. URL http://ahds.ac.uk/linguistic-corpora/.

Wajdi Zaghouani et Kais Dukes : Can Crowdsourcing be used for Effective Annotation of Arabic ? *In Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 2014)*, Reykjavik, Islande, mai 2014.

Othman Zennaki, Nasredine Semmar et Laurent Besacier : Inducing Multilingual Text Analysis Tools Using Bidirectional Recurrent Neural Networks. *In Proceedings of the 26th International Conference on Computational Linguistics (COLING 2016)*, Osaka, Japon, décembre 2016. URL https://hal.archives-ouvertes.fr/hal-01374205.

Carlo Zoli et Silvia Randaccio : The context of use of e-dictionaries for the minority languages of Italy (case study). *In Proceedings of the European Network of e-Lexicography (Enel) COST Action (WG3 meeting)*, Barcelone, Espagne, mars 2016.

Annexes

Les annexes qui suivent sont composées :

- des matrice de confusion
- des ressources que nous avons collectées *via* nos plateformes :
  - les corpus bruts pour l'alsacien et le créole mauricien (Annexe B),
  - les listes de variantes graphiques myriadisées pour l'alsacien et le créole mauricien (Annexe C).
- des règles de transposition déduites des variantes graphiques alignées pour l'alsacien (Annexe D),
- du guide d'installation et d'adaptation des plateformes développées à une nouvelle langue (Annexe E),
- des questionnaires des enquêtes et leurs résultats (tels que produits par `framaforms`)[163] (Annexes F et G).

Elles ne contiennent pas les éléments suivants, mis à disposition en ligne :

- les codes source des plateformes P__Ann et P__Prod__Var, publié sur GitHub, à l'adresse https://github.com/alicemillour/Bisame :
  - `Bisame` : branche `bisame` ;
  - `Krik !` : branche `krik` ;
  - `Recettes de Grammaire` : branche `recipes` ;
  - `Ayo !` : branche `recipes_ayo`.
- réponses brutes aux enquêtes réalisées :
  - `L'alsacien, Internet, et vous`, disponible au téléchargement : alicemillour.github.io/assets/survey_gsw_2018.tsv ;
  - `Le créole mauricien et sa présence en ligne`, disponible au téléchargement : alicemillour.github.io/assets/survey_mfe_2018.tsv ;
- le corpus annoté de référence du créole guadeloupéen, disponible à la page : alicemillour.github.io/assets/gcf_annotated_corpus.csv et distribué sur Ortolang : https://repository.ortolang.fr/api/content/krik_gcf/head/ ;
- le corpus annoté de référence et le corpus myriadisé du créole mauricien, disponible à la page : alicemillour.github.io/assets/mfe_annotated_corpus.csv, alicemillour.github.io/assets/mfe_annotated_corpus_myriadise.csv et distribués sur Ortolang : https://repository.ortolang.fr/api/content/ayo_mfe/head/ ;
- le corpus myriadisé de de l'alsacien, disponible à la page : alicemillour.github.io/assets/gsw_annotated_corpus_myriadise.csv et distribués sur Ortolang : https://repository.ortolang.fr/api/content/bisame_gsw/head/ ;

163. Voir : framaforms.org/.

Annexe A

Matrices de confusion des annotations de références produites pour l'alsacien

	ADJ	ADP	ADV	AUX	CONJ	DET	INTJ	NOUN	NUM	PART	PRON	PROPN	PUNCT	SCONJ	SYM	VERB	X
ADJ	7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ADP	0	1 8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ADV	0	0	1 9	0	0	0	0	0	0	0	0	0	0	0	0	0	0
AUX	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	2	0
CONJ	0	0	1	0	7	0	0	0	0	0	0	0	0	0	0	0	0
DET	0	1	0	0	0	1 0	0	1	0	0	0	0	0	0	0	0	0
INTJ	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0
NOUN	0	0	0	0	0	0	0	2 2	0	0	0	0	0	0	0	0	0
NUM	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
PART	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0
PRON	0	0	0	0	1	0	0	1	0	0	3 0	0	0	0	0	0	0
PROPN	0	0	0	0	0	0	0	1	0	0	0	8	0	0	0	0	0
PUNCT	0	0	0	0	0	0	0	0	0	0	0	0	5 0	0	0	0	0
SCONJ	0	0	2	0	2	0	0	0	0	0	0	0	0	3	0	0	0
SYM	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
VERB	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	2 5	2
X	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	4

TABLEAU A.1 – Matrices de confusion des deux annotations manuelles fournies par les chercheuses du LiLPa (corpus `Hoflieferant_P53`).

	ADJ	ADP	ADV	AUX	CONJ	DET	INTJ	NOUN	NUM	PART	PRON	PROPN	PUNCT	SCONJ	SYM	VERB	X
ADJ	2 1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ADP	2	3 1	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0
ADV	6	4	7	0	1	0	0	0	0	0	0	0	0	0	0	0	0
AUX	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
CONJ	0	3	3	0	2 2	0	0	0	0	0	0	0	0	0	0	0	0
DET	1	0	0	0	0	3 4	0	3	0	0	0	0	0	0	0	0	0
INTJ	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
NOUN	0	0	0	0	0	0	0	7 7	0	0	0	0	0	0	0	0	0
NUM	0	0	0	0	0	0	0	1	9	0	0	0	0	0	0	0	0
PART	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0
PRON	1	0	0	0	0	4	0	0	0	0	4	0	0	0	0	0	0
PROPN	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
PUNCT	0	0	0	0	0	0	0	0	0	0	0	0	5 1	0	0	0	0
SCONJ	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0
SYM	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
VERB	0	9	0	0	0	0	0	0	0	0	0	0	0	0	0	5 8	1
X	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0

TABLEAU A.2 – Matrices de confusion des deux annotations manuelles fournies par les chercheuses du LiLPa (corpus `recettes`).

	ADJ	ADP	ADV	AUX	CONJ	DET	INTJ	NOUN	NUM	PART	PRON	PROPN	PUNCT	SCONJ	SYM	VERB	X
ADJ	9	0	0	0	0	0	0	0	1	0	0	2	0	0	0	0	0
ADP	0	3 9	0	0	0	0	0	3	1	0	0	0	0	0	0	0	0
ADV	2	0	9	0	0	0	0	0	0	0	0	0	0	0	0	0	1
AUX	0	0	0	2 3	0	0	0	0	0	0	0	0	0	0	0	3	0
CONJ	0	0	1	0	9	0	0	0	0	0	0	0	0	0	0	0	0
DET	0	0	0	0	0	4 6	0	0	0	0	3	0	0	0	0	0	0
INTJ	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
NOUN	0	0	0	0	0	0	0	6 6	0	0	0	1	0	0	0	0	1
NUM	0	0	0	0	0	0	0	0	1 6	1	0	0	0	0	0	0	1
PART	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
PRON	0	0	1	0	0	0	0	0	0	0	5	0	0	4	0	1	1
PROPN	0	0	0	0	0	0	0	0	0	0	0	3 7	0	0	0	0	0
PUNCT	0	0	0	0	0	0	0	0	0	0	0	0	4 8	0	0	0	1
SCONJ	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
SYM	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0
VERB	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	3 4	0
X	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1 9

TABLEAU A.3 – Matrices de confusion des deux annotations manuelles fournies par les chercheuses du LiLPa (corpus `wikipedia1`).

	ADJ	ADP	ADV	AUX	CONJ	DET	INTJ	NOUN	NUM	PART	PRON	PROPN	PUNCT	SCONJ	SYM	VERB	X
ADJ	1 7	1	1	0	0	0	0	0	0	0	0	0	0	0	0	2	0
ADP	0	5 2	0	0	0	0	0	5	0	0	0	1	1	0	0	0	0
ADV	2	0	2 1	0	0	0	0	1	0	0	0	0	0	0	0	1	0
AUX	0	0	0	2 7	0	0	0	0	0	0	0	0	0	0	0	5	0
CONJ	0	1	4	0	1 1	0	0	0	0	0	0	0	0	0	0	0	0
DET	0	1	0	0	0	4 5	0	0	0	0	2	0	0	0	0	0	0
INTJ	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
NOUN	3	0	0	0	0	0	0	7 1	0	0	0	0	0	0	0	1	0
NUM	0	0	0	0	0	0	0	0	1 7	0	0	2	0	0	0	0	0
PART	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0
PRON	0	0	0	0	0	0	0	2	0	0	2 0	0	0	0	0	0	0
PROPN	0	0	0	0	0	0	0	0	0	0	0	4 5	2	0	0	0	0
PUNCT	0	0	0	0	0	0	0	0	0	0	0	0	7 8	0	0	0	1
SCONJ	0	3	3	0	1	0	0	0	0	0	0	0	0	0	0	0	0
SYM	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0
VERB	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	3 9	1
X	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	5

TABLEAU A.4 – Matrices de confusion des deux annotations manuelles fournies par les chercheuses du LiLPa (corpus `wikipedia2`).

Annexe B

Corpus bruts myriadisés

B.1 Corpus brut myriadisé sur Recettes de Grammaire (gsw)

Participant	Titre	Contenu
		Poèmes
Gamer	Serwelà	Serwelà schelle en zwei Schnitte en d'länge jeder einschnitte ùf e Daller lâje Vinaigrette d'rùf màche met de g'schellte ùn g'hàkte Zewel
Schang2	Fìngerliedla	Dàs ìsch d'r Düüma Da schìtelt d'Pflüüma Da hebt sa uf Da drehjt sa Heim Un da kleina Stumpa Frìsst sa àlla allei
Schang2	Fìngerliedla	Dàs ìsch d'r Düüma Da schìtelt d'Pflüüma Da hebt sa uf Da drehjt sa Heim Un da kleina Stumpa Frìsst sa àlla allei
Schang2	Kukus wu bìsch ?	Kukus wu bìsch ? Ìm Wàld ! Wàs hàsch ? A Frosch ! Gìb mìr eins ! Nei nei ! Gickshàlls ! Dü oï !

Schang2	Hàns ìm Schnockaloch	D'r Hans ìm Schnockaloch hàt àlles wàs ar wìll ! Un wàs ar hàt, dàs wìll ar nìt, Un wàs ar wìll, dàs hàt ar nìt. D'r Hans ìm Schnockaloch hàt àlles wàs ar wìll ! D'r Hàns ìm Schnòckaloch sajt àlles wàs ar wìll ! Un wàs ar sajt, dàs dankt 'r nìt, Un wàs ar dankt, dàs sajt 'r nìt, D'r Hans ìm Schnockaloch sajt àlles wàs ar wìll ! D'r Hàns ìm Schnòckaloch màcht àlles wàs ar wìll ! Un wàs ar màcht, dàs soll ar nìt, Un wàs ar soll, dàs màcht ar nìt. D'r Hans im Schnòckaloch màcht àlles wàs ar will ! D'r Hàns ìm Schnòckaloch geht ànna wu n ar wìll ! Un wu n ar ìsch, dò bliebt ar nìt, Un wu n ar bliebt, dò gfàllts ìhm nìt. D'r Hàns ìm Schnockaloch geht ànna wu n ar wìll ! Jetzt hàt d'r Hàns sò sàtt Un ìsch vum Elànd màtt. Un lawa, sajt ar, kàt ar nìt, Un starwa, sajt ar, wìll ar nìt, Ar sprÌngt üssem Fanschter üs, Un kummt ìn's Nàrrahüs !
Schang2	Ritta, ritta Ross	Ritta, ritta Ross Z' Bàsel ìsch a Schloss Z' Rom, ìsch a Glockahüs Do lüega drëi Jumpfra üs Eina spìnnt Siida D àndera kàt 's nìt liida D' drìtta spìnnt Hàwerstràui Hìlf dìr Gott mi liewa Fràui !
Textes libres		
Schang2	As gìt kei Tod !	As gìt kei Tod ! Àlls ìsch e n ewig Dosì ! Kei Harzschlàg vo dr chà verlore geh.- Er wìrd no witerschloh ìn dana Garta Wenn dü scho d'nìde lìgsch ìm fichte Grung. Wàs ìn dr gschräuie het dur d'làngi Nacht, Wìrd witerlabe n ìn de Büechewall Ìn jederem brietige churze Summerwatter.- Un aller Trieb vo Liebi vo dim Harz Wìr àlls no dosì ìn de Maienacht Ìn jederem Lockrüef ìn de fäischtere Baim. Dank's wie de wìtt : àlls ìsch e n ewig Dosì. Nathan Katz

Schang2	Un dü bìsch nìt emol mìt mr gànge	Un dü bìsch nìt emol mìt mr gànge So het mi Harz d'gànzi Nàcht üfgrine, wìl de nìt e Rung mìt mr gànge bìsch - Un s' war mr doch so wohl gsì, numme fir chänne z'seh wie d'Sunne ìn dim Hoor zwistzeret hatt, wie s' Fall so wit un grien vor dr glage wär, wie d'Wall so stìll gsì wäre n um di umme, Maidle. So àber hàni spot ìn dr Nàcht no gschtümt, un dankt àn e Zit wu mr nemmi Labe, - wu d'Wìrem ìn is nàge - D'Wìrem ! D'unge n ìn dr Fichtigkheit un ìn dr Fäischteri ! Un dü bìsch nìt emol e Rung mìt mr gànge hìt. - Nathan Katz
Proverbes / phrases		
Gamer	proverbe	A Kàtz mìt Hanschig fàngt kè Miis
Gamer	proverbe	A mang gschèit Hüen lìegt s Ei nawe's Nascht.
Gamer	proverbe	A Stìck Brot ìm Sàck ìsch meh Wart àss a Fadra àm Hüet.
Gamer	proverbe	As ìsch noch kè Glehrter vum Hìmmel ghèit.
Schang2	proverbe	Wàs d'r Büür nìt kennt, frìsst ar nìt.
Schang2	proverbe	Liewer Dumm àss Stumm.
Schang2	proverbe	Wàs da nìt weisch, màcht dìr nìt heiss.
Recettes		
OLCA	Spàrichel-Flan	D'Spàrichle schäle ùn d'grìene Spìtze sowie 's zàrte Teil vùm Stiel bhàlte. Se wäsche ùn se in 2 cm-làngi Stìckle schniide. D'Spàrichle ìn e Sìbb üss Metàll màche ùn àlles zàmme 5 Minüte làng ìn kochendem Sàlzwàsser àbkoche. Se àbtropfe Ion, àbkiehle. wìdder àbtropfe Ion. Ìnere Form 's Mëhl mìt de Eier beàrweite, mìt de Mìlich ùn'em licht gschmolzene Bùtter ànrìehre. sàlze, pfeffere, Mùschkàtnùss dezügenn. De Bodde vùnere Plàtt mìt Bùtter inschmìere. d'Spàrichle-Stìckle scheen ànneleje, de Teig drìwwer màche ùn ìn heissem Offe 210° (th7) 35 bis 40 Minùte koche lon.

OLCA	Gespìckter Làchsrìcke mìt grìene Lìnse ùn Meerretti-Rahm	D'griene Lìnse 12 Stùncle Iàng inweiche Ion. 's Gemìes ln kleìni Vìereck odder Drejeck schniide. Sie ìm Bütter schwìtze Ion, dànn d' Lìnse dezügenn. Ànderthàlb sovìel Wàsser dezüschìtte ùn e Stùnd làng kechle Ion. De Làchs rìschte ùn e Träschel geraicherter Speck ùm jedes Filee rùmwìckle. De Rahm gùssàrtig inkoche Ion. Mìt Meerretti, Sàlz ùn Pfeffer üss de Mìehl ànmàche. D'gespìckte Làchsfilee 3 Minüte Iàng ànbrote. Ùf e Grìeni-Lìnse Bert leje Ùn d'Sauce drum erùm nàppìere. Mìteme bìssele Kerwelskrüt, Tomàte-Wìrfel ùn zwei Gemies-Strissle denäwe ziere. (Rezept vùm Michel HUSSER "LE CERF" ìn Màrle)
Delpa	Galrìewleküeche	1. De Offe ùff 180°C (Th. 6) stelle un vorheize. 2. In ere Sàlàtschìssel 's Mahl, 's Bàckpulver, d Gewìrz, 's Sàlz ùn de Zùcker undernànder màche. 3. In e ànderi Sàlàtschìssel 's Eel, 's Àpfelmües ùn d Mìlich undernànder rüehre. 4. De flìssig Teig ùff d Mahlmìschung schìtte ùn güet undernander rüehre. D Nùsse ùn d Galerìewle dezüe gann. Wieder emol undernànder rüehre ùn de Teig in d' Form schìtte, wie mìt Eel ingerìwwe worre ìsch. 5. De Küeche 40 bìs 50 Minüte ìn de Offe schìewe. Ùffeme Gìtter kàlt ware lon.
mlibmann	Griaßpflütta	Dr Liter Mìlch ìn a Pfànna schìtta. D Mìlch wärma, bis sa kocht. Wenn d Mìlch kocht, dràb macha vum Fiir un s Griaß driischìtta, wahrend àss ma mìt'ma Holzläffel dreiht (zìmlig energisch). 's gìtt schnall a dìcka Brüehja - do müeß jeder lüega wia-n-ar's garn hàt : - weniger Griaß gìtt weicha Griaßpflütta - mehr Griaß gìtt härta Griaßpflütta Noh-n-era Minüta Dreiha ìsch's güet. Ma kàt's eifàch a so assa, noch wàrm (zum Beispiel mìt Äpfelmüeß oder Zwatschgamüeß). Oder ma kààt Bolla ("Pflütta") drüsmàcha un sa ìn'ra Pfànna mìt Eel lo braagla.

OLCA	Erdbeere ùn Rhàbàrwer-Sabayon	d'Erdbeere wäsche. entstìele ùn ìn Vìertel schniide. Fer de Sabayon : ìnere Kàssroll mìt dìckem Bodde (odder Wàsserbàd) d'Eigelbe ùm de Puderzùcker schlaawe. Schlaawe bis dàss es wiss wùrd. Zwei Sùppeleffel Wisswin dezügenn ùn wittersch zuem e Schüm schlaawe. Licht ànkoche Ion (odder im Wàsserbàd) ùn debi wittersch schlaawe. Nooch ùn nooch de Rescht Wisswin dezügenn. D'Mìschùng muess schümisch wëre ùn sich verdopple, sie muess àwwer nìt àbkoche. Vùm Fiir ewegnamme soboel dàss de Sabayon dìck genue wùrre ìsch, kàlt Ion wëre. În e Sàlàdschìssel de Rahm ùn de Vanill-Zùcker schìtte. Zuere e Chantilly schlaawe ùn mìtem kàlte Sabayon mìsche. d'Rhàbàrwer-Stìckle koche. Se ìn kleini Pfannele odder ìn kleini Plättle Ieje, dänn d'Erdbeere drùff Ieje. Mìtem Sabayon ìwwerzìehje ùn mìt de g'hobelte Màndle bstraie.
mlibmann	Kugelhopf	Umanànder màcha : s Mahl, dr Zucker, dr Sàlz, d Eier, d Bierhewa ìn dr Mìlch, küüm leih. (Àchtung : d Mìlch sott jo nìt züe wàrm see, sunscht geht d Hewa kàpütt !) Wenn àlles güet umanànder ìsch, màcht ma dr Butter dràà. (Dr Butter müeß nìt kàlt see, ìnter weiech.) Wìder güet umanànder màcha. D Meertriiwala derzüe màcha un riahra. A Kugelhopfmodel mìt Butter ischmeera un Màndla drii lìega. Dr Teig ìns Model inaschìtta. A pààr Stund loo hàwa, ìn'ma Ort wo's ehnder wàrm ìsch un ohna Durzug. (Do kààt ma dr Teig àwaschlàga, dàs heißt : mìt da Hand liicht drufschlàga, so àss'r wìder amol uffa kummt ; 's ìsch a Tràdition, àwer 's ìsch nìt notwandig.) Wenn dr Teig güet uffakumma ìsch, dr Bàchofa uf 180°C ààmàcha un, wenn ar aso wàrm ìsch, dr Kugelhopf inamàcha un 40 Minüta loo bàcha.

Antoine67	Làmmel Kotlette mit Zitronne un Thymian (ùf'em Grill)	1) In ere Schessel, unter-n-ander mìsche,'s Olive Eel, D'Zitronne Sàft un dr Tymian 2) s'Ànmàchegewertz : Sàlz, Pfaffer 4) d'Kotlette inbaitze im Olive-n-Eel, e ganzi Stund 5) uf'em Grill (barbecue) d'Holz Kohle guet wärme 5) Guet d'Kotlette in schmerre un uf de Grill màche 6) 10 Minute e jedi Sitt koche.
OLCA	Mimosa Bibbele	12 Eier koche ùn kàlt lon wäre. Mìteme Messer, e klein Stìck vùm Ùnterteil vùn de Eier àbschnide fer àss se guet stehn bliiwe. 's Käppele abschnide. Ihr kenne 's gràd schnide odder Zàcke màche àwwer ìhr mìen e bìssele Gëls debi hàn. Mìteme Leffele, 's Eigël rüsshole ùn ìn e Schìssel màche. Mìr verdrùckt se mìtere Gàwwel ùn mìscht se mìt de àndere Zuetàte : Majonääs, Seneft, Cornichon Stìckle, Meerrëtti, Cornichonbréi, Sàlz ùn Pfeffer. Guet rìehre ùn alles ìn e Plastiksäckel lëëre. 's Eck vùn dem Säckel abschnide ùn 's benùtze wie e Sprìtz. D'Eier fille ùn e bìssel meh 'nin màche fer Plàtz hàn fer d'Awwe ùn de Schnàwwel. Üssere Gälerueb, 12 Schnäwwele ùn 24 Fiessle schnide ùn demit, d'Bibbele zìere. Mìteme kleines Stìckel Olive gìbt's 2 Giggele. Àm End, màcht m'r de Deckel wìdder drùf. Voila ! Los geht's !
oblise	Serwelàsàlàt	Serwelà schelle en zwei Schnitte en d'länge jeder einschnitte ùf e Daller lâje Vinaigrette d'rùf màche met de g'schellte ùn g'hàkte Zewel
Schang2	Lammala	'S Eigaala vum Eiwiss trenna 'S Eigaala mìt'm Zucker schlàga un 'S Eiwiss zu Schnee schlàga un mìt'm Rascht mìscha 'S Mahl d'Särka un 's Bàchpulver derzüa màcha În d'Furm mìt Butter gstrìcha drii schìtta Kochzitt 30 Minuta 180 Gràd Àchtung gah dàss ma d'Furm nìt züaviel uffìllt denn as geht ufa

B.2 Corpus brut myriadisé sur Ayo ! (mfe)

Participant	Titre	Contenu
		Poèmes
Rimena Deborah	Enn pep, Enn nasion (1)	Kot pase,inn gagn bate,inn gagn viole Intel inn kokin, lot Dan petrin Tou kout mem refrin. Ala mo pe dir, mwa mo'nn plin. Trouv dimounn pe brile, to filme. Trouv dimounn tonbe , to riye. Trouv dimounn pe nwaye, To contign gete ? Eski se sa to limanite ? Enn pep, enn nasion Bizin partaz mem vizion. Enn pep, enn nasion Bizin na pena divizion. Ki to seve drwat, boukle ouswa frize Ki to met sari, ijab ou enn bout sort Ki to koz angle franse oubien kreol Pe inport to kominote, ki to ete isi Dan nou ti zil, nou tou rasanble. Anou viv an akorite Anou viv parski nou lib Anou rasanble dan nou diferans Ek aret tir vanzans. Nou sel relizion se limanite Anou viv dan enn mond kot dimounn nepli get ras ni kas pou tir twa dan dife. Kot nou tou kontribie Ki pou per Laval , Maha shivaratree , fet sinwa ou mem mars lor dife. Nou tou kiltir melanze Divizion nou na pa le Enn sel Pep, enn sel nasion nou ete. Nou pa kritike , nou pa zize.

Rimena Deborah	Enn pep, Enn nasion (2)	Parski nou viv dan enn diversite Kot nou lespri oblize devlope. Anou get divan Ansam , solider Pou fer enn nasion meyer. Anou aret zize Aret kritike Aksepte nou diferans E donn nou tou enn sans Pou briye. Pou exprime ki to zenn, ki to vie nou tou met lame dan lapat pou nou pei avanse Pou nou nasion briye Rouz, ble ,zonn ,ver anou less nou pavion flote avek fierte Anou kiltiv lamour Pou ki nou pep sirviv a tou bann fleo lasosiete. Anou kre enn pep kot nou kouler nou longer Nou ras nou kas Pa inportan pou fer nou respekte. Kot to diversite panse pa enn rezon pou fer twa kritike. Anou tou ansam kre enn Pep ek nasion pli bon, san koripsion , san malediksion, kot tou problem ena so solision. Kot lamour ranplas laenn Kot divizion fer plas a linion Wi nou le enn pep enn nasion koumsa mem Tan ki nou ena mem disan ki koule dan nou lavenn nou pa bizin ena laenn Okontrer pran la penn Aksepte ki nou diferans se pa enn problem.

Absc0nse	PAROL SOLEY	Saler soley transpers mo lapo Li resof frwader mo andan Frwader mo santiman Li ekler mwa dan mo tonbo Mo get lao Li dir mwa sa so travay depi toultan Li resof bann mor vivan Li tir zot depi zot kaso Li dan so natir resofe Kouma li dan mo natir tonbe-leve Enn niaz pase trouble nou konversasion Nou perdi koneksion me li pou revini Nou perdi koneksion me mo rapel so parol beni
FLOEZI	Tras to bann pa	Lor laplaz tras to bann pa pe efase To lavwa pe anvole, fonn dan nwarte Fey badamie tom enn a enn dan silans Melankoli fer tom lapli to labsans Vag apre vag pe kares to souvenir Lor mo lapo zot finn les zis enn soupir Bann niaz gri ape trayi mo malsans To silwet pe disparet, mo dezesperans Soley, retourne ! Souy tou bann larm lor mo vizaz detranpe Amenn klarte dan nwar Lor laplaz tras to bann pa finn efase Aster mo nam anflame anvole !

FLOEZI	Les twa ale	Inn ariv ler pou to ale Pou to bann pa zot anvole Pou to lespri li repoze Kan to lizie zot referme Mo ti krwar to finn amare Ki to nam perdi dan nwarte Me to sourir kontign danse To lavwa plin limansite Aster to nam, li libere Tou to douler inn efase Mo'nn pran letan pou mo konpran Ki se mwa ki finn ansene Inn ariv ler pou mo avanse Pou mo bann pa swiv zot sime Pou konkretiz mo destine Dan enn lespwar demezire Mo get devan, kit tou deryer Mo vizaz reflet lalimier Tou sa soufrans, mem to labsans Mo'nn aksepte, mo les ale Mo pou rapel tou to koze Sak fwa ki nou de ti riye To sourir pou inond mo lazourne Aster ki mo les twa ale
Marine D. Henri	Lavi court	Ou lavi li defile koumadir enn bobinn difil, alor viv li avan ki li fini !
		Textes libres
laural	Reconesens	mo nouri lisien, ler voler vini mwa ki bisin zape
begnan	La rivier Tanier	Mo passe la rivier Mo zoine ein vier gran mama Mo dir li ki li fer la Li dir mwa li lapess cabo wai wai mo zenfan fo travail pou gagne son pain (bis)
laural	sirandane a siroter	Ena ene mamzelle, li suivre mo partout, mais zames mo capav embras li ?
laural	Diab marie ena pie pima	gard dan zoin nourri so lever pe met choula to enn Lalo to lagel garte mord to Lalang avan to koze mopie dourri roupi kare

AnnesyB	Mo lanfans	Souvan kan mo asize, mo mazine kouma lavi ti ete kan mo ti ankor tipti. Bann kouzin ek kouzinn ti res dan mem lakour avek nou e nou gran mama ti ankor lamem pou vey nou. Dan konze lekol, nou pa ti al pas vakans kot lot fami parski ti ena deza zanfan dan lakour pou nou zwe. Nou ti lev boner toulezour, fer nou louvraz vit-vit pou nou gagn nou lazourne pou nou zwe. Nou ti zwe lakaz zouzou, kout maye, sot lakord, lamarel, kanet. Nou ti ousi aranz boul avek enn ta plastik ki nou ti kol avek latres kolant, pou zwe foutborl, kokin balie koko pou aranz servolan ou rod bouson ek enn ti plans pou zwe badminton.
Proverbes / phrases		
hani	proverbe	Zafer mouton napa zafer kabri
begnan	proverbe	Zot dir dan vier caraill ki ena bon lasos
Rimena Deborah	proverbe	Si to le viv dan lape, met to lespri avek lalimier -Natty Jah
FLOEZI	proverbe	Ti koson riy nene so mama
FLOEZI	proverbe	Labou riy lamar
begnan	proverbe	zamai mo pou blier letan lontan Non zamai mo pou blier qui nou kapav ale pli loin zamai mo pou blier ki nu kapav depass nu limit si dan vant nou mama nu finn grandi pena simin ki nu rest tipti kan nu oule resi nu la vie
laural	proverbe	Lalang pena lezo
Recettes		
hani	Pima farsi	Fann pima an de, tir lagrin e apre rins andan pima la me les lake pima la. Sot sanpignon ase dan diber, gard pou pli tar. Kraz lay. Dan enn gran bol, mett laviann ase, laser sosiss, azout lay, zwanion ase, sampignon kwi, disel ek dip-wav e dizef-la. Batt tou sa la ensam ziska li vinn enn lapat konpak. Farsi bann pima-la avek preparation presedan. Dan enn pwal ase ot, kwi bann pima farsi-la ziska ki zot bien dore lor tou kote. Egoutt zot lor enn sopalin et servi desuit

hani	Poutou	Gres enn moul 8 opuses ki pass dan microwave avek enn tigit diber. Fonn leress diber-la dan enn deksi e azout dile so. Azout tou leress ingredian e melanz bien avek enn spatil. Lapat ki pe forme bizin ena enn konsistans versan e imid. Vers lapat-la dan moul. Pa kouver li. Kwi dan microwave lor reglaz temperatir maximal pendan 8-9 minit. Pran enn ver ou enn zafer ron pour dekoup bann gato an form rond.
hani	Boulet sousou	Plis sousou, tir so leker ek rap li. Met disel e melanze ar lame pandan 3 a 4 minit pou tir tou delo ki ena sousou-la. Apre sa, pers li e met li apar. Rousi laviann ase (oubien krevet) lor ti dife. Azout sousou rape. Azout lapoud kanz, lasos soza, ek enn parti lake zwanion-la. Apre nek zis bizin petri melanz-la. Avan fer bann boul, vers delwil sezam dan kre lame. Fer bann boul. Kwi bann boulet-la dan steamer pendant 15 minit. Servi avek inpe lake zwanion dan enn bouyon oubien avek enn lasos pima ek lasos soza.
Marine D. Henri	Rougaille saucisse kreol	Bouil saucisse la pendan 5 minits Les li egoute inpe Fer saut inpe zonion, lail ek zinzam Azout ou pomme damour ziska li rann deluile Met ou dithin apre larg ou saucisse ladan Pou fini, met inpe cotomili lor la Bon appétit

Vaness	Di riz saffrané ek poisson salé	Dan ene casrol met di riz boui dan 4 tass de lo lor gran difé. Azout saffron, disel ek ti lanis. laiss boui pou 10 minit. di riz bizin rest ferm. (cui mé en grain) Apré 10 minit, transfer di riz dans paspirer, laisse refrawdir pendant ankor 10 minit. Reserv pou pli tar. Sof karail lor ti difé, azout 1/4 tass de lhuil, azout laye, zwaynon asé, ti poi ek poisson sale. laiss frir pendant 3 minit, pass couyere 4 a 5 fwa. Azout di riz ek rezin sek, pass couyere pou melanz tou ensam environ 5 a 6 fwa, pendant 5 a 6 minit, ziska ki parfin poisson sale ek zwanyon komens fané. Teingn difé. Azout kotomili asé, aroz partout lor di riz. (mo rekomann akompagn sa pla la ek ene bon ti satini pom damur bien for)
evey	cari poisson ek brinzelle	1. couple brinzelles en ti morceaux apres laisse li trempe dans de lo avec ene peu di sel pou ene 5-10 minutes 2. tire di lo ek commence frire brinzelles batch par batch dans ene pe de l'huile apres mette dans ene bol - reserver 3. faire marine poisson dans ene peu di sel ek di poivre apres faire frire - reserver 4. frire zonions ek l'ail dans de l'huile - azoute tomates, meti - di thym - piment - melanze tout ek cuit ziska la sauce la vinne ene peu epais 5. azoute poisson (deja frire) ek brizelles (deja frire) - mete di sel/di poivre laisse cuit lor ti di fe 6. taigne di feu ek servi chaud avec du riz
natsumoanpo	Gali dossi	Dans une tasse - versez le gali - ensuite ajoutez le sucre - ensuite le lait (liquide ou en poudre) - puis l'arachide NB : Si lait en poudre alors ajoutez de l'eau + glace si vous le souhaitez

Annexe C

Variantes graphiques myriadisées

C.1 Variantes graphiques myriadisées sur Recettes de Grammaire (gsw)

token original	variante 1	variante 2	variante 3	variante 4	variante 5
’r	er				
a	e				
abschnide	àbschniida				
alles	àlles				
ànbrote	ààbrota				
ànmàche	ààmàcha				
Àpfelmües	Àpfelmùs	Àpfelmüas	Äpfelmüeß	Äpfelmùß	
assa	asse				
Bàckpulver	Bàckbulver				
Beispiel	Bispììl				
Bibbele	Bibbala				
bis	bés				
bìssele	bìtzala				
bìtsi	bessel	béssel	bìtzi		
Bolla	Bolle				
braagla	braadle				
Brüehja	Brej				
Bùtter	Butter				
com	vom				
d’Kotlette	d’Kotlett				
dänn	dernoh				
dàss	àss				
de	da				
debi	derbii				
demit	dermìt				
denäwe	dernawa				
dezügenn	derzüagawa	dezugann			
dìcka	dégi				
Dr	D’r	De	Der		
Dr	der				
dràb	ràb	eràb			
Dreiha	Drahja	Dreihe	draje		

dreiht	draaït	drahjt			
driischìtta	dränschitte	néngschéde			
drüs	d'rüs				
e	a				
Eel	Eil				
eifàch	emfàch				
Eigelbe	Eigal				
energisch	énèrgisch				
Erdbeere	Arbeere	Erdbéére	arbere	Ardbeera	
ere	'ra				
ewegnamme	awagnamma				
g'hobelti	g'robti				
Galrìewle	Galerewle	Galerieble	Galriawla	Galerìewle	Galriawla
Galrìewleküeche	Galerewlekùche	Galeriebleküecha	Galrìawlaküacha		
gann	gawa	màche			
Gewìrz	Gwìrz				
Gewìrz	Gwìrz				
gìtt	gebt	gébt	gétt	gìbt	
Griaß	Grees	Gress	Greß		
Griaßpflütta	Greespflüdde	Greßpflütte	Griesbap	Griespflüdde	GrussFlutta
güet	gùt	güat	guet		
härta	herdi	härti			
hàt	het				
Holzläffel	Holzleffel				
Ieje	lega				
ìn	én				
Ion	lo				
ìsch	esch				
ìwwerzìehje	ìwwerziaga				
jeder	jéder				
kàt	kàät	kànt	kànn	kàt's	
kenne	känna				
kleini	kleina				
Kotlette	Kotlett				
Küeche	Küacha				
Kugelhopf	Kugelhupf				
Kugelhopfmodel	Kugelhupmodell				
Labküechegewìrz	Labkùchegwìrz				
Làmmel	Làmmel				
Leffele	Läffala				
Liter	Lidder				
lüega	lüeje	lüje			
Ma	Mer				
ma	m'r	mer			
màcha	màche	macha	mache	mâche	màcha
mehr	mee	meh			
Messer	Masser				
mìen	mian				
Mìlich	Melech	Mìlch	Mélisch	melich	
Minüta	Minüt				
mìt	met				
mìt'ma	mét'm				
Mìteme	Mìt'ma				
mìtere	mìt'ra				

müeß	muß	mües	muess	müass	
Päckel	Packel				
Pfànna	Pfànn				
Pfeffer	Pfaffer				
Pflütta	Pflüdde				
Rescht	Rascht				
Rezept	Rezapt				
rüehre	rehre	rüahra			
rüsshole	üssahola				
s	es				
sa	se				
Sabayon	Zabayon				
Sauce	Soßa				
schìewe	schtelle	schìawa			
schìtta	schéde	schìtte			
Schlaawe	Schlàre				
schlaawe	schlàre				
schniide	schneide				
Schüm	Schaum				
see	sìì				
sie	sìe				
sö	so				
Spàrichle	Sparkla	Spàrkla	Spàrgla	Spàrichel	Spàrigla
Sùppeleffel	Suppaläffel				
Teig	Deg	deg			
Tymian	Tymiàn				
Ùffeme	Uff'ma				
ùn	und	un			
undernànder	undernander				
Üssere	üss'ra				
vorheize	vorhaïze	vorheiza			
vùn	vu				
vùnere	vun'ra				
wärma	werme				
wäsche	wasche				
weicha	waïschi	wechi			
weniger	wenier	wenijer			
Wenn	Wànn				
wia	wii	wier			
wia-n-ar's	wier-er's				
Wieder	wìdder				
wiss	weiss				
Wisswin	Weisswein	Wisswii			
wittersch	witterscht				
wo	wu				
worre	worra				
wùrd	ward				
wùrre	worra				
Zewel	Zìwwala				
zìmlig	zimili	zémlisch			
Zìmmet	Zemt				
Zuetàte	Züatàt				
Zwatschgamüeß	Zwatschgamüas				

C.2 Variantes graphiques myriadisées sur `Ayo!` (`mfe`)

token original	variante 1	variante 2
pouss	opuses	
zien	zuin	
ze	Ze	
lane	lannee	
ron	rond	
pou	pour	
leres diber	leress diber	
leres	leress	
tanperatir	temperatir	
pandan	pendan	
kouyer	kuyer	couyere
vie	vier	
karay	caraill	
kontign	contign	
les	less	
Angle	angle	
Franse	franse	
Kreol	kreol	
Per	per	
lapenn	la penn	
tranpe	trampe	
lake-zwanion	lake-zwanion	
Zien	zuin	
pe	p	
Ze	jeux	
ant	ent	
lane	lannee	
lane-la	lannee la	
bizin	bisin	
zwin	zoin	
nouri	nourri	
di riz	diri	
pwasson	poisson	
poi	pwa	
conzelé	konzele	
lapoud	la poud	

Annexe D

règles de transposition déduites des variantes graphiques alignées pour l'alsacien

'ho ↔ 'ro
mas ↔ mùs
ˆà ↔ ˆàà
me ↔ mee
ˆà ↔ ˆàb
mel ↔ mìl
ˆà ↔ ˆdà
mel ↔ mél
ˆa$ ↔ ˆe$
mél ↔ mìl
ˆab ↔ ˆàl
mer ↔ m'r
ˆab ↔ ˆàb
met ↔ mesmìt
ˆal ↔ ˆén
met ↔ mìt
ˆal ↔ ˆàl
mét ↔ mìt
ˆàp ↔ ˆäp
mian ↔ mìen
ˆar ↔ ˆer
müas ↔ nd$
ˆaw ↔ ˆew
müas ↔ re$
ˆco ↔ ˆza
müas ↔ mes
ˆco ↔ ˆvo
müas ↔ mues
ˆde ↔ ˆùf
müass$ ↔ mùs
ˆde ↔ ˆte
ˆrü ↔ ˆü
ra$ ↔ raa
ˆsa ↔ ˆte
ra$ ↔ rìe
ˆsa ↔ ˆdà
rah ↔ mùs
ˆuf ↔ ˆvo
rah ↔ reih
ˆuf ↔ ˆü
rahja ↔ reiha
ˆun ↔ ˆùn
rahja$ ↔ richel
adladl ↔ agl
rahja$ ↔ reihe$
ann$ ↔ awa$
ras ↔ reihe$
ànn$ ↔ àt's$
ras ↔ res
ànn$ ↔ àt$
rb ↔ rdb
ànn$ ↔ àt's$
re ↔ rüe
ànt ↔ àt
re ↔ rie
ànt$ ↔ àt's$
re ↔ rüe
àntànn$ ↔ àt
re ↔ rie
àt$ ↔ àt$
re ↔ res
àt$ ↔ àt's$
dän ↔ de$
ries ↔ rus
dän ↔ d'r
sa$ ↔ se$
dégi$ ↔ dern
sauce$ ↔ rkl
dégi$ ↔ dìcka$
sauce$ ↔ sö$
der ↔ dr
sbap$ ↔ soßa$
dr ↔ er
see$ ↔ sùp
dr ↔ dern
see$ ↔ spflüdde$
eb ↔ er
ser ↔ rìew
eb ↔ dìcka$
ser ↔ sìì$
ébt ↔ àt's$
sie ↔ sìe
ébt ↔ étt
sie ↔ s'r
eel ↔ eil
so$ ↔ zìm
eel ↔ erb
so$ ↔ sìe
eer ↔ bùt
sup ↔ étt
ees ↔ ess
sup ↔ spflüdde$
ees$ ↔ eß$
gebt ↔ gìtt
üas$ ↔ üeß$
gebt ↔ gétt
üass ↔ üeß$
gebt ↔ ˆs
üass ↔ ües
gebt ↔ gétt
üass$ ↔ ües
gébt ↔ gìtt
üass$ ↔ üeß$
gét ↔ gìt
üat ↔ üeß$
gétgétt ↔ gìtgìbt
üat ↔ ües
gétt ↔ gìbt
üdde$ ↔ ije
gew ↔ gw
üdde$ ↔ sö$
güat ↔ guet
üega$ ↔ ìew
güat ↔ gùt
üega$ ↔ reih
gue ↔ güe
üej ↔ üeh
gue ↔ güeeß$
üej ↔ üj
guet ↔ gùt
uess$ ↔ tzi
güet ↔ gùt
uess$ ↔ uß$
h$ ↔ hr$
iaßp ↔ iesp
wo$ ↔ wii
iaßpflütta$ ↔ iesbap$
wo$ ↔ wùr
ìaw ↔ üec
wor ↔ wu$
ìaw ↔ ìew
wor ↔ wùr
ìbt ↔ ütta$
za$ ↔ wùr
ìbt ↔ ìtt
za$ ↔ zep
ichel ↔ igl
zap ↔ ze$
ichel ↔ ije
zap ↔ zìm
idde ↔ ite
zem ↔ ze$
ie ↔ ige
zem ↔ zìm
ie ↔ ije
zém ↔ weic
ieige ↔ ige
zém ↔ zim
ier$ ↔ ii$
zém ↔ zìm
ier$ ↔ iglii$
zew ↔ zep
ige ↔ eil
zew ↔ zìww
ìssel ↔ ìtt

müass$ ↔ muß$
ˆdr ↔ ˆr
mue ↔ müe
ˆdr ↔ ˆer
mües ↔ muß$
ˆdr ↔ ˆr
mües ↔ mues
ˆdr ↔ ˆnén
müeß ↔ mùs
ˆdràn ↔ ˆnén
müeß ↔ nii
ˆdriis ↔ ˆnéngs
n$ ↔ jéd
ˆen ↔ ˆìn
n$ ↔ mùß
ˆén ↔ ˆìn
n$ ↔ na$
ˆén ↔ ˆìs
n$ ↔ nd$
ˆenˆdràn ↔ ˆén
na$ ↔ na$
ˆer ↔ ˆ'r
na ↔ mùß
ˆes ↔ ˆza
nei ↔ ni$
ˆes ↔ ˆùf
nn$ ↔ ni$
ˆes ↔ ˆerˆs
nn$ ↔ niint$
ˆes ↔ ˆìs
pac ↔ nt$
ˆgann$ ↔ ˆmàche$
pac ↔ päc
ˆie ↔ ˆle
par ↔ pàr
ˆio ↔ ˆlo
màc ↔ mâc
ˆrü ↔ agl
ra ↔ raa

ree ↔ rie
bal ↔ bel
rees ↔ rus
beer ↔ béér
rees ↔ re$
beer ↔ bìtzi$
reesp ↔ rusriaßp
bei ↔ bi
reesp ↔ riaßp
bes ↔ bés
reh ↔ rüah
bés ↔ bis
rerie ↔ rìe
besbés ↔ bés
reß ↔ riaß
bess ↔ bìts
ress$ ↔ riaß$
bess ↔ béér
reßp ↔ riesp
béss ↔ bis
reßpflütte$ ↔ riesbap$
béss ↔ bìts
rg ↔ rig
bessel$ ↔ bìtzi$
rgl ↔ rìe
bessel$ ↔ bìts
rgl ↔ richel
béssel$ ↔ bìtzi$
riaßpf ↔ russf
béssel$ ↔ bìts
riaw ↔ rìe
but ↔ bùt
riaw ↔ rìew
but ↔ bìtzi$
richel ↔ rkl
da$ ↔ d'r
ra ↔ päc
da$ ↔ de$
rie ↔ sùp

t$ ↔ te$
eesp ↔ eßp
t$ ↔ ti$
ega$ ↔ eje$
t$ ↔ t'r
eif ↔ emf
ts$ ↔ tes$
em ↔ erm
ta$ ↔ zim
em ↔ erberm
ta$ ↔ ta$
en ↔ ern
ta$ ↔ te$
eneer ↔ ern
ta$ ↔ ti$
erz ↔ ez
tem ↔ ta$
erzˆun ↔ ezˆùn
tem ↔ te$
es$ ↔ gel
ter ↔ t'm
es$ ↔ eß$
tsi ↔ t'r
faf ↔ fef
tsi ↔ wec
gal ↔ gìtt
üac ↔ ìtzal
gal ↔ gel
üac ↔ üet
geb ↔ géb
üah ↔ üec
geb ↔ gìb
üah ↔ üje$
geb ↔ géb
üas ↔ üeh
geb ↔ gìb
richel ↔ s'r
géb ↔ gìb
üas$ ↔ ües

var1 ↔ var2
h$ ↔ ht$
vu$ ↔ uß$
h$ ↔ hr$
vu ↔ vùn
h$ ↔ ht$
vun ↔ vùn$
ha$ ↔ he$
vun ↔ vùn
haïz ↔ heiz
waïsc ↔ wäs
här ↔ her
waïsc ↔ zìww
härhaïz ↔ herheiz
waïsc ↔ weic
hàt ↔ het
waïsc ↔ tzi
haum ↔ hüm
wàn ↔ iàn
héde ↔ hìtte
wàn ↔ wu$
héde$ ↔ hìtta$
war ↔ wied
hel ↔ hl
wär ↔ werwen
hìawa$ ↔ htelle$
was ↔ wen
hìe ↔ hte
was ↔ vùn$
hop ↔ hup
wei ↔ wi
ia-n- ↔ ier-
wi ↔ wer
ia$ ↔ ier$
wi ↔ wied
ian ↔ iàn
üas ↔ üj
ianwär ↔ ütta$
wìdd ↔ wec

ziaga$ ↔ wùrwäs
ìssel ↔ ìtzal
ziaga$ ↔ zìehje$
jed ↔ sìì$
zim ↔ zìehje$
jed ↔ üje$
zim ↔ zìm
kà ↔ kàà
züat ↔ zuet
kän ↔ kpu
kbu ↔ jéd
kbu ↔ ken
kùc ↔ ken
kùc ↔ küec
la$ ↔ küec
la$ ↔ le$
laawe ↔ làre
läf ↔ lef
ler ↔ lr
ma$ ↔ m'r$
mac ↔ màc
mac ↔ mâc
wìdd ↔ wii
mas ↔ le$

Annexe E

Guides d'installation et d'adaptation des plateformes développées à une nouvelle langue

E.1 Installation du projet Laravel :

```
$ git clone https://github.com/alicemillour/Bisame.git
$ cd Bisame
$ cp .env.example .env
$ composer install
$ composer update
$ php artisan key:generate
```

Éditez .env et remplacez les attributs de la base de données avec les vôtres.

Remarque : en production au lieu de `composer install`, écrire : `composer install --no-dev`

E.1.1 Avant de commencer

Vous devez publier les assets de laravel-filemanager :

```
php artisan vendor:publish --tag=lfm_public
```

Vous devez lancer la migration pour créer les tables de la base de données :

```
$ php artisan migrate
```

Seed de la base de données :

```
$ php artisan db:seed
```

Cela crée un utilisateur avec lequel vous pourrez vous connecter :

```
Email : admin@admin.com
Password : 4dm1n
```

Enfin, compilez les assets :

```
$ npm install
$ npm run dev
```

Génération de fake data :

```
$ php artisan db:seed --class=DevDatabaseSeeder
```

Pour faire un rollback complet de la base de données :

```
$ php artisan migrate:refresh --seed
```

Installation de Dusk (Browser Tests) :

```
$ php artisan dusk:install
```

Ajout user Admin (changer le *password* dans l'interface après connexion)

```
$ php artisan db:seed --class=UserTableSeeder
```

Import avatars, corpus d'entraînement et translations

```
$ php artisan avatars:import
$ php artisan corpus:import
$ php artisan translations:import
```

E.2 Adaptation à une nouvelle langue

E.2.1 Adaptation de la partie Recettes + variantes (sans les annotations)

Ce guide est destiné à lister les différentes modifications à apporter au code source de Recettes de Grammaire pour créer une instance pour une nouvelle langue.

La valeur de la variable 'locale' (config/app.php) détermine quelle version de l'application est affichée. Pour l'alsacien, 'locale' : 'bisame' Pour le créole guadeloupéen, 'locale' : 'krik' Elle est utilisée ailleurs pour récupérer les éléments propres à cette application. Le guide ci-dessous s'applique dans le cas où la variable 'locale' vaut 'new_language'.

Pour créer une instance avec une nouvelle langue, il suffit de remplacer la valeur de la variable locale, ainsi que de créer les éléments détaillés ci-dessous.

E.2.1.1 Éléments de design

Les éléments de design propres à une instance données sont : (attention aux extensions)

- l'image de fond : `public/images/back-recipes-new_language.jpg`.
  En fonction de l'image de fond, il peut être utile d'adapter la valeur de transparence du bandeau supérieur dans `public/css/new_language.css`.
- le favicon visible sur l'onglet du site web : `public/images/favicon-new_language.png`
- l'image d'aperçu apparaissant lorsque l'URL est transmis par certains canaux (par exemple FACEBOOK
  `public/images/ppic-new_language.png`.

E.2.1.2 Textes et *wording*

- Les fichiers `resources/views/partials/new_language-intro.php` (explication du projet qui s'affiche pages /register et /info)
  et `resources/views/partials/new_language-charte.php` (charte affichée page /register) sont à adapter à vos langue et projet. Les logos visibles en pied de page peuvent être modifiés dans le fichier `resources/views/partials/footer.php`.
- Les différentes versions des textes affichés sur le site sont modifiables dans les fichiers du répertoire /resources/lang/new_language/. Voir en particulier home.php 'app-name', 'langue' etc.

E.2.1.3 Langues des participants

Les participants peuvent renseigner les langues qu'ils parlent dans leur profil. Celles-ci sont enregistrées dans la base de données au moyen du fichier
`database/seeds/LanguagesTableSeeder.php`. Pour *seeder* les langues en bases, exécuter :
`$ php artisan db :seed --class=LanguagesTableSeeder`

E.2.2 Adaptation de la partie Annotation

Pour activer/désactiver les fonctionnalités liées à l'annotation des recettes, commenter/décommenter les parties de codes contenues dans les balises /* Fonctionnalité d'annotation */, dans les fichiers :

- `app/Http/Controllers/RecipeController.php`
- `resources/views/welcome.blade.php`
- `resources/views/recipes/show.blade.php`
- `resources/views/recipes/_show-*.blade.php`
- `resources/views/partials/nav.blade.php`
- `resources/views/layouts/app.blade.php`

E.2.2.1 Tagset

Pour entrer le *tagset* en base de données, il faut 1. créer un fichier de *seed*
`database/seeds/csvs/new_language/postags.csv` au format suivant :

name ;full_name ;description
PROPN ;Nom propre ;"Les noms propres ne sont pas des noms communs." NOUN ;Nom commun ;"Les noms communs ne sont pas des noms propres. Ex : . . . " etc.

Remarques : - Les descriptions correspondent au guide d'annotation qui sera présenté aux participants, dans nos versions nous donnons une liste d'exemples correspondant à la catégorie ainsi qu'éventuellement des consignes additionnelles (balises `<u ATTENTION</u ne pas confondre avec... etc.`) - Les descriptions sont en HTML ce qui permet de leur ajouter des éléments de style.

2. Exécuter `$ php artisan db :seed --class=PostagTableSeeder`
   (voir le fichier `database/seeds/PostagTableSeeders.php` au besoin.)

E.2.2.2 Scripts de prétraitements

Les corpus ajoutés au moyen de la plateforme ainsi que les versions préannotées sont stockées dans le dossier `storage/app/new_language`. Veillez à accorder les bonnes autorisations d'écriture dans ce dossier.

Avant de pouvoir être annotées, les recettes saisies sont soumises à un certain nombre de prétraitements effectués par des scripts situés dans le dossier `scripts/new_language/`

Tous les appels aux scripts sont réalisés dans le fichier
`app/Http/Controllers/RecipeController.php` et peuvent y être modifés au besoin.

Les scripts de prétraitements utilisés par notre version sont :

1. **Tokenisation** : `scripts/new_language/tokenize.sh`

   Entrée : *file.txt* (dossier : `storage/app/new_language/corpus/raw/`)

   Sortie : *file.txt.tok* (dossier : `storage/app/new_language/corpus/tokenized/`)

   exemple de sortie *recettes.txt.tok* :
   `D' grìene Lìnse 12 Stùnde làng inweiche lon .`

2. **Transformation en "*seed*"** pour « peupler » la base de données avec le corpus brut :
   `scripts/new_language/word_to_seed.sh`
   Entrée : *file.txt.tok* (dossier : `storage/app/new_language/corpus/tokenized/`)

   Sortie : *file.txt.word_seed* (dossier : `storage/app/new_language/corpus/word_seed/`)

   exemple de sortie *recettes.txt.word_seed*

   ```
   corpus_name;sentence_position;position;value
   recettes;1;1;D'
   recettes;1;2;grìene
   recettes;1;3;Lìnse
   recettes;1;4;12
   recettes;1;5;Stùnde
   recettes;1;6;làng
   recettes;1;7;inweiche
   recettes;1;8;lon
   recettes;1;9;.
   ```

3. **Préannotation** (réalisée dans notre cas avec MElt entraîné pour la langue considérée)
   `scripts/new_language/preannotate.sh`
   Entrée :

   file.txt.tok (dossier : `storage/app/new_language/corpus/tokenized/`)

   Sortie : *file.txt.preannotated*
   (dossier : `storage/app/new_language/corpus/preannotation/preannotation-1/`)

   Exemple de sortie *recettes.txt.preannotated* :

   ```
   D'/DET grìene/ADJ Lìnse/NOUN 12/NUM Stùnde/NOUN làng/ADV inweiche/VERB
   lon/VERB ./PUNCT
   ```

4. **Transformation en "*seed*"** pour peupler la base de données avec le corpus pré-annoté : `scripts/new_language/preannotation_to_seed.sh` (transforme le format brown en format *seed*, le score de confiance `confidence_score` étant fixé arbitrairement à 10)

   Entrée : *file.txt.preannotated*
   Sortie : *file.txt.preannotation_seed*
   exemple de sortie : *recettes.txt.preannotation_seed*

```
corpus_name;sentence_position;word_position;
           value;postag_name;confidence_score;tagger
recettes;1;1;D';DET;10;MElt
recettes;1;2;grìene;ADJ;10;MElt
recettes;1;3;Lìnse;NOUN;10;MElt
recettes;1;4;12;NUM;10;MElt
recettes;1;5;Stùnde;NOUN;10;MElt
recettes;1;6;làng;ADV;10;MElt
recettes;1;7;inweiche;VERB;10;MElt
recettes;1;8;lon;VERB;10;MElt
recettes;1;9;.;PUNCT;10;MElt
```

E.2.2.3 Ressources à fournir pour la formation

Il est possible de configurer une formation obligatoire pour certaines parties du discours à annoter, dans ce cas, il faut fournir pour chaque étiquette `TAG` un corpus de formation `corpus_TAG` tel qu'il existe des occurrences de mots à étiqueter `TAG` et de mots pouvant être confondus avec des mots de catégorie `TAG`.

Exemple en français :

Soit la tâche d'annotation en parties du discours avec le *tagset* suivant :
[ADJ;ADP;PUNCT;ADV;AUX;SYM;INTJ;CCONJ;X;
NOUN;DET;PROPN;NUM;VERB;PART;PRON;SCONJ]

et la phrase : « Franck et Théo sont donc partis. ».

La phrase doit être annotée :

« Franck/PROPN et/**CCONJ** Théo/PROPN sont/VERB donc/**ADV** partis/VERB ./PUNCT ».
Le mot *donc* est ambigu, car pourrait être annoté CCONJ dans un autre contexte.

Considérant cet exemple, les fichiers à fournir pour l'entraînement de la catégorie CCONJ seraient les suivants :

1. *CCONJ_corpus.csv*
   Remarque : is_training = 1, is_active = 1

```
name;is\_training;is\_active
corpus\_CCONJ;1;1
```

2. *CCONJ_words.csv*

```
corpus_name;sentence_position;position;value
corpus_CCONJ;1;1;Franck
corpus_CCONJ;1;2;et
corpus_CCONJ;1;3;Théo
```

```
corpus_CCONJ;1;4;sont
corpus_CCONJ;1;5;donc
corpus_CCONJ;1;6;partis
corpus_CCONJ;1;7;.
```

3. *CCONJ_annotations.csv*
   Remarques :
   - la valeur de confidence score n'est pas prise en compte
   - les annotations proposées au participant durant la phase de formation (`tagger = training`) doivent être suivies par les annotations correctes attendues (`tagger = solution`).

```
corpus_name;sentence_position;word_position;
                        value;postag_name;confidence_score;tagger
corpus_CCONJ;1;1;Franck;10;PROPN;training
corpus_CCONJ;1;2;et;10;CCONJ;training
corpus_CCONJ;1;3;Théo;10;PROPN;training
corpus_CCONJ;1;4;sont;10;VERB;training
corpus_CCONJ;1;5;donc;10;CCONJ;training
corpus_CCONJ;1;6;partis;10;VERB;training
corpus_CCONJ;1;7;.;10;PUNCT;training
corpus_CCONJ;1;1;Franck;10;PROPN;solution
corpus_CCONJ;1;2;et;10;CCONJ;solution
corpus_CCONJ;1;3;Théo;10;PROPN;solution
corpus_CCONJ;1;4;sont;10;VERB;solution
corpus_CCONJ;1;5;donc;10;ADV;solution
corpus_CCONJ;1;6;partis;10;VERB;solution
corpus_CCONJ;1;7;.;10;PUNCT;solution
```

Une fois ces fichiers créés et placés dans le dossier correspondant au nouveau langage exécuter :
`$ php artisan corpus :import CCONJ`
(voir au besoin : `app/Console/Commands/ImportTrainingCorpus.php`).

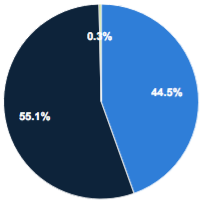
Annexe F

Résultats des enquêtes menées sur la pratique en ligne de l'alsacien et du créole mauricien

F.1 L'alsacien, Internet, et vous

Êtes-vous...

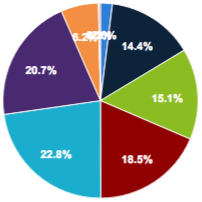
Chart options »



Un homme	541
Une femme	670
Ne souhaite pas répondre	4

Quel âge avez-vous ?

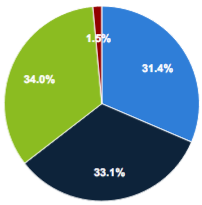
Chart options »



Moins de 20 ans	24
Entre 20 et 29 ans	176
Entre 30 et 39 ans	184
Entre 40 et 49 ans	226
Entre 50 et 59 ans	278
Entre 60 et 69 ans	253
Entre 70 et 79 ans	76
Plus de 80 ans	1
Ne souhaite pas répondre	3

Votre langue maternelle est...

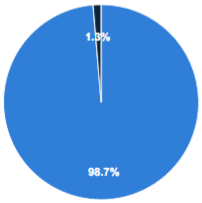
Chart options »



le français	384
l'alsacien	404
les deux	415
autre	18

Précisez

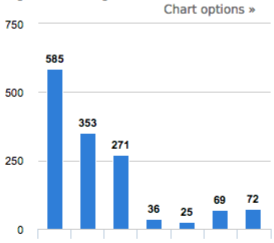
Chart options »



Laissé vide	1206
Valeur entrée par l'utilisateur	16
Longueur moyenne des soumissions en mots (sans les blancs)	1.81

Vous parlez... (plusieurs réponses possibles)

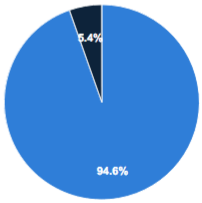
Chart options »



le bas alémanique du nord (bas-rhinois)	585
le bas alémanique du sud (haut-rhinois)	353
l'alsacien de Strasbourg	271
le francique rhénan lorrain (platt)	36
le francique rhénan méridional (palatin)	25
autre (précisez)	69
je ne sais pas	72

Précisez

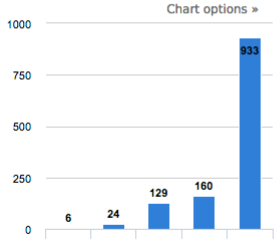
Chart options »



Laissé vide	1158
Valeur entrée par l'utilisateur	66
Longueur moyenne des soumissions en mots (sans les blancs)	5.05

Avez-vous un investissement professionnel et/ou associatif lié à l'alsacien ?

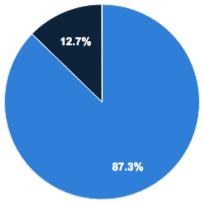
Chart options »



Oui, en tant que fonctionnaire d'une organisation pour la promotion de	6

Précisez le cadre de votre investissement

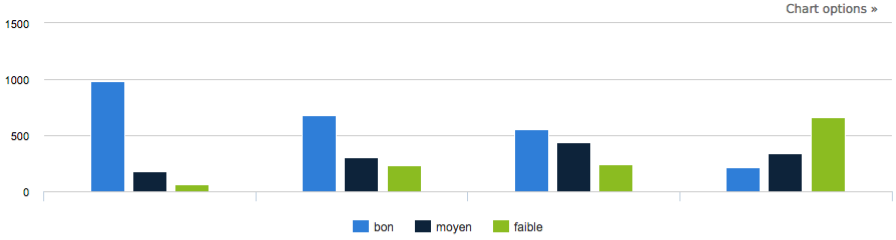
Chart options »



Laissé vide	1068
Valeur entrée par l'utilisateur	156
Longueur moyenne des soumissions en mots (sans les blancs)	6.24

Évaluez votre niveau en alsacien

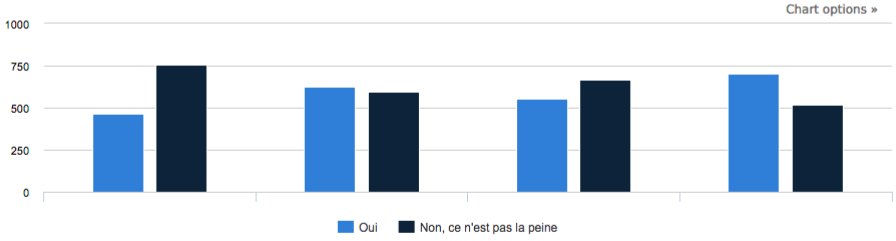
Chart options »



	bon	moyen	faible
Compréhension orale	981	182	59
Production orale	681	306	235
Compréhension écrite	550	435	237
Production écrite	215	343	664

Vous ameriez améliorer votre...

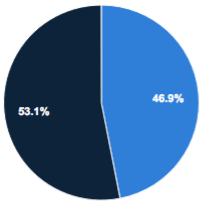
Chart options »



	Oui	Non, ce n'est pas la peine
Compréhension orale	464	757
Production orale	627	594
Compréhension écrite	553	668
Production écrite	702	520

Utilisez vous l'alsacien sur Internet ?

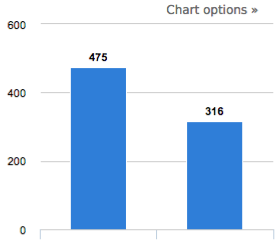
Chart options »



Oui (même rarement)	574
Non, jamais	649

Si oui, que faites-vous en alsacien ?

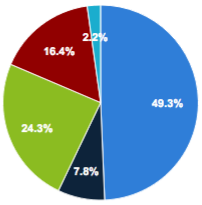
Chart options »



Je lis des contenus (articles, publications, commentaires) en alsacien	475
J'écris des contenus (articles, publications, commentaires) en alsacien	316

Écrivez-vous l'alsacien ?

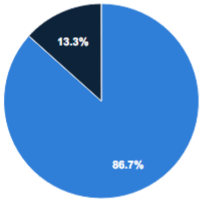
Chart options »



Oui (même rarement)	603
Non, car l'alsacien est une langue orale que je ne souhaite pas écrire	95
Non, car je ne saurais pas comment l'écrire	297
Non, car je n'en ai pas l'occasion	200
Non, pour une autre raison	27

Précisez

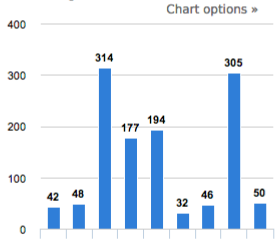
Chart options »



Laissé vide	1059
Valeur entrée par l'utilisateur	163
Longueur moyenne des soumissions en mots (sans les blancs)	7.70

Lorsque vous écrivez en alsacien (en ligne ou non), qu'écrivez-vous ?

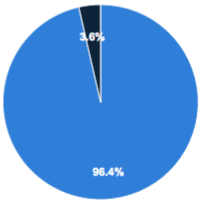
Chart options »



Des articles informatifs (informations, articles de blog, histoire de l'Alsace, histoire familiale etc.)	42
Des contenus littéraires (pièce de théâtre, nouvelle, roman, poésie etc.)	48
Des commentaires à des publications en alsacien	314
Des blagues	177
Des lettres ou des e-mails	194
Des recettes de cuisine	32
Des avis politiques	46
Vous conversez par écrit en alsacien sur les réseaux sociaux (t'chat, Twitter, Facebook messenger etc.)	305
Autres (précisez ci-dessous)	50

Donnez des exemples de contenus que vous écrivez en alsacien (en ligne ou non)

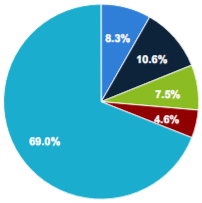
Chart options »



Laissé vide	1180
Valeur entrée par l'utilisateur	44
Longueur moyenne des soumissions en mots (sans les blancs)	4.75

Lorsque vous écrivez, utilisez vous la graphie issue de la méthode ORTHAL (http://www.orthal.fr/) ?

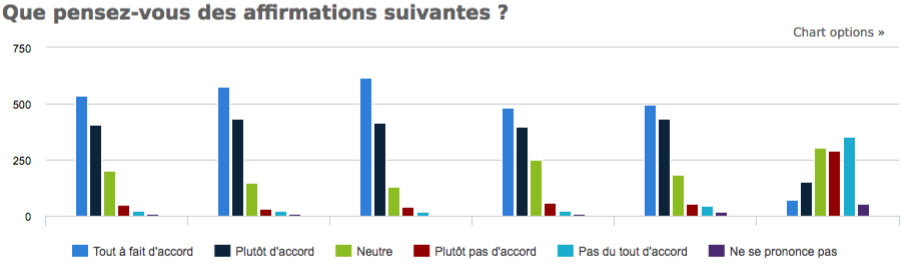
Chart options »



Oui, toujours.	50
Oui, ça m'arrive.	64
Non, j'aimerais l'utiliser mais je ne la maîtrise pas.	45
Non, je connais cette méthode mais refuse de l'utiliser.	28
Non, je n'avais jamais entendu parler de cette graphie.	416

Que pensez-vous des affirmations suivantes ?

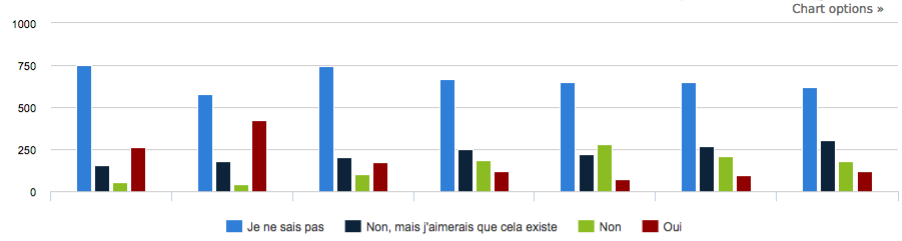
Chart options »



	Tout à fait d'accord	Plutôt d'accord	Neutre	Plutôt pas d'accord	Pas du tout d'accord	Ne se prononce pas
J'aimerais pouvoir utiliser l'alsacien dans tous les aspects de ma vie	534	406	203	47	21	11
Utiliser l'alsacien sur Internet permet de montrer que c'est une langue dynamique	574	433	148	33	23	11
Une meilleure présence de l'alsacien sur Internet le rendrait plus attractif aux nouvelles générations	616	415	128	40	17	6
Cela vaut la peine d'utiliser l'alsacien sur Internet	481	396	251	60	23	10
C'est plus facile d'utiliser le français que l'alsacien en ligne	494	434	181	53	44	16
Il est impossible d'utiliser l'alsacien sur Internet	70	151	303	290	352	55

À votre connaissance, les ressources suivantes sont-elles disponible en ligne ?

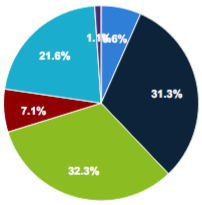
Chart options »



	Je ne sais pas	Non, mais j'aimerais que cela existe	Non	Oui
dictionnaire monolingue en ligne	748	156	56	259
dictionnaire bilingue en ligne	575	180	44	422
dictionnaire de prononciation en ligne	745	202	99	175
correcteur orthographique	664	252	186	119
clavier de saisie adapté à l'alsacien	647	223	277	74
correcteur orthographique	647	268	210	96
traducteur automatique	619	304	180	118

Participer à la production collaborative de ressources en ligne pour l'alsacien vous paraît

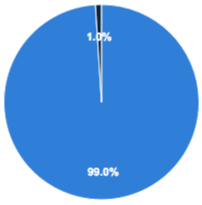
Chart options »



Une bonne idée, je le fais déjà !	81
Une bonne idée, mais je ne sais pas comment faire.	382
Une bonne idée, mais je n'ai pas le temps.	395
Trop compliqué pour moi, je ne suis pas à l'aise avec l'informatique	87
Trop compliqué pour moi, mon niveau d'alsacien n'est pas assez bon	264
Une mauvaise idée (expliquez pourquoi ci-dessous)	13

Pourquoi ?

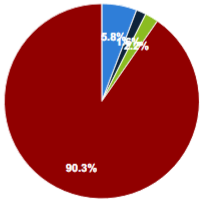
Chart options »



Laissé vide	1212
Valeur entrée par l'utilisateur	12
Longueur moyenne des soumissions en mots (sans les blancs)	11.83

Connaissez-vous l'existence de la plate-forme Bisame (http://bisame.paris-sorbonne.fr) ?

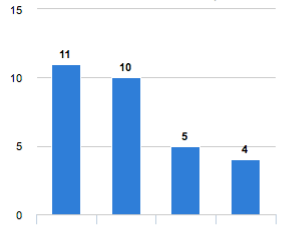
Chart options »



Oui, mais je ne me suis pas inscrit	71
Oui, je me suis inscrit mais je n'ai pas participé	20
Oui, j'y ai participé	27
Non, je n'en ai jamais entendu parler	1104

Vous avez participé, et cela vous a paru...

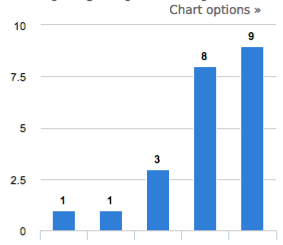
Chart options »



facile	11
difficile	10
amusant	5
ennuyeux	4

Vous n'avez pas participé à ce projet parce que...

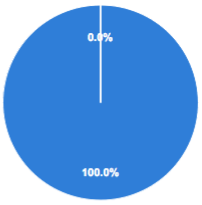
Chart options »



ça avait l'air trop difficile	1
le projet ne vous a pas plu	1
vous n'avez pas compris ce qu'il fallait faire	3
vous n'avez pas eu le temps	8
vous avez pensé participé mais vous avez oublié	9

Précisez pourquoi

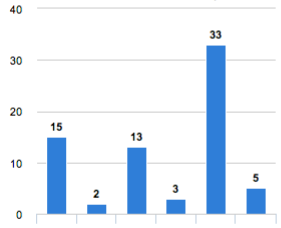
Chart options »



Laissé vide	1224
Valeur entrée par l'utilisateur	0
Longueur moyenne des soumissions en mots (sans les blancs)	0

Vous ne vous êtes pas inscrit parce que...

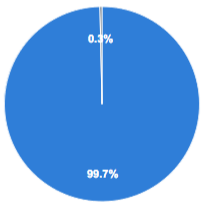
Chart options »



vous n'aviez pas envie de participer	15
vous n'aviez pas confiance dans le site Internet	2
vous n'avez pas compris de quoi il s'agissait	13
vous n'avez pas réussi	3
vous avez pensé vous inscrire plus tard mais vous avez oublié	33
autre (précisez)	5

Précisez pourquoi

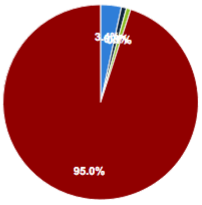
Chart options »



Laissé vide	1220
Valeur entrée par l'utilisateur	4
Longueur moyenne des soumissions en mots (sans les blancs)	6.75

Connaissez-vous l'existence de la plate-forme Recettes de grammaire (http://bisame.paris-sorbonne.fr/recettes) ?

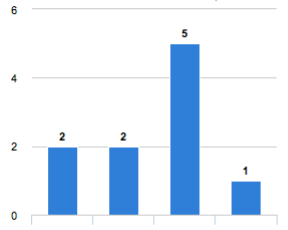
Chart options »



Oui, mais je ne me suis pas inscrit	42
Oui, je me suis inscrit mais je n'ai pas participé	10
Oui, j'y ai participé	9
Non, je n'en ai jamais entendu parler	1161

Vous avez participé, et cela vous a paru...

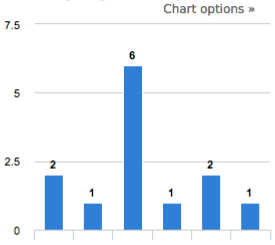
Chart options »



facile	2
difficile	2
amusant	5
ennuyeux	1

Vous n'y avez pas participé parce que...

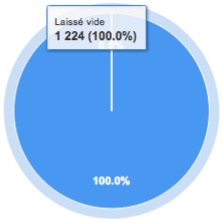
Chart options »



ça avait l'air trop difficile	2
vous n'avez pas compris ce qu'il fallait faire	1
vous n'avez pas eu le temps	6
le site fonctionne mal sur téléphone mobile	1
vous avez pensé participé mais vous avez oublié	2
vous n'avez pas de recette de cuisine	1

Précisez pourquoi

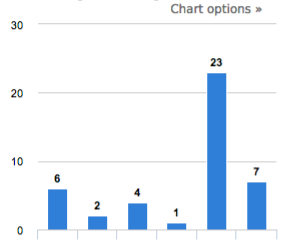
Chart options »



Laissé vide	1224
Valeur entrée par l'utilisateur	0
Longueur moyenne des soumissions en mots (sans les blancs)	0

Vous ne vous êtes pas inscrit parce que...

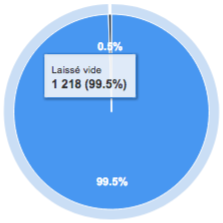
Chart options »



vous n'aviez pas envie de participer	6
vous n'aviez pas confiance dans le site Internet	2
vous n'avez pas compris de quoi il s'agissait	4
vous n'avez pas réussi	1
vous avez pensé vous inscrire plus tard mais vous avez oublié	23
autre (précisez)	7

Précisez pourquoi

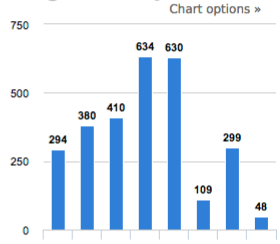
Chart options »



Laissé vide	1218
Valeur entrée par l'utilisateur	6
Longueur moyenne des soumissions en mots (sans les blancs)	4.33

Vous aimeriez que votre participation à la création de ressources en ligne vous permette...

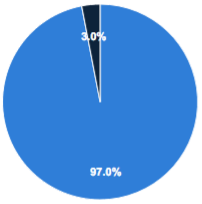
Chart options »



de rentrer en contact avec d'autres personnes parlant alsacien pour pouvoir discuter	294
de partager des contenus, des conseils, des astuces, des avis	380
de vous amuser en ligne	410
d'améliorer votre alsacien	634
d'apprendre des choses (de manière générale)	630
de gagner des bons de réduction (librairies ou événements culturels par exemple)	109
de valoriser votre connaissance de l'alsacien en participant à un projet de recherche	299
autre (préciser)	48

Précisez

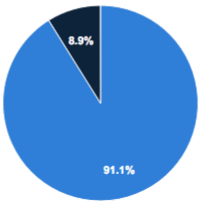
Chart options »



Laissé vide	1187
Valeur entrée par l'utilisateur	37
Longueur moyenne des soumissions en mots (sans les blancs)	7.76

Un commentaire, une suggestion, un témoignage à partager ?

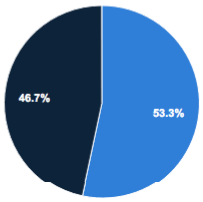
Chart options »



Laissé vide	1115
Valeur entrée par l'utilisateur	109
Longueur moyenne des soumissions en mots (sans les blancs)	38.08

Vous souhaitez être tenu(e) au courant des résultats de ce sondage ? Inscrivez votre e-mail ci-dessous.

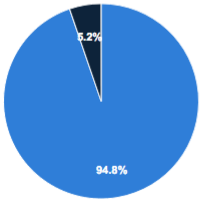
Chart options »



Laissé vide	653
Valeur entrée par l'utilisateur	571
Longueur moyenne des soumissions en mots (sans les blancs)	3.54

Vos réponses vont nous permettre d'améliorer nos sites de création collaborative de ressources en alsacien. Si vous souhaitez recevoir des nouvelles des projets Bisame et Recettes de grammaire, inscrivez votre e-mail ci-dessous. (maximum 2 mails par an)

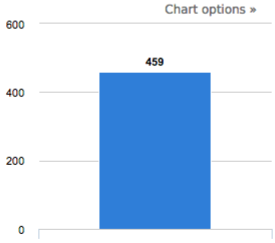
Chart options »



Laissé vide	1158
Valeur entrée par l'utilisateur	64
Longueur moyenne des soumissions en mots (sans les blancs)	3.58

Vos réponses vont nous permettre d'améliorer les sites de création collaborative de ressources en alsacien que nous développons. Si vous souhaitez également recevoir des nouvelles des projets Bisame et Recettes de grammaire, cochez cette case

Chart options »

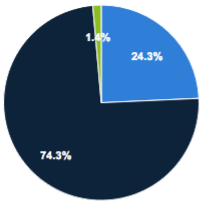


Je souhaite recevoir des nouvelles par mail (maximum 2 mails par an)	459

F.2 Le créole mauricien et sa présence en ligne

Êtes-vous...

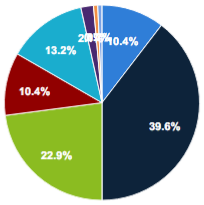
Chart options »



Un homme	35
Une femme	107
Ne souhaite pas répondre	2

Quel âge avez-vous ?

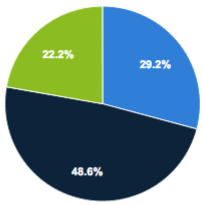
Chart options »



Moins de 20 ans	15
Entre 20 et 29 ans	57
Entre 30 et 39 ans	33
Entre 40 et 49 ans	15
Entre 50 et 59 ans	19
Entre 60 et 69 ans	3
Entre 70 et 79 ans	1
Ne souhaite pas répondre	1

Comment avez-vous eu connaissance de ce sondage ?

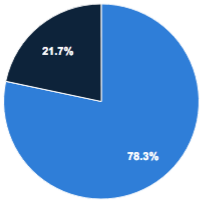
Chart options »



bouche à oreille	42
réseaux sociaux	70
autre	32

Précisez

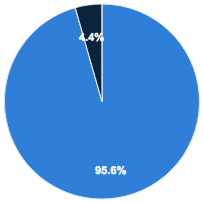
Chart options »



Laissé vide	112
Valeur entrée par l'utilisateur	31
Longueur moyenne des soumissions en mots (sans les blancs)	2.32

Sur les réseaux sociaux, grâce...

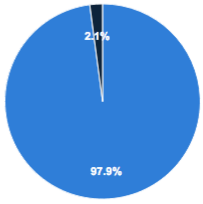
Chart options »



au partage d'un ami	65
au partage d'une page spécifique (précisez laquelle)	3

Précisez la page

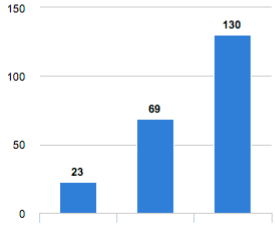
Chart options »



Laissé vide	140
Valeur entrée par l'utilisateur	3
Longueur moyenne des soumissions en mots (sans les blancs)	2.00

La ou les langue(s) que vous considérez comme maternelle(s) est/sont...

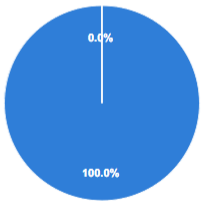
Chart options »



l'anglais	23
le français	69
le créole mauricien	130

Précisez votre autre langue maternelle

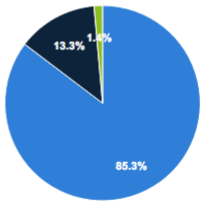
Chart options »



Laissé vide	144
Valeur entrée par l'utilisateur	0
Longueur moyenne des soumissions en mots (sans les blancs)	0

Avez-vous vécu à l'île Maurice ?

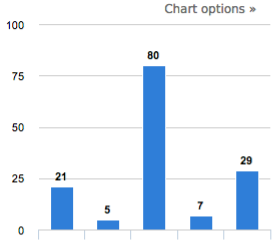
Chart options »



Oui, j'y habite actuellement	122
Oui, mais je n'y habite plus	19
Non	2

Et dans quelle région de l'île ?

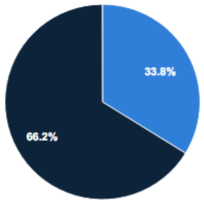
Chart options »



Nord	21
Sud	5
Centre	80
Est	7
Ouest	29

Avez-vous vécu une partie de votre vie dans un autre pays que l'île Maurice ?

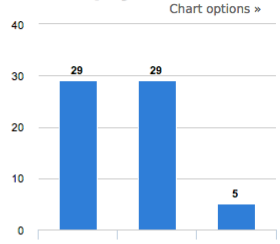
Chart options »



Oui	48
Non	94

Si oui, vous avez vécu dans un pays...

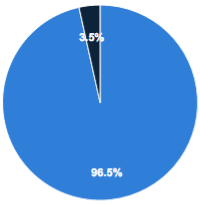
Chart options »



anglophone	29
francophone	29
autre	5

Précisez

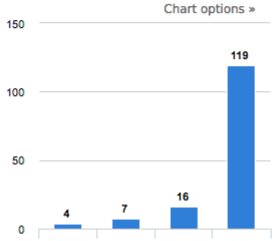
Chart options »



Laissé vide	139
Valeur entrée par l'utilisateur	5
Longueur moyenne des soumissions en mots (sans les blancs)	1.20

Avez-vous un investissement professionnel et/ou associatif lié au créole mauricien ? ?

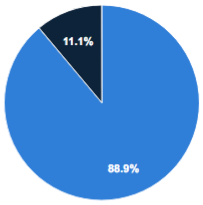
Chart options »



Oui, en tant que fonctionnaire d'une organisation pour la promotion du créole mauricien	4
Oui, comme enseignant fonctionnaire	7
Oui, dans un autre cadre (précisez ci-dessous)	16
Non	119

Précisez le cadre de votre investissement

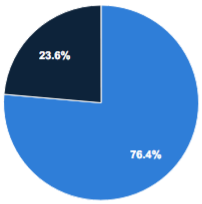
Chart options »



Laissé vide	128
Valeur entrée par l'utilisateur	16
Longueur moyenne des soumissions en mots (sans les blancs)	6.13

Pensez-vous que le créole mauricien possède des variantes ?

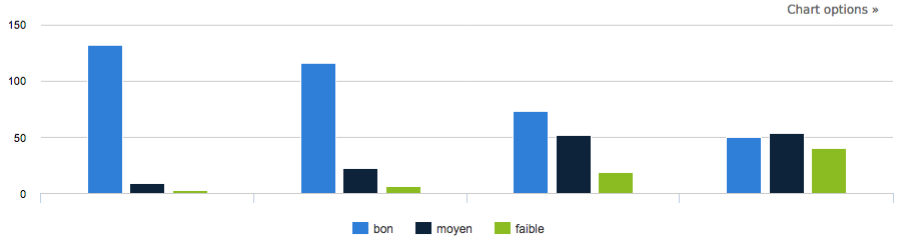
Chart options »



Oui, le créole parlé varie selon les régions	110
Non, le créole parlé est le même à travers l'île	34

Évaluez votre niveau en créole mauricien

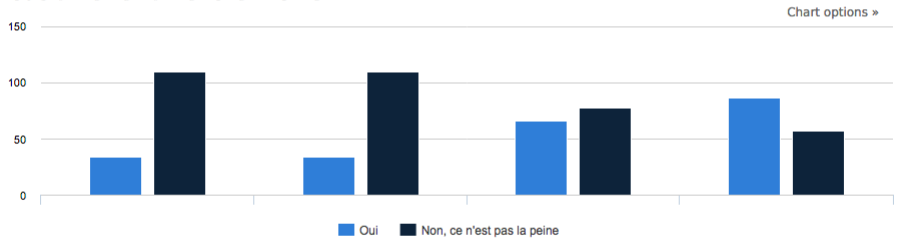
Chart options »



	bon	moyen	faible
Compréhension orale	132	9	3
Production orale	116	22	6
Compréhension écrite	73	52	19
Production écrite	50	54	40

Vous ameriez améliorer votre...

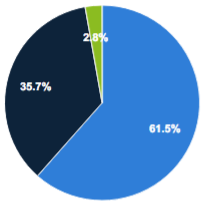
Chart options »



	Oui	Non, ce n'est pas la peine
Compréhension orale	34	110
Production orale	34	110
Compréhension écrite	66	78
Production écrite	87	57

Laquelle de ces phrases vous semble plus facile à lire

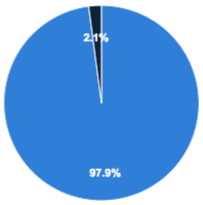
Chart options »



monn al aste dipin	88
monne alle acheté dipain	51
je ne suis à l'aise avec aucune de ces orthographes	4

Proposez une orthographe alternative : comment auriez-vous écrit "Je suis allé(e) acheter du pain" en créole mauricien ?

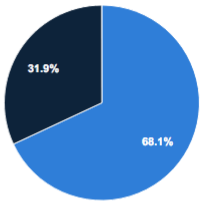
Chart options »



Laissé vide	140
Valeur entrée par l'utilisateur	3
Longueur moyenne des soumissions en mots (sans les blancs)	4.33

Utilisez vous le créole mauricien sur Internet ?

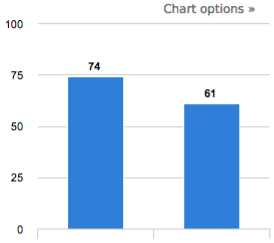
Chart options »



Oui (même rarement)	98
Non, jamais	46

Si oui, que faites-vous en créole mauricien ?

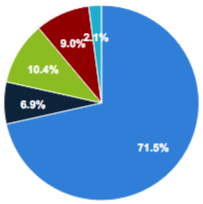
Chart options »



Je lis des contenus (articles, publications, commentaires) en créole	74
J'écris des contenus (articles, publications, commentaires) en créole	61

Écrivez-vous le créole mauricien (en ligne ou non) ?

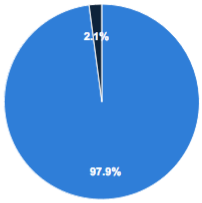
Chart options »



Oui (même rarement)	103
Non, car le créole est une langue orale que je ne souhaite pas écrire	10
Non, car je ne saurais pas comment l'écrire	15
Non, car je n'en ai pas l'occasion	13
Non, pour une autre raison	3

Précisez

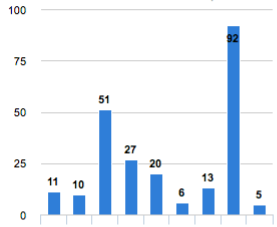
Chart options »



Laissé vide	141
Valeur entrée par l'utilisateur	3
Longueur moyenne des soumissions en mots (sans les blancs)	5.33

Lorsque vous écrivez en créole mauricien (en ligne ou non), qu'écrivez-vous ?

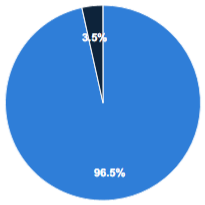
Chart options »



Des articles informatifs (informations, articles de blog, histoire de l'île Maurice, histoire familiale etc.)	11
Des contenus littéraires (pièce de théâtre, nouvelle, roman, poésie etc.)	10
Des commentaires à des publications en créole	51
Des blagues	27
Des lettres ou des e-mails	20
Des recettes de cuisine	6
Des avis politiques	13
Vous conversez par écrit en créole sur les réseaux sociaux (t'chat, Twitter, Facebook messenger etc.)	92
Autres (précisez ci-dessous)	5

Donnez des exemples de contenus que vous écrivez en créole mauricien (en ligne ou non)

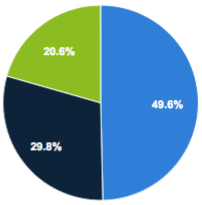
Chart options »



Laissé vide	139
Valeur entrée par l'utilisateur	5
Longueur moyenne des soumissions en mots (sans les blancs)	4.60

Quand vous n'êtes pas sûr de l'orthographe d'un mot, vous l'écrivez...

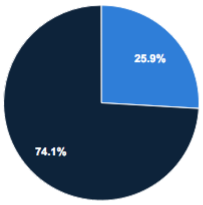
Chart options »



comme je l'entends	70
avec la graphie française	42
ça dépend	29

Sauriez-vous expliquer de quoi cela dépend ?

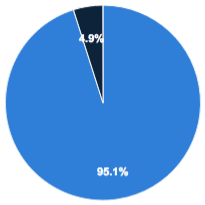
Chart options »



oui	7
non	20

Ça dépend de...

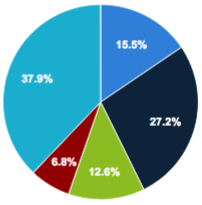
Chart options »



Laissé vide	136
Valeur entrée par l'utilisateur	7
Longueur moyenne des soumissions en mots (sans les blancs)	13.00

Lorsque vous écrivez, utilisez vous la graphie préscrite par l'Akademi Kreol Morisien dans "Lortograf Kreol Morisien" ?

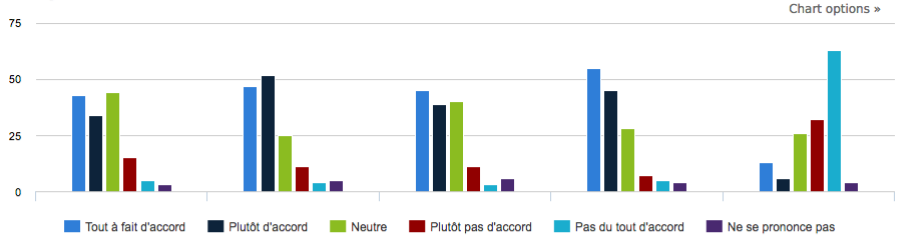
Chart options »



Oui, toujours.	16
Oui, ça m'arrive.	28
Non, j'aimerais l'utiliser mais je ne la maîtrise pas.	13
Non, je connais cette graphie mais refuse de l'utiliser.	7
Non, je n'avais jamais entendu parler de cette graphie.	39

Que pensez-vous des affirmations suivantes ?

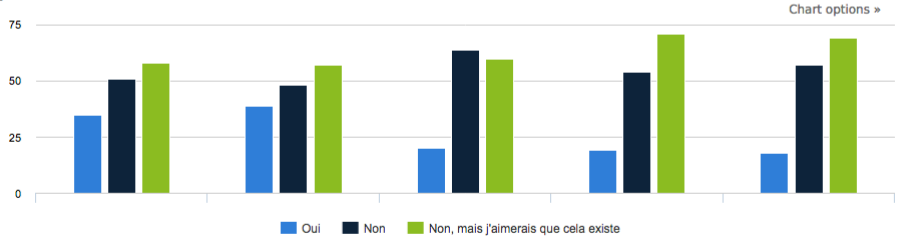
Chart options »



	Tout à fait d'accord	Plutôt d'accord	Neutre	Plutôt pas d'accord	Pas du tout d'accord	Ne se prononce pas
J'aimerais pouvoir utiliser le créole dans tous les aspects de ma vie	43	34	44	15	5	3
Une meilleure présence du créole sur Internet favoriserait la diffusion d'une graphie uniformisée	47	52	25	11	4	5
Cela vaut la peine d'utiliser le créole mauricien sur Internet	45	39	40	11	3	6
C'est plus facile d'utiliser le français et l'anglais que le créole en ligne	55	45	28	7	5	4
Il est impossible d'utiliser le créole sur Internet	13	6	26	32	63	4

À votre connaissance, les ressources suivantes sont-elles disponible en ligne pour le créole mauricien ?

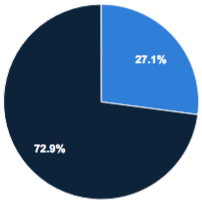
Chart options »



	Oui	Non	Non, mais j'aimerais que cela existe
dictionnaire monolingue en ligne	35	51	58
dictionnaire bilingue en ligne	39	48	57
dictionnaire de prononciation en ligne	20	64	60
correcteur orthographique	19	54	71
traducteur automatique	18	57	69

Connaissez-vous le principe de la production participative (aussi appelé "crowdsourcing") ?

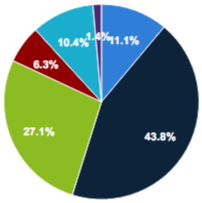
Chart options »



Oui	39
Non	105

Participer à la production collaborative de ressources en ligne pour le créole mauricien vous paraît

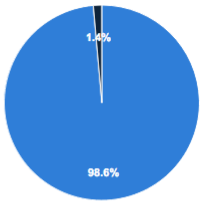
Chart options »



Une bonne idée, je le fais déjà !	16
Une bonne idée, mais je ne sais pas comment faire.	63
Une bonne idée, mais je n'ai pas le temps.	39
Trop compliqué pour moi, je ne suis pas à l'aise avec l'informatique	9
Trop compliqué pour moi, mon niveau de créole n'est pas assez bon	15
Une mauvaise idée (expliquez pourquoi ci-dessous)	2

Pourquoi ?

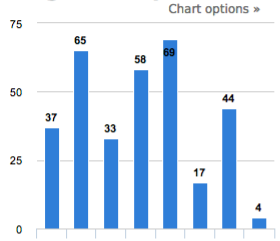
Chart options »



Laissé vide	142
Valeur entrée par l'utilisateur	2
Longueur moyenne des soumissions en mots (sans les blancs)	17.50

Vous aimeriez que votre participation à la création de ressources en ligne vous permette...

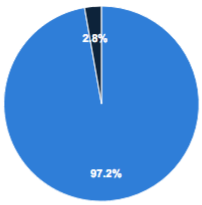
Chart options »



de rentrer en contact avec d'autres personnes parlant le créole mauricien pour pouvoir discuter	37
de partager des contenus, des conseils, des astuces, des avis	65
de vous amuser en ligne	33
d'améliorer votre créole	58
d'apprendre des choses (de manière générale)	69
de gagner des bons de réduction (librairies ou événements culturels par exemple)	17
de valoriser votre connaissance du créole en participant à un projet de recherche	44
autre (préciser)	4

Précisez

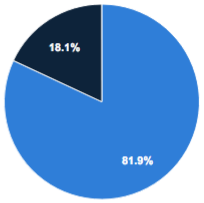
Chart options »



Laissé vide	140
Valeur entrée par l'utilisateur	4
Longueur moyenne des soumissions en mots (sans les blancs)	12.75

Un commentaire, une suggestion, un témoignage à partager ?

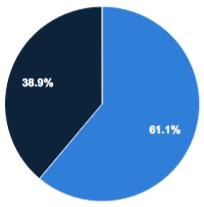
Chart options »



Laissé vide	118
Valeur entrée par l'utilisateur	26
Longueur moyenne des soumissions en mots (sans les blancs)	26.08

Vous souhaitez être tenu(e) au courant des résultats de ce sondage ? Inscrivez votre e-mail ci-dessous.

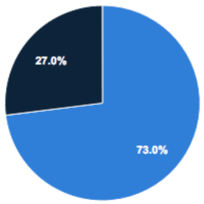
Chart options »



Laissé vide	88
Valeur entrée par l'utilisateur	56
Longueur moyenne des soumissions en mots (sans les blancs)	3.16

Vous souhaitez être averti(e) lorsque nous aurons mis en place un site de production participative pour le créole mauricien ? Inscrivez votre mail ci-dessous"

Chart options »



Laissé vide	46
Valeur entrée par l'utilisateur	17
Longueur moyenne des soumissions en mots (sans les blancs)	2.76

Annexe G

Questionnaires sur la pratique en ligne de l'alsacien et du créole mauricien

G.1 L'alsacien, Internet, et vous

Êtes-vous...

- ○ Un homme
- ◉ Une femme
- ○ Ne souhaite pas répondre

Quel âge avez-vous ?

- ○ Moins de 20 ans
- ◉ Entre 20 et 29 ans
- ○ Entre 30 et 39 ans
- ○ Entre 40 et 49 ans
- ○ Entre 50 et 59 ans
- ○ Entre 60 et 69 ans
- ○ Entre 70 et 79 ans
- ○ Plus de 80 ans
- ○ Ne souhaite pas répondre

Votre langue maternelle est...

- ○ le français
- ○ l'alsacien
- ○ les deux
- ◉ autre

Précisez

Vous parlez... (plusieurs réponses possibles) *

- ☐ le bas alémanique du nord (bas-rhinois)
- ☐ le bas alémanique du sud (haut-rhinois)
- ☐ l'alsacien de Strasbourg
- ☐ le francique rhénan lorrain (platt)
- ☐ le francique rhénan méridional (palatin)
- ☑ autre (précisez)
- ☐ je ne sais pas

Précisez

Avez-vous un investissement professionnel et/ou associatif lié à l'alsacien ? *

- ☐ Oui, en tant que fonctionnaire d'une organisation pour la promotion de l'alsacien
- ☐ Oui, comme enseignant fonctionnaire
- ☐ Oui, en tant que membre d'une association locale ou nationale de défense de l'alsacien
- ☑ Oui, dans un autre cadre (précisez ci-dessous)
- ☐ Non

Précisez le cadre de votre investissement

Évaluez votre niveau en alsacien *

	bon	moyen	faible
Compréhension orale *	◉	○	○
Production orale *	○	◉	○
Compréhension écrite *	◉	○	○
Production écrite *	○	◉	○

Vous ameriez améliorer votre... *

	Oui	Non, ce n'est pas la peine
Compréhension orale *	◉	○
Production orale *	◉	○
Compréhension écrite *	○	◉
Production écrite *	◉	○

Utilisez vous l'alsacien sur Internet ? *

- ◉ Oui (même rarement)
- ○ Non, jamais

Si oui, que faites-vous en alsacien ? *

- ☑ Je lis des contenus (articles, publications, commentaires) en alsacien
- ☑ J'écris des contenus (articles, publications, commentaires) en alsacien

Écrivez-vous l'alsacien ? *

- ◉ Oui (même rarement)
- ○ Non, car l'alsacien est une langue orale que je ne souhaite pas écrire
- ○ Non, car je ne saurais pas comment l'écrire
- ○ Non, car je n'en ai pas l'occasion
- ○ Non, pour une autre raison

Précisez

Lorsque vous écrivez en alsacien (en ligne ou non), qu'écrivez-vous ? *

- ☐ Des articles informatifs (informations, articles de blog, histoire de l'Alsace, histoire familiale etc.)
- ☐ Des contenus littéraires (pièce de théâtre, nouvelle, roman, poésie etc.)
- ☐ Des commentaires à des publications en alsacien
- ☑ Des blagues
- ☐ Des lettres ou des e-mails
- ☐ Des recettes de cuisine
- ☑ Des avis politiques
- ☐ Vous conversez par écrit en alsacien sur les réseaux sociaux (t'chat, Twitter, Facebook messenger etc.)
- ☑ Autres (précisez ci-dessous)

Donnez des exemples de contenus que vous écrivez en alsacien (en ligne ou non)

Lorsque vous écrivez, utilisez vous la graphie issue de la méthode ORTHAL (http://www.orthal.fr/) ? *

- ○ Oui, toujours.
- ○ Oui, ça m'arrive.
- ◉ Non, j'aimerais l'utiliser mais je ne la maîtrise pas.
- ○ Non, je connais cette méthode mais refuse de l'utiliser.
- ○ Non, je n'avais jamais entendu parler de cette graphie.

Que pensez-vous des affirmations suivantes ? *

	Tout à fait d'accord	Plutôt d'accord	Neutre	Plutôt pas d'accord	Pas du tout d'accord	Ne se prononce pas
J'aimerais pouvoir utiliser l'alsacien dans tous les aspects de ma vie *	○	◉	○	○	○	○
Utiliser l'alsacien sur Internet permet de montrer que c'est une langue dynamique *	◉	○	○	○	○	○
Une meilleure présence de l'alsacien sur Internet le rendrait plus attractif aux nouvelles générations *	○	○	◉	○	○	○
Cela vaut la peine d'utiliser l'alsacien sur Internet *	○	○	○	◉	○	○
C'est plus facile d'utiliser le français que l'alsacien en ligne *	○	◉	○	○	○	○
Il est impossible d'utiliser l'alsacien sur Internet *	○	○	○	◉	○	○

À votre connaissance, les ressources suivantes sont-elles disponible en ligne ? *

	Je ne sais pas	Non, mais j'aimerais que cela existe	Non	Oui
dictionnaire monolingue en ligne *	○	◉	○	○
dictionnaire bilingue en ligne *	○	○	◉	○
dictionnaire de prononciation en ligne *	○	○	○	◉
correcteur orthographique *	○	○	◉	○
clavier de saisie adapté à l'alsacien *	○	○	◉	○
correcteur orthographique *	○	◉	○	○
traducteur automatique *	◉	○	○	○

Participer à la production collaborative de ressources en ligne pour l'alsacien vous paraît *

- ○ Une bonne idée, je le fais déjà !
- ○ Une bonne idée, mais je ne sais pas comment faire.
- ○ Une bonne idée, mais je n'ai pas le temps.
- ○ Trop compliqué pour moi, je ne suis pas à l'aise avec l'informatique
- ○ Trop compliqué pour moi, mon niveau d'alsacien n'est pas assez bon
- ◉ Une mauvaise idée (expliquez pourquoi ci-dessous)

Pourquoi ?

Bisame et Recettes de Grammaire

Ces deux sites ont été développés dans le cadre d'un travail de thèse à Sorbonne Université pour permettre aux locuteurs de participer à la production collaborative de ressources en ligne pour l'alsacien

Connaissez-vous l'existence de la plate-forme Bisame (http://bisame.paris-sorbonne.fr) ? *

- ◉ Oui, mais je ne me suis pas inscrit
- ○ Oui, je me suis inscrit mais je n'ai pas participé
- ○ Oui, j'y ai participé
- ○ Non, je n'en ai jamais entendu parler

Vous ne vous êtes pas inscrit parce que... *

- [ ] vous n'aviez pas envie de participer
- [ ] vous n'aviez pas confiance dans le site Internet
- [ ] vous n'avez pas compris de quoi il s'agissait
- [ ] vous n'avez pas réussi
- [ ] vous avez pensé vous inscrire plus tard mais vous avez oublié
- [x] autre (précisez)

Précisez pourquoi *

Connaissez-vous l'existence de la plate-forme Recettes de grammaire (http://bisame.paris-sorbonne.fr/recettes) ? *

- (•) Oui, mais je ne me suis pas inscrit
- ( ) Oui, je me suis inscrit mais je n'ai pas participé
- ( ) Oui, j'y ai participé
- ( ) Non, je n'en ai jamais entendu parler

Recettes de Grammaire est un site collaboratif de recettes en alsacien qui permet également de produire des ressources linguistique pour la langue ! Voir l'article paru dans les DNA : https://www.dna.fr/actualite/2018/06/11/audio-comment-automatiser-l-alsacien-a-travers-des-recettes-de-cuisines (ou en cherchant "DNA Bisame" dans Google)

Vous ne vous êtes pas inscrit parce que... *

- [ ] vous n'aviez pas envie de participer
- [ ] vous n'aviez pas confiance dans le site Internet
- [ ] vous n'avez pas compris de quoi il s'agissait
- [ ] vous n'avez pas réussi
- [ ] vous avez pensé vous inscrire plus tard mais vous avez oublié
- [x] autre (précisez)

Précisez pourquoi

Vous aimeriez que votre participation à la création de ressources en ligne vous permette... *

- [ ] de rentrer en contact avec d'autres personnes parlant alsacien pour pouvoir discuter
- [ ] de partager des contenus, des conseils, des astuces, des avis
- [ ] de vous amuser en ligne
- [x] d'améliorer votre alsacien
- [x] d'apprendre des choses (de manière générale)
- [ ] de gagner des bons de réduction (librairies ou événements culturels par exemple)
- [ ] de valoriser votre connaissance de l'alsacien en participant à un projet de recherche
- [ ] autre (préciser)

Un commentaire, une suggestion, un témoignage à partager ?

Vous souhaitez être tenu(e) au courant des résultats de ce sondage ? Inscrivez votre e-mail ci-dessous.

Vos réponses vont nous permettre d'améliorer nos sites de création collaborative de ressources en alsacien. Si vous souhaitez recevoir des nouvelles des projets Bisame et Recettes de grammaire, inscrivez votre e-mail ci-dessous. (maximum 2 mails par an)

G.2 Le créole mauricien et sa présence en ligne

LE CRÉOLE MAURICIEN ET VOUS

Êtes-vous... *
- ○ Un homme
- ◉ Une femme
- ○ Ne souhaite pas répondre

Quel âge avez-vous ? *
- ○ Moins de 20 ans
- ○ Entre 20 et 29 ans
- ○ Entre 30 et 39 ans
- ○ Entre 40 et 49 ans
- ○ Entre 50 et 59 ans
- ○ Entre 60 et 69 ans
- ○ Entre 70 et 79 ans
- ◉ Plus de 80 ans
- ○ Ne souhaite pas répondre

Comment avez-vous eu connaissance de ce sondage ?
- ○ bouche à oreille
- ○ réseaux sociaux
- ◉ autre

Précisez *

La ou les langue(s) que vous considérez comme maternelle(s) est/sont...
- ☐ l'anglais
- ☑ le français
- ☐ le créole mauricien
- ☑ autre

Précisez votre autre langue maternelle *

Avez-vous vécu à l'île Maurice ?
- ◉ Oui, j'y habite actuellement
- ○ Oui, mais je n'y habite plus
- ○ Non

Et dans quelle région de l'île ?
- ☑ Nord
- ☐ Sud
- ☐ Centre
- ☐ Est
- ☐ Ouest

Avez-vous vécu une partie de votre vie dans un autre pays que l'île Maurice ?
- ◉ Oui
- ○ Non

Si oui, vous avez vécu dans un pays...

- [ ] anglophone
- [ ] francophone
- [x] autre

Précisez *

Avez-vous un investissement professionnel et/ou associatif lié au créole mauricien ? ? *

- [ ] Oui, en tant que fonctionnaire d'une organisation pour la promotion du créole mauricien
- [x] Oui, comme enseignant fonctionnaire
- [x] Oui, dans un autre cadre (précisez ci-dessous)
- [ ] Non

Précisez le cadre de votre investissement *

Pensez-vous que le créole mauricien possède des variantes ? *

- (•) Oui, le créole parlé varie selon les régions
- ( ) Non, le créole parlé est le même à travers l'île

Évaluez votre niveau en créole mauricien *

	bon	moyen	faible
Compréhension orale *	(•)	( )	( )
Production orale *	( )	(•)	( )
Compréhension écrite *	( )	(•)	( )
Production écrite *	( )	( )	(•)

Vous ameriez améliorer votre... *

	Oui	Non, ce n'est pas la peine
Compréhension orale *	( )	(•)
Production orale *	(•)	( )
Compréhension écrite *	(•)	( )
Production écrite *	( )	( )

Laquelle de ces phrases vous semble plus facile à lire

- ( ) monn al aste dipin
- ( ) monne alle acheté dipain
- (•) je ne suis à l'aise avec aucune de ces orthographes

Proposez une orthographe alternative : comment auriez-vous écrit "Je suis allé(e) acheter du pain" en créole mauricien ?

Utilisez vous le créole mauricien sur Internet ? *

- (•) Oui (même rarement)
- ( ) Non, jamais

Si oui, que faites-vous en créole mauricien ? *

- [x] Je lis des contenus (articles, publications, commentaires) en créole
- [x] J'écris des contenus (articles, publications, commentaires) en créole

Écrivez-vous le créole mauricien (en ligne ou non) ? *

- (•) Oui (même rarement)
- ( ) Non, car le créole est une langue orale que je ne souhaite pas écrire
- ( ) Non, car je ne saurais pas comment l'écrire
- ( ) Non, car je n'en ai pas l'occasion
- ( ) Non, pour une autre raison

Lorsque vous écrivez en créole mauricien (en ligne ou non), qu'écrivez-vous ? *

- [ ] Des articles informatifs (informations, articles de blog, histoire de l'île Maurice, histoire familiale etc.)
- [ ] Des contenus littéraires (pièce de théâtre, nouvelle, roman, poésie etc.)
- [ ] Des commentaires à des publications en créole
- [x] Des blagues
- [ ] Des lettres ou des e-mails
- [ ] Des recettes de cuisine
- [x] Des avis politiques
- [ ] Vous conversez par écrit en créole sur les réseaux sociaux (t'chat, Twitter, Facebook messenger etc.)
- [x] Autres (précisez ci-dessous)

Donnez des exemples de contenus que vous écrivez en créole mauricien (en ligne ou non) *

Quand vous n'êtes pas sûr de l'orthographe d'un mot, vous l'écrivez...

- ( ) comme je l'entends
- ( ) avec la graphie française
- (•) ça dépend

Sauriez-vous expliquer de quoi cela dépend ?

- (•) oui
- ( ) non

Ça dépend de... *

Lorsque vous écrivez, utilisez vous la graphie préscrite par l'Akademi Kreol Morisien dans "Lortograf Kreol Morisien" ? *

- ( ) Oui, toujours.
- (•) Oui, ça m'arrive.
- ( ) Non, j'aimerais l'utiliser mais je ne la maîtrise pas.
- ( ) Non, je connais cette graphie mais refuse de l'utiliser.
- ( ) Non, je n'avais jamais entendu parler de cette graphie.

Orthographe du créole mauricien, éditée par l'Akademi Kreol Morisien (2011) : cliquez ici

Que pensez-vous des affirmations suivantes ? *

	Tout à fait d'accord	Plutôt d'accord	Neutre	Plutôt pas d'accord	Pas du tout d'accord	Ne se prononce pas
J'aimerais pouvoir utiliser le créole dans tous les aspects de ma vie *	( )	(•)	( )	( )	( )	( )
Une meilleure présence du créole sur Internet favoriserait la diffusion d'une graphie uniformisée *	( )	( )	( )	(•)	( )	( )
Cela vaut la peine d'utiliser le créole mauricien sur Internet *	( )	(•)	( )	( )	( )	( )
C'est plus facile d'utiliser le français et l'anglais que le créole en ligne *	(•)	( )	( )	( )	( )	( )
Il est impossible d'utiliser le créole sur Internet *	( )	(•)	( )	( )	( )	( )

À votre connaissance, les ressources suivantes sont-elles disponible en ligne pour le créole mauricien ? *

	Oui	Non	Non, mais j'aimerais que cela existe
dictionnaire monolingue en ligne *	○	◉	○
dictionnaire bilingue en ligne *	◉	○	○
dictionnaire de prononciation en ligne *	○	◉	○
correcteur orthographique *	○	○	◉
traducteur automatique *	○	○	◉

PARTICIPER AU DÉVELOPPEMENT DU CRÉOLE MAURICIEN GRÂCE A VOS CONNAISSANCES

Connaissez-vous le principe de la production participative (aussi appelé "crowdsourcing") ? *

- ○ Oui
- ○ Non

Participer à la production collaborative de ressources en ligne pour le créole mauricien vous paraît *

- ○ Une bonne idée, je le fais déjà !
- ○ Une bonne idée, mais je ne sais pas comment faire.
- ○ Une bonne idée, mais je n'ai pas le temps.
- ○ Trop compliqué pour moi, je ne suis pas à l'aise avec l'informatique
- ○ Trop compliqué pour moi, mon niveau de créole n'est pas assez bon
- ○ Une mauvaise idée (expliquez pourquoi ci-dessous)

Vous aimeriez que votre participation à la création de ressources en ligne vous permette... *

- ☐ de rentrer en contact avec d'autres personnes parlant le créole mauricien pour pouvoir discuter
- ☐ de partager des contenus, des conseils, des astuces, des avis
- ☐ de vous amuser en ligne
- ☐ d'améliorer votre créole
- ☐ d'apprendre des choses (de manière générale)
- ☐ de gagner des bons de réduction (librairies ou événements culturels par exemple)
- ☐ de valoriser votre connaissance du créole en participant à un projet de recherche
- ☐ autre (préciser)

Un commentaire, une suggestion, un témoignage à partager ?

Vous souhaitez être tenu(e) au courant des résultats de ce sondage ? Inscrivez votre e-mail ci-dessous.

Vous souhaitez être averti(e) lorsque nous aurons mis en place un site de production participative pour le créole mauricien ? Inscrivez votre mail ci-dessous"

Myriadisation de ressources linguistiques pour le traitement automatique de langues non standardisées

Résumé

Les sciences participatives, et en particulier la myriadisation (*crowdsourcing*) bénévole, représentent un moyen peu exploité de créer des ressources langagières pour certaines langues encore peu dotées, et ce malgré la présence de locuteurs sur le Web. Nous présentons dans ce travail les expériences que nous avons menées pour permettre la myriadisation de ressources langagières dans le cadre du développement d'un outil d'annotation automatique en parties du discours. Nous avons appliqué cette méthodologie à trois langues non standardisées, en l'occurrence l'alsacien, le créole guadeloupéen et le créole mauricien. Pour des raisons historiques différentes, de multiples pratiques (ortho)graphiques co-existent en effet pour ces trois langues. Les difficultés posées par l'existence de cette variation nous ont menée à proposer diverses tâches de myriadisation permettant la collecte de corpus bruts, d'annotations en parties du discours, et de variantes graphiques.
L'analyse intrinsèque et extrinsèque de ces ressources, utilisées pour le développement d'outils d'annotation automatique, montrent l'intérêt d'utiliser la myriadisation dans un cadre linguistique non standardisé : les locuteurs ne sont pas ici considérés comme un ensemble uniforme de contributeurs dont les efforts cumulés permettent d'achever une tâche particulière, mais comme un ensemble de détenteurs de connaissances complémentaires. Les ressources qu'ils produisent collectivement permettent de développer des outils plus robustes à la variation rencontrée.
Les plateformes développées, les ressources langagières, ainsi que les modèles de *taggers* entraînés sont librement disponibles.

Mots-clés : Myriadisation ; Traitement automatique des langues ; Langues peu dotées ; Langue non standardisées ; Corpus annoté ; morphosyntaxe ; annotation manuelle

Crowdsourcing linguistic resources for natural non-standardised languages processing

Summary

Citizen science, in particular voluntary crowdsourcing, represents a little experimented solution to produce language resources for some languages which are still little resourced despite the presence of sufficient speakers online.
We present in this work the experiments we have led to enable the crowdsourcing of linguistic resources for the development of automatic part-of-speech annotation tools. We have applied the methodology to three non-standardised languages, namely Alsatian, Guadeloupean Creole and Mauritian Creole. For different historical reasons, multiple (ortho)-graphic practices coexist for these three languages. The difficulties encountered by the presence of this variation phenomenon led us to propose various crowdsourcing tasks that allow the collection of raw corpora, part-of-speech annotations, and graphic variants.
The intrinsic and extrinsic analysis of these resources, used for the development of automatic annotation tools, show the interest of using crowdsourcing in a non-standardized linguistic framework: the participants are not seen in this context a uniform set of contributors whose cumulative efforts allow the completion of a particular task, but rather as a set of holders of complementary knowledge. The resources they collectively produce make possible the development of tools that embrace the variation.
The platforms developed, the language resources, as well as the models of trained taggers are freely available.

Keywords : Crowdsourcing ; Natural language processing ; Less-resourced languages ; Non-standardized languages ; Annotated corpora ; Part-of-speech ; Manual annotation

UNIVERSITÉ SORBONNE UNIVERSITÉ

ÉCOLE DOCTORALE :
École Doctorale V – Concepts et langage
Maison de la Recherche, 28 rue Serpente, 75006 Paris, FRANCE

DISCIPLINE : Mathématiques, informatique et applications aux sciences de l'Homme