

À l'écoute des locuteurs : production participative de ressources langagières pour des langues non standardisées

Alice Millour, Karën Fort

▶ **To cite this version:**

Alice Millour, Karën Fort. À l'écoute des locuteurs : production participative de ressources langagières pour des langues non standardisées. Revue TAL : traitement automatique des langues, 2018. hal-01995758

HAL Id: hal-01995758
https://hal.science/hal-01995758

Submitted on 13 Feb 2019

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

À l'écoute des locuteurs : production participative de ressources langagières pour des langues non standardisées

Alice Millour* — **Karën Fort****

Sorbonne Université, STIH - EA 4509, 28 rue Serpente, 75006 Paris, France
** alice.millour@etu.sorbonne-universite.fr, ** karen.fort@sorbonne-universite.fr*

RÉSUMÉ. *Les sciences participatives, et en particulier la production participative (*crowdsourcing*) bénévole, sont un moyen encore peu exploité de créer des ressources langagières pour les langues peu dotées dont suffisamment de locuteurs sont présents sur le Web. Nous présentons ici nos expériences concernant l'annotation en parties du discours pour des langues non standardisées, en l'occurrence l'alsacien et le créole guadeloupéen. Nous détaillons la méthodologie utilisée, montrons qu'elle est adaptable à plusieurs langues, puis nous présentons les résultats obtenus. L'analyse des limites de la plateforme d'origine nous a conduites à en développer une nouvelle, qui, outre l'annotation en parties du discours, permet la création de corpus bruts et d'un lexique de variantes alignées. Les plateformes créées, les ressources langagières, et les modèles de* taggers *entraînés sont librement disponibles.*

ABSTRACT. *Citizen science, in particular voluntary crowdsourcing, is still little experimented solution to produce language resources for less-resourced languages with enough connected speakers. We present here experiments we led on part-of-speech annotation for non standardized languages, namely Alsatian and Guadeloupean Creole. We detail the methodology we used and show that it is adaptable to other languages, then we present the results we obtained. An analysis of the limits of this platform led us to develop a new one, that allows the creation of raw corpora and part-of-speech annotations, and the construction of a multivariant lexicon. The created platforms, language resources and tagging models are all freely available.*

MOTS-CLÉS : *langues non standardisées, production participative, annotation en parties du discours.*

KEYWORDS: *non-standardized languages, crowdsourcing, part-of-speech annotation.*

1. Pourquoi et comment créer des ressources pour des langues non standardisées

1.1. *Un enjeu culturel majeur*

Alors que les communications numériques connaissent un essor sans précédent et que l'accès aux technologies de communication moderne se démocratise[1], le monde numérique reste très peu représentatif des communautés linguistiques y ayant accès (Prado, 2012). Or, la diversité linguistique fait partie du patrimoine culturel à préserver, et la recherche en traitement automatique des langues (TAL), en permettant la saisie et la diffusion de contenus numériques, la présence en ligne de ressources linguistiques, ou encore le développement d'outils pédagogiques pour un nombre croissant de langues, peut participer à enrayer l'érosion à l'œuvre.

Les travaux portant sur de nouvelles langues présentent l'intérêt de confronter les chercheurs à des problématiques linguistiques nouvelles. Pour autant, la recherche en TAL ne concerne encore qu'une extrême minorité de langues : d'après Benjamin (2018), une majorité d'êtres humains a aujourd'hui pour langue maternelle l'une des 7 000 langues n'étant pas ou très peu considérées dans nos recherches. En cause, notamment, les politiques de financement et les opportunités professionnelles moindres découlant des recherches sur les langues peu dotées (Branco, 2018).

En outre, l'essor des communications virtuelles pose la question de l'intégration des langues non standardisées aux technologies du langage. Alors que l'UNESCO (Diki-Kidiri, 2007) 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, force est de constater que certaines de ces langues sont d'ores et déjà en usage sur Internet. L'absence de norme orthographique n'empêche pas l'utilisation de ces langues à l'écrit, en témoigne par exemple le cas de « l'*entre-soi* des groupes Facebook » (Rivron, 2012), qui favorise par exemple le dépassement de la gêne à écrire l'éton (langue de la région du Centre au Cameroun, autour de 250 000 locuteurs).

1.2. *Les sciences participatives, une solution encore peu exploitée*

La construction de ressources annotées de qualité par des linguistes est notoirement coûteuse[2]. De nombreuses langues, ne présentant pas un intérêt économique

1. Voir, par exemple, l'évolution de la couverture d'Internet donnée par Internet World Stats (`https://www.internetworldstats.com/stats.htm`), évaluée à 54,4 % de la population mondiale au 31 décembre 2017.

2. Voir par exemple (Böhmová *et al.*, 2001), l'une des rares publications donnant un coût approximatif pour une ressource langagière, le *Prague Dependency Treebank*, de l'ordre de 600 000 dollars.

immédiat, ou dont le nombre de locuteurs est faible, en sont par conséquent privées. Par ailleurs, ces locuteurs représentent un recours potentiel insuffisamment exploité pour la construction de ressources langagières. Notre hypothèse est qu'il est possible de pallier le manque de ressources langagières brutes et annotées en mettant à contribution les locuteurs *via* une plateforme de production participative adaptée intégrant des outils de TAL (y compris ceux créés à l'aide de ces ressources).

En outre, il apparaît que l'intervention des locuteurs dans la construction de ressources pour une langue non standardisée soit une condition nécessaire à la production de ressources de qualité représentatives des variétés de la langue.

1.3. *Une démarche expérimentale itérative*

L'hypothèse formulée pose un certain nombre de questions scientifiques : peut-on atteindre une qualité d'annotation suffisante de la part de locuteurs non familiers de la linguistique ? Quelle stratégie adopter pour assurer la qualité des ressources produites pour une tâche difficile pour les participants, susceptibles de commettre des erreurs ? Comment intégrer le TAL à la constitution de ressources de façon transparente pour le participant, tout en s'assurant qu'il ait conscience de l'impact de sa participation ? Enfin, comment optimiser, faciliter et rendre agréable une telle participation ?

Pour répondre à ces questions et évaluer notre hypothèse, nous avons développé deux plateformes de production participative permettant de recueillir des ressources de différentes natures. Ainsi, après avoir présenté la nécessité de construire des ressources annotées qui soient indépendantes de tout outil d'annotation, libres de droits, numérisées et accessibles, pour assurer la pérennité des avancées technologiques pour les langues peu dotées, nous présentons ici les deux expériences menées.

Notre première approche, décrite en partie 3 présente une méthodologie d'annotation participative en parties du discours. Si cette tâche est considérée comme résolue pour l'anglais et les langues dites « supercentrales » (Calvet, 2002)[3], de nombreuses langues, notamment l'alsacien et le créole guadeloupéen, ne bénéficient pas de corpus annotés de qualité ni d'outils performants. Nous détaillons dans la section 3.4 le processus d'instanciation de cette méthodologie pour ces deux langues de France non standardisées, ainsi que les résultats obtenus en termes de participation, de qualité des annotations et de performances des outils entraînés avec celles-ci. Cette première initiative de production participative pour deux langues peu dotées nous a amenées à développer une seconde méthodologie. Celle-ci place cette fois le locuteur au cœur de la production de corpus bruts représentatifs de la réalité des variétés existantes pour chaque langue avant même leur annotation, notamment grâce à la construction col-

3. Ce qui n'est vrai que dans une certaine mesure, comme le montre, entre autres, la quantité de travaux concernant l'adaptation au domaine, aux communications virtuelles (*computer-mediated communication*), ou aux contenus générés par les utilisateurs (*user-generated contents*), qui sont autant de catégories de productions langagières pouvant être considérées comme moins dotées au regard de certaines tâches.

laborative d'un lexique de variantes alignées. L'instanciation de cette méthodologie pour l'alsacien ainsi que les résultats préliminaires obtenus sont décrits en section 4.

2. État de l'art

Sont considérées comme peu dotées les langues qui, comparativement à d'autres, disposent de moins de ressources et outils favorisant leur intégration dans le monde numérique (voir notamment (Berment, 2004)). Cette appellation recouvre des réalités très variées : des langues ayant ou non le statut de langue officielle (par exemple, l'islandais et l'igbo), parlées par un nombre réduit ou important de locuteurs (par exemple l'inuktitut et le lao), présentant ou non une parenté avec une langue mieux dotée (par exemple l'occitan languedocien au regard du catalan, et l'arménien, isolé), etc. En résulte une grande diversité de caractéristiques linguistiques et de ressources, au sens large, disponibles pour chacune de ces langues.

Dans cette partie, nous présentons, dans un premier temps, les stratégies mises en place pour tirer parti de ces différents paramètres au regard de la tâche d'annotation en parties du discours. Puis, nous présentons les travaux existants quant à la production participative comme moyen de produire à bas coût des ressources de qualité.

2.1. *Annotation en parties du discours des langues peu dotées*

2.1.1. *Pallier le manque de ressources, approches existantes et limitations*

Les travaux existants quant à l'annotation en parties du discours présentent un éventail de stratégies visant à pallier le déficit de ressources annotées nécessaires au développement d'approches statistiques supervisées classiques. Elles diffèrent par la nature des ressources qu'elles requièrent. On compte notamment (i) les approches non supervisées tirant profit de la disponibilité de corpus parallèles permettant la projection d'annotations (Agić *et al.*, 2016), ou de la parenté de la langue considérée avec une langue mieux dotée (Hana *et al.*, 2004 ; Scherrer et Sagot, 2013 ; Bernhard *et al.*, 2018), (ii) les approches semi-supervisées telles que celle décrite par Garrette *et al.* (2013), intégrant par exemple des transducteurs à états finis pour analyser les mots inconnus, (iii) les approches faiblement supervisées, telles que l'utilisation du `Wiktionnaire` comme ressource complémentant un corpus annoté de taille réduite (Li *et al.*, 2012). Une solution pour limiter les coûts de développement est d'intégrer un lexique externe à l'entraînement d'un outil d'annotation supervisé de manière à augmenter la qualité d'annotation tout en limitant le coût de construction de la ressource (voir en particulier la figure 8.1 de (Sagot, 2018)).

Or, pour un grand nombre de langues, aucune de ces ressources n'est disponible en quantité suffisante. Par ailleurs, et quelle que soit la méthode employée, l'existence d'un corpus annoté est nécessaire, *a minima* comme référence pour évaluer les outils développés. Notons également que la libre disponibilité de ressources pérennes garantit la réutilisabilité de celles-ci, indépendamment des avancées technologiques.

2.1.2. *Le cas particulier des langues non standardisées*

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, notamment celui des langues anciennes, par exemple le moyen allemand (Barteld, 2017), des contenus générés par les utilisateurs, comme `Wikipédia` (Krumm *et al.*, 2008), ou des communications médiées par ordinateur (Melero *et al.*, 2012). Des langues bien dotées, à l'instar du chinois mandarin (de Chine continentale, de Hong Kong et de Taïwan) (Tseng *et al.*, 2005) ou du portugais (brésilien et du Portugal) (Garcia *et al.*, 2014), sont également sujettes à cette variabilité. Or, à ce jour, les outils développés et évalués pour une langue donnée sont en réalité conçus de manière peu robuste à toutes formes de variations (Plank, 2016).

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. La variabilité peut principalement être prise en compte de deux manières :

– soit en entraînant un outil pour chaque variété de langue considérée (*language adaptation*), ce qui implique, outre la nécessité de pouvoir identifier les variétés, une démultiplication du nombre de corpus d'entraînement nécessaires ;

– soit en normalisant les corpus (voir par exemple (Ljubešic *et al.*, 2016 ; Samardzic *et al.*, 2015), ou (Cox, 2010), pour une discussion sur la *rentabilité* de la normalisation). Cela suppose de définir une norme, et de connaître les variétés ainsi que les mécanismes de normalisation pour chacune d'elles. Est également envisagée l'utilisation de techniques de translittération (Pingali *et al.*, 2017), ou de dictionnaires de prononciation pour l'entraînement de modèles de transcription phonétique réduisant la variabilité scripturale (Steiblé et Bernhard, 2018).

Quelle que soit la méthode employée, celle-ci requiert soit une description de la variation existante, soit un large corpus permettant d'en inférer les motifs.

2.2. *Production participative de ressources langagières*

2.2.1. *Des productions participatives*

La production participative (*crowdsourcing*) s'est imposée depuis une dizaine d'années comme l'une des solutions aux freins que constitue le manque de moyens et de linguistes disponibles pour la construction de ressources langagières. Elle consiste à lancer un appel ouvert à participation (aujourd'hui principalement *via* le Web) pour faire réaliser une tâche, par des bénévoles (comme sur `Wikipédia`) ou en échange d'une (micro) rémunération (*microworking crowdsourcing*, comme sur `Amazon Mechanical Turk`[4]). Il existe bien entendu un continuum entre ces

4. Voir : `https://www.mturk.com/`.

deux extrêmes, avec des applications bénévoles offrant des « récompenses » variées, depuis le simple divertissement, à l'instar des jeux ayant un but comme JeuxDeMots (Lafourcade et Joubert, 2008) ou ZombiLingo (Guillaume *et al.*, 2016) jusqu'aux bons d'achat, comme sur Phrase Detectives (Chamberlain *et al.*, 2009).

Si le travail parcellisé à la Amazon Mechanical Turk permet 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 trouver plus facilement des experts pour certaines langues peu dotées (Callison-Burch et Dredze, 2010) et pose des problèmes éthiques et de qualité produite (Fort *et al.*, 2011). En effet, il est actuellement impossible, sur ces plateformes, de former les travailleurs à la tâche (il n'est possible que de les évaluer). Or, pour une tâche comme l'annotation en parties du discours, une formation est nécessaire : en témoignent les résultats obtenus par Hovy *et al.* (2014) pour l'annotation *via* CrowdFlower[5] de *tweets* en anglais (84 % d'exactitude), et de Jamatia et Das (2014) et Zaghouani et Dukes (2014) déplorant tous deux 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[6], 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. Cependant, attirer et retenir les participants sur ces plateformes constituent un exercice complexe (Tuite, 2014), encore qualifiable d'« alchimie ».

2.2.2. *Productions participatives pour les langues peu dotées*

Les travaux concernant la production participative pour les langues peu dotées s'articulent selon différents axes. L'un concerne la production de données orales, dans un but de documentation des langues en danger, à l'instar des travaux de Bettinson et Bird (2017) et de Blachon *et al.* (2016) développant des outils de collecte de la parole *à l'attention des chercheurs*, ou de collecte d'informations sur la variation dialectale, dont par exemple (Leemann *et al.*, 2015). D'autres travaux utilisent la production participative pour recueillir des données géolinguistiques (voir par exemple (Avanzi et Stark, 2017) pour la variation du français, ou (Boula de Mareüil *et al.*, 2018) pour la construction d'un atlas sonore des langues régionales de France). À notre connaissance, sloWCrowd est la seule plateforme existante visant la production collaborative de ressources pour le traitement automatique des langues peu dotées. Si elle permet la validation de ressources annotées (par exemple, le sloWNet (Fišer *et al.*, 2014)) elle

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

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

est inadaptée à la résolution de tâches relativement complexes telles que l'annotation en parties du discours (Klubička et Ljubešić, 2014).

3. Expérimenter la production participative pour l'annotation en parties du discours

Notre première expérience de production participative a concerné l'annotation en parties du discours : à travers une plateforme dédiée, le participant annote des séquences de quatre phrases (figure 1) dont une provient d'un corpus de référence annoté par des linguistes et sert à l'évaluation du participant[7].



Figure 1. *Interface d'annotation en parties du discours pour l'alsacien*

3.1. *Une hypothèse de départ forte*

Les locuteurs des langues régionales de France sont très attachés à leur langue, à sa survie, voire à son développement. Notre hypothèse de départ a supposé qu'il n'était pas nécessaire de développer un jeu autour de la tâche pour motiver les locuteurs à y participer. Un tel développement étant en effet coûteux, nous avons privilégié la création d'une plateforme librement disponible légère, facile à maintenir et à adapter à d'autres langues. Nous avons par conséquent limité la ludification de la plateforme à un simple système de points permettant de classer les participants.

3.2. *Un processus cyclique intégrant un contrôle de la qualité*

3.2.1. *Préannotation et intégration continue des annotations*

Nous nous sommes inspirées de la méthodologie utilisée avec succès pour la syntaxe en dépendances du français dans `ZombiLingo` (Guillaume *et al.*, 2016), notam-

7. Le nombre de phrases a été choisi de manière à limiter la durée d'annotation d'une séquence (moins de dix minutes en moyenne), tout en assurant la collecte de suffisamment de données d'évaluation.

ment en intégrant i) des outils de préannotation, ii) une formation obligatoire pour les participants, et iii) une méthodologie d'évaluation continue des ressources produites.

La préannotation des corpus par deux outils intégrés à la plateforme est facultative mais permet de réduire la complexité de la tâche en ne proposant au participant que les étiquettes les plus probables. Si aucun outil, même imparfait, n'est disponible, une possibilité consiste à en entraîner un avec un corpus de taille très réduite dans un premier temps. Afin d'améliorer la qualité de la préannotation (ici, en parties du discours) au fur et à mesure de l'expérimentation, nous avons mis en place un processus vertueux (illustré par la figure 2), qui consiste à réentraîner régulièrement l'outil supervisé avec les annotations produites par les participants. Un deuxième outil de préannotation minimaliste peut être obtenu en tirant parti des caractéristiques linguistiques de la langue considérée (nous en donnons des exemples dans la section 3.4.4).

La double préannotation est utilisée de la manière suivante : lorsque les deux outils proposent la même étiquette, celle-ci est proposée en priorité au participant. Lorsqu'ils sont en désaccord, ce sont les deux étiquettes qui sont suggérées. Ce mécanisme permet d'accélérer la tâche de validation tout en laissant au participant la possibilité de choisir une autre étiquette dans la liste complète. L'outil supervisé étant réentraîné régulièrement, le taux d'accord entre les deux outils est amené à évoluer selon le processus vertueux lié aux performances croissantes de l'outil, facilitant à chaque itération la tâche au participant.

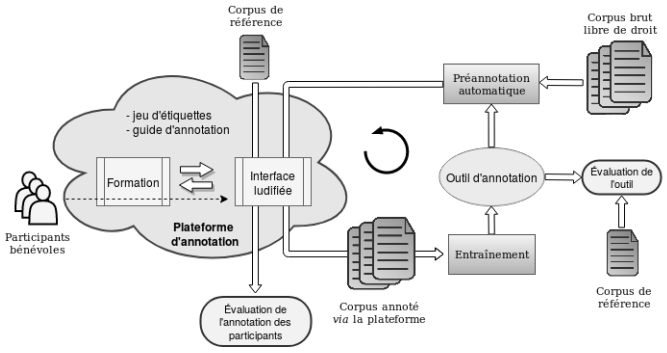


Figure 2. *Processus cyclique d'intégration continue des annotations produites*

3.2.2. *Une formation obligatoire*

Nous avons mis en place une phase de formation obligatoire pour tous les participants consistant à annoter intégralement quatre phrases issues d'un corpus de référence. Elle est conçue pour être la plus proche possible de la phase de production d'annotations, à ceci près que le participant ne peut valider une phrase que lorsque les annotations qu'il propose sont correctes. En cas d'erreur, le *token* mal annoté est mis en évidence mais l'étiquette attendue n'est pas divulguée. Cette première phase permet de confronter le participant aux difficultés de la tâche tout en le familiarisant aux

catégories existantes. Qu'il s'agisse de la phase de formation ou de production d'annotations, le participant à toujours accès à un guide d'annotation simplifié, organisé sous forme de listes d'exemples servant d'aide-mémoire pour chacune des catégories.

3.2.3. *Une évaluation continue des participants et des annotations*

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 *token* T avec la catégorie C_i : $ScoreAnn_{T,P,C_i}$ 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 *token* T par plusieurs participants. Nous déterminons ainsi pour chaque étiquette attribuée au *token* T un score de confiance $Score_{T,C_i}$ correspondant à la moyenne des scores des annotations Ann_{T,P_j,C_i} produites par différents participants : $Score_{T,C_i} = \frac{\sum_j ScoreAnn_{T,P_j,C_i}}{\sum_{i,j} ScoreAnn_{T,P_j,C_i}}$.

Nous choisissons enfin l'étiquette unique la plus probable pour chaque *token* : $C_T = \arg\max_i(Score_{T,C_i})$. Le corpus ainsi annoté est utilisé pour entraîner des *taggers*, utilisés à leur tour comme outils de préannotation dès lors que leurs performances dépassent l'outil précédent.

3.3. *Une méthodologie réplicable et des ressources produites librement disponibles*

Afin de proposer une méthodologie qui soit facilement réplicable[8], nous avons choisi d'utiliser le jeu d'étiquettes universel (Petrov *et al.*, 2012), contenant initialement 17 catégories (tableau 1), et facilement adaptable aux besoins spécifiques de chaque langue. Les modifications apportées à ce jeu d'étiquettes lors de l'instanciation de la plateforme sont présentées en partie 3.4.2.

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

Tableau 1. *Liste des étiquettes utilisées selon le classement de ses créateurs*

Par ailleurs, afin de garantir la disponibilité des ressources produites, la plateforme d'annotation est alimentée exclusivement de corpus libres de droits redistribuables.

8. C'est-à-dire que le processus peut être reproduit, sans nécessairement que les résultats le soient (reproductibilité) (Cohen *et al.*, 2018).

Pour des langues pouvant être peu dotées en termes même de ressources brutes, ce qui est souvent le cas des langues non standardisées, cela revient à suivre une démarche pragmatique aboutissant à la création de « corpus opportunistes » (McEnery et Hardie, 2011), représentant « [...] ni plus ni moins que les ressources qui ont pu être recueillies pour une tâche donnée. »[9].

Le code de la plateforme développée est quant à lui librement disponible sur `GitHub`[10] sous licence CeCILL v2.1[11]. La méthodologie décrite peut ainsi être adaptée facilement, la plateforme étant prête à être instanciée comme illustré dans la section 3.4.

3.4. *Une instanciation pour deux langues : l'alsacien et le créole guadeloupéen*

Nous avons mis en place deux instances de la plateforme décrite en section 3[12] : `Bisame`[13] (« *bisame* », ou « *bisanme* », « *bisàmme* », « *bisamme* », soit « ensemble » est employé dans l'expression « *Salü bisame !* », soit « Bonjour à tous ! ») pour l'alsacien, et `Krik`[14] (« *Krik* » est un terme intraduisible utilisé dans la tradition créole par les conteurs avant leur prise de parole) pour le créole guadeloupéen[15].

3.4.1. *Deux langues de France aux profils très différents*

L'alsacien est un terme générique pour le continuum de sous-systèmes dialectaux germaniques (Malherbe, 1983) parlé en Alsace et dans une partie de la Moselle. Le bas alémanique, variété principale de l'alsacien, est lui-même divisé en deux sous-ensembles : le bas alémanique du nord (NV) et du sud (SV). On trouve à Strasbourg une variété du bas alémanique du nord légèrement teintée de francique (STRV). En dépit du déclin de la transmission familiale, une étude comptabilise 550 000 locuteurs en 2004 (Barre et Vanderschelden, 2004). Le créole guadeloupéen, à base lexicale française et africaine, compte pour sa part environ 600 000 locuteurs (400 000 en Guadeloupe, 200 000 ailleurs dans le monde (Colot et Ludwig, 2013)). L'atlas des langues en danger établi par l'UNESCO[16] ne donne pas le degré de vitalité de l'alsacien pris

9. "*[...] nothing more nor less than the data that it was possible to gather for a specific task.*" (McEnery et Hardie, 2011).

10. Voir : `https://github.com/alicemillour/Bisame`.

11. Voir : `http://www.cecill.info/`.

12. Ces plateformes ont fait l'objet de publications spécifiques (Millour et Fort, 2018a ; Millour et Fort, 2018b), nous les présentons ici en regard, avec de nouveaux éléments d'analyse, notamment une enquête sur les participants.

13. Voir : `http://bisame.paris-sorbonne.fr`.

14. Voir : `http://krik.paris-sorbonne.fr`.

15. Alsacien et créole guadeloupéen évoluant dans un contexte de diglossie avec le français, c'est la langue que nous avons utilisée pour 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.

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

isolément, mais donne celui du groupe des « langues alémaniques », comprenant également le souabe et le haut valaisan. Ce groupe, au nombre de locuteurs de l'ordre du million, est classé comme vulnérable par l'UNESCO. Le créole guadeloupéen, plus dynamique, est pour sa part absent de l'atlas.

Aucune de ces langues n'a été lissée par l'usage d'une forme normative écrite, bien que des initiatives de graphies unifiées existent, notamment l'orthographe `ORTHAL` (Crévenat-Werner et Zeidler, 2008) pour l'alsacien, et celle du GEREC-F (Groupe d'études et de recherches en espace créolophone et francophone) (Ludwig *et al.*, 1990) modifiée plus tard par Bernabé (2001) et coexistant avec le système introduit par Hazaël-Massieux (2000) pour le créole guadeloupéen. En résulte une variabilité scripturale qui s'additionne à la variabilité dialectale, multipliant les graphies existantes pour un élément de lexique donné. Enfin, les deux langues coexistant avec le français, les éventuels trous lexicaux sont parfois remplis par des termes français.

3.4.2. *Tokénisation et mises à jour des jeux d'étiquettes*

La tokénisation de langues non standardisées représente une gageure, dans la mesure où l'ensemble des pratiques scripturales n'est pas connu en amont de la conception du tokéniseur. Dans le cadre de nos expériences, nous avons utilisé un script `Python`, initialement développé pour l'alsacien (Bernhard *et al.*, 2017) que nous avons adapté au créole guadeloupéen. Dans les deux langues, le script a été mis à jour tout au long de l'expérimentation, de nouvelles formes orthographiques remettant en cause nos règles de tokénisation apparaissant au fur et à mesure que nos corpus augmentaient en taille.

Parallèlement à ces ajustements du tokéniseur, le jeu d'étiquettes (voir la section 3.3) a été complété pour les deux langues, afin de faciliter l'annotation sans affecter la bonne lisibilité des textes proposés. En particulier, nous avons dû effectuer un certain nombre de choix arbitraires assurant la bonne intelligibilité des séquences proposées pour les participants, *a priori* non experts.

Dans le cas de l'alsacien, nous avons ajouté la catégorie ADP+DET pour les contractions n'étant pas automatiquement séparées, par exemple *am*, contraction de *an* et *dem* (« au »). Dans le cas du créole guadeloupéen, la tokénisation, par exemple, de la contraction *k'ay*, regroupant *ka*, particule du présent, et *ay*, (3e personne du singulier du verbe « avoir »), sous forme de deux *tokens k'* et *ay*, rendait la lecture et la compréhension difficiles pour les locuteurs. Ces considérations nous ont amenées à ajouter la catégorie PART+VERB. Pour les mêmes raisons, les *tokens* contenant des pronoms tels que *ba'y* (« pour lui/elle »), *trapé'y* (« l'attraper »), ou *sa'w* (contraction de *sa* (« ce/cette ») et *ou* (« tu »), littéralement « ce que tu ») nous ont amenées à l'ajout des catégories ADP+PRON, VERB+PRON et PRON+PRON.

3.4.3. *Collecte des corpus bruts et construction d'une référence minimale*

Les corpus bruts alimentant la plateforme ont été construits suivant la méthodologie décrite dans la section 3.3. Nous avons recueilli pour ces deux langues l'ensemble

des corpus libres de droits et accessibles à notre connaissance. Ceux-ci proviennent par conséquent de sources hétérogènes, telles que les projets de la Wikisphère (Wikipédia, Wiktionnaire, incubateurs Wikimédia), des textes non soumis au droit d'auteur produits par les organismes locaux de promotion de la langue, notamment l'OLCA[17], ou gracieusement fournis par des participants, ou de bases de données telles COCOON[18] qui contient des transcriptions de conversations en créole guadeloupéen[19]. Les contenus et tailles de ces corpus sont détaillés dans le tableau 2. Les variétés des corpus alsaciens (section 3.4.1) sont données en indice. Dans le cas de COCOON, la taille est donnée en nombre de groupes de souffle transcrits, pas de phrases.

	Nom	Nb. phrases (Nb. *tokens*)	Source
Alsacien	$T_{\texttt{gsw,Sv}}$	267 (5 110)	Wikipédia
	$T_{\texttt{gsw,Strv}}$	66 (1 768)	Nouvelle
Créole guadeloupéen	$T_{\texttt{gcf,Wiki}}$	74 (873)	Wikisphère
	$T_{\texttt{gcf,COCOON}}$	1 080 (9 175)	COCOON

Tableau 2. *Corpus collectés pour l'alsacien et le créole guadeloupéen*

Les corpus de référence, annotés manuellement par des chercheuses du laboratoire LiLPa de Strasbourg pour l'alsacien et par une étudiante guadeloupéenne, deux expertes de l'annotation et un dialectologue créolophone pour le créole guadeloupéen, contiennent respectivement 102 et 100 phrases. Leurs contenus sont détaillés dans le tableau 3.

	Nom	Nb. phrases (Nb. *tokens*)	Source
Alsacien	E_{Sv}	47 (875)	Wikipédia
	$E_{\text{Nv,1}}$	26 (362)	Pièce de théâtre
	$E_{\text{Nv,2}}$	29 (231)	Recettes
Créole guadeloupéen	$E_{\texttt{gcf,Wiki}}$	17 (238)	Wikisphère
	$E_{\texttt{gcf,COCOON}}$	83 (1 385)	COCOON

Tableau 3. *Corpus de référence annotés par des experts linguistes*

3.4.4. *Outils de préannotation*

Comme décrit dans la section 3.2.1, nous avons intégré deux outils de préannotation 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, selon la méthodologie définie par Bernhard et Ligozat (2013), ainsi que MElt (Denis et Sagot, 2010),

17. Office pour la langue et la culture d'Alsace, voir https://www.olcalsace.org/.
18. Collection de corpus oraux numériques, voir https://cocoon.huma-num.fr/.
19. Voir par exemple, sous licence CC BY-NC-SA : https://cocoon.huma-num.fr/exist/crdo/meta/crdo-GCF_1022.

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 *tokens* 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 *token*-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.

Nous savons que la préannotation 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 a un impact sur nos participants. Nous avons néanmoins observé, dans le cas de l'alsacien, que si les outils proposent la même étiquette dans 50 % des cas en moyenne, celle-ci est rejetée par les participants dans 12 % des cas.

3.5. *Résultats obtenus et discussion*

3.5.1. *Participation*

	Alsacien	Créole guadeloupéen
Nombre d'inscrits	208	35
Participants ayant finalisé la phase d'entraînement	75	17
Participants ayant produit des annotations	47	11
Jours d'annotation	109	9
Nombre d'annotations produites	24 588	1 205
Taille du corpus annoté (*tokens*)	7 973	933
Qualité des annotations produites (F-mesure)	0,93	0,87

Tableau 4. *Participation sur les deux plateformes*

La participation sur les plateformes est détaillée dans le tableau 4. L'écart entre les deux plateformes (plus de 200 participants pour l'alsacien et 35 pour le créole guadeloupéen) s'explique à notre avis par la différence d'énergie déployée à communiquer sur chacune des instances : la publicité de la plateforme `Bisame` a été réalisée à travers des communications sur les groupes Facebook de locuteurs, par le bouche-à-oreille, grâce au relais d'organisations comme le FILAL[20] ou d'entreprises telles que la Marque Alsace[21], par le biais d'une chronique diffusée sur France Bleu Elsass[22], et *via* la page Facebook du projet `Bisame`[23]. La plateforme `Krik` n'a pas bénéficié

20. Fonds international pour la langue alsacienne, voir `https://filalsace.net/`.
21. Voir : `http://www.marque-alsace.fr/`.
22. Voir : `https://www.francebleu.fr/elsass`.
23. Voir : `https://www.facebook.com/bisame.elsass/`.

d'un tel effort de communication, nos contacts étant moindres et l'expérimentation ayant été rapidement interrompue. Nous avons en effet constaté que 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 utilisé n'était pas adapté à l'annotation de l'oral. En effet, le corpus du créole est constitué en majorité de transcriptions et est découpé en groupes de souffle. Ces séquences à annoter se sont révélées inutilisables en l'état car 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 un autre projet incluant la production bénévole de corpus bruts étant en cours de développement (voir la section 4), nous avons mis cette instance en pause.

Nous avons également observé, et cela est valable pour les deux instances, qu'environ 40 % des participants ne produisent aucune annotation après avoir finalisé la phase de formation. On peut supposer que la durée de la formation (huit minutes en moyenne) ainsi que la nature de la tâche, difficile et rébarbative sont la cause de cette démotivation. Par ailleurs, bien que l'application soit conçue pour être utilisable sur téléphone mobile, son inconfort d'utilisation a été évoqué par plusieurs participants comme un facteur de découragement.

Nous avons réalisé une enquête auprès des participants de la plateforme `Bisame`[24] afin de recueillir des informations sur leurs genres, âges, niveaux d'études, et langues maternelles. Cette enquête montre qu'il existe une marge de progression quant à la participation, notamment des femmes. En effet, 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 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, les participants ont un niveau d'études élevé, 60 % d'entre eux ayant atteint au moins le niveau BAC + 4, ce qui participe à expliquer la bonne qualité des annotations obtenues.

3.5.2. *Ressources produites*

Les difficultés liées à la nature du corpus se ressentent dans la qualité des annotations produites sur la plateforme `Krik` (voir le tableau 4) : celles-ci atteignent une exactitude de 87 %, très en deçà de ce que nous observons sur `Bisame` (93 %). Notons tout de même que ces résultats sont supérieurs à celui obtenu pour une tâche semblable réalisée par travail parcellisé avec `CrowdFlower` : 84 % d'exactitude pour de l'annotation de *tweets* en anglais (Hovy *et al.*, 2014).

Pour comprendre la source des erreurs commises par les participants, nous avons corrigé manuellement le corpus annoté *via* la plateforme `Krik`. Outre les difficultés liées au corpus, évoquées plus haut, cette analyse nous a permis de révéler des li-

24. 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.

mitations de notre tokéniseur dues à l'apparition d'habitudes scripturales inconnues. 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. Ils doivent par conséquent être regroupés sous la forme *anba_la* pour être annotés comme adverbe. Ces cas de figure ont été intégrés progressivement au script de tokénisation grâce à de nouvelles règles. Nous avons par ailleurs analysé les erreurs commises par les participants de manière à en identifier les motifs récurrents. Par exemple, de nombreuses confusions existent entre les catégories ADJ et VERB dans le cas de l'alsacien, ou le cas de *té*, pouvant désigner le verbe « être » ou la particule désignant le passé en créole guadeloupéen. Ont ainsi été mis en évidence les cas les plus intrinsèquement ambigus requérant une vigilance particulière et devant être intégrés à la phase de formation ainsi qu'au guide d'annotation mis à disposition.

3.5.3. *Outils entraînés*

Nous avons utilisé les ressources produites pour entraîner le *tagger* `MElt`, et observé deux types de difficultés propres à chacune des plateformes[25]. Dans le cas du créole guadeloupéen, nous avons dû compenser la mauvaise qualité des annotations recueillies : la correction manuelle du corpus annoté a permis d'augmenter de 10 % les performances de l'outil entraîné, passant de 76 à 84 % d'exactitude. Dans le cas de l'alsacien, l'entraînement de différentes instances de `MElt` avec différents sous-corpus correspondant à des variétés spécifiques de l'alsacien a permis de mettre en avant la nécessité de prendre en compte ces variétés. L'intégration de deux lexiques préexistants (environ 40 000 entrées, décrits dans (Millour et Fort, 2018b)) à l'entraînement de `MElt`, permet d'améliorer les performances de l'outil sur les mots inconnus de près de 30 % en moyenne, mais ne suffit pas à produire une couverture lexicale suffisante pour compenser les variabilités dialectale et lexicale.

Ce cas est illustré dans le tableau 5 : avec trois variétés présentes en tout (SV et STRV dans le corpus d'entraînement, SV et NV dans le corpus d'évaluation), on constate que les meilleures performances sont obtenues lorsque le corpus d'entraînement et le corpus de test appartiennent à la même variété SV (83,7 %). D'autre part, on observe que les performances sur l'ensemble des corpus d'évaluation E_{SV}+$E_{\text{NV},1}$+$E_{\text{NV},2}$ augmentent très faiblement malgré l'ajout du corpus T_{STRV} (1 768 *tokens*). Nous observons que l'augmentation globale des performances (+ 1,2 point) peut se faire au détriment de la qualité d'annotation sur certains corpus d'évaluation pris séparément, ici E_{SV} (- 1,4 point). En effet, malgré une augmentation de 30 % de la taille du corpus d'entraînement, le pourcentage de mots inconnus est stable pour ce corpus et certains *tokens*, en quantité insuffisante ici pour conclure sur la nature de la baisse de performances, se retrouvent mal annotés. Ce phénomène met en avant l'importance de la prise en compte des différentes variétés de l'alsacien dans le développement d'outils d'annotation performants.

25. Une analyse complète des résultats et des performances des outils entraînés a été présentée dans (Millour et Fort, 2018a ; Millour et Fort, 2018b).

	E_{Sv}	$E_{\text{Nv},1}$	$E_{\text{Nv},2}$	E_{Sv}+$E_{\text{Nv},1}$+$E_{\text{Nv},2}$
T_{Sv}	**83,7**	78,7	71,3	77,9
Unk. *tokens*	40 %	65 %	62 %	52 %
$T_{\text{Sv}} + T_{\text{STRV}}$	82,3	**82,8**	**71,8**	**79,1**
Unk. *tokens*	40 %	37 %	61 %	47 %

Tableau 5. *Exactitude des outils entraînés pour l'alsacien*

3.5.4. *Enseignements tirés*

En ce qui concerne la participation et les quantités limitées de ressources produites, deux éléments d'analyse nous semblent importants à considérer.

D'une part, notre hypothèse de départ concernant la motivation des locuteurs à développer des outils pour leur langue s'est révélée insuffisante. En effet, si contribuer à la création de ressources langagières pour leur langue motive certains à venir participer, ils ne restent pas (Millour et Fort, 2017). La même observation a été faite dans un cadre extrême, celui d'une mission humanitaire visant à traduire des SMS pour aider les rescapés du tremblement de terre à Haïti en 2010 (Munro, 2013) : les volontaires se sont épuisés (dans tous les sens du terme) au bout de quelques semaines. Cette constatation rejoint les résultats de l'enquête concernant `ZombiLingo` : si certains jouent pour « aider les scientifiques », ceux qui reviennent et participent le plus le font pour le jeu (Fort *et al.*, 2017). Il est donc nécessaire d'ajouter des éléments ludiques pour favoriser la rétention des participants.

Par ailleurs, les variétés scripturales et dialectales inhérentes aux langues non standardisées posent plusieurs types de 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 perdre de contributeurs. 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.* »

La création d'une instance pour le créole guadeloupéen a montré la facilité technique de l'adaptation de la plateforme. Il est néanmoins indispensable de disposer de corpus bruts de taille et de qualité suffisantes pour permettre une annotation de qualité. En outre, la collaboration avec des locuteurs influents au sein de la communauté linguistique est apparue comme un facteur important de réussite pour ce type de projet participatif.

Enfin, l'absence de prise en compte de la variation peut conduire à une stagnation voire à 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 disponibles, ni de description des mécanismes de variabilité à l'œuvre. Le processus de collecte de corpus ne peut donc se faire sans l'intervention des locuteurs, ceux-ci étant les seuls à même de produire des corpus écrits qui soient *représentatifs* des variétés scripturales et dialectales en usage. Les ressources produites *via* les plateformes `Bisame` et `Krik` sont disponibles sous différentes licences libres fonctions des corpus bruts correspondants[26]. Les modèles de *taggers* entraînés sont également disponibles, sous licence CC BY-NC-SA[27].

4. `Recettes de Grammaire` : une plateforme autonome de création de ressources pour les langues non standardisées

La première plateforme a permis de valider une méthodologie tout en montrant ses limites. La suivante tire les enseignements de cette expérience, notamment en permettant la production de corpus (dans un premier temps, des recettes de cuisine) et la construction d'un lexique de variantes alignées. Nous avons également largement renforcé la fluidité du processus d'annotation et la ludification des tâches proposées.

4.1. *Produire et annoter des corpus dans différentes variétés de la langue*

La plateforme `Recettes de Grammaire` répond à trois objectifs : i) faire produire du corpus brut sous forme de recettes de cuisine et d'anecdotes, ii) faire corriger les annotations produites par un outil état de l'art et iii) faire produire des variantes scripturales et dialectales pour les mots des recettes. L'ajout de recettes, d'anecdotes et de commentaires est réalisé par le biais d'une interface classique inspirée de sites de cuisine existants. Outre son intérêt linguistique, la plateforme a donc également un rôle culturel, puisqu'elle permet de produire une base de données de recettes qui, on peut l'espérer, seront typiques de la région.

4.2. *Fluidifier le processus d'annotation*

Afin d'encourager un participant ayant ajouté une recette à l'annoter dans la foulée, nous avons mis en place le cheminement suivant :

1) le participant ajoute une recette ;

2) cette recette est préannotée à la volée par un outil état de l'art et le résultat de cette annotation est montré au participant ;

26. Les articles `Wikipédia` sous licence CC BY-SA, les autres textes sous licence CC BY-NC-SA.
27. Voir : `https://bisame.paris-sorbonne.fr/downloads` (pour l'alsacien) `https://krik.paris-sorbonne.fr/downloads` (pour le créole guadeloupéen).

3) si le participant accepte de corriger ces préannotations, il est renvoyé vers l'interface correspondante.

L'annotation, qui est en réalité une correction de la préannotation fournie par l'outil, se fait de manière séquentielle *par catégorie* : en cliquant sur une catégorie, le participant fait apparaître les préannotations correspondantes qu'il peut valider ou rejeter. La phase d'annotation est structurée en différentes étapes correspondant à trois niveaux de difficulté des catégories établis grâce à l'étude des annotations produites par les participants (voir 3.5.1) :

– *facile*, ayant obtenu plus de 0,95 de F-mesure (pour l'alsacien, NOUN et DET) ;

– *intermédiaire*, ayant obtenu une F-mesure comprise entre 0,90 et 0,95 (pour l'alsacien, ADP+DET, NUM, INTJ, PROPN, VERB et SYM) ;

– *difficile*, pour les catégories restantes (pour l'alsacien, ADJ, ADV, ADP, AUX, CONJ, PRON, PART, SCONJ et X).

Les trois étapes de l'annotation sont les suivantes :

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 droite du texte saisi (voir la figure 3). Les catégories hachurées sont les catégories *difficiles* qui requièrent une formation préalable. La formation à une catégorie consiste à présenter au participant une séquence de phrases dont certains *tokens* ont été annotés manuellement avec la catégorie donnée. Le participant doit valider les préannotations correctes et rejeter celles qui sont erronées pour pouvoir valider sa formation. Les catégories blanches sont les *intermédiaires* entre lesquelles le participant peut naviguer librement, les grisées sont celles qui n'ont pas été utilisées lors de la préannotation ;

3) une fois toutes les étiquettes issues de la préannotation examinées par le participant, il lui reste à annoter les *tokens* dont l'étiquette a été rejetée, en choisissant une étiquette parmi la liste complète.

Ce découpage permet de diminuer la complexité de la tâche par rapport à une annotation séquentielle. La formation est en outre plus ciblée. 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.

4.3. *Évaluer les participants, une gageure*

La nature des textes annotés ne permet pas d'introduire des phrases de référence sur lesquelles évaluer les participants. Il nous a donc été impossible de reproduire la méthodologie d'évaluation, pourtant efficace, présentée dans la section 3.2.3. Néanmoins, nous pouvons évaluer les participants en introduisant des préannotations volontairement erronées, pour un jeu de *tokens* connus et non ambigus. Par exemple,

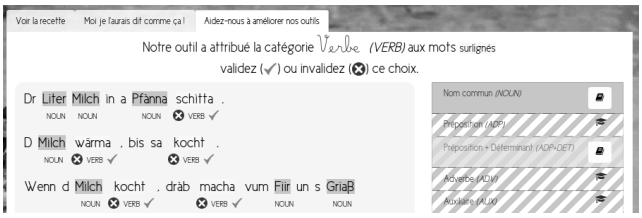


Figure 3. *Extrait de l'interface d'annotation pour la catégorie* VERB

la préannotation du mot « avec » *mìt*/ADV doit être corrigée en *mìt*/ADP. Les performances du participant sur ces *tokens* de test définissent son niveau de confiance. Cette méthode, sans doute moins efficace que la précédente, fera l'objet d'une attention particulière pour être améliorée au besoin.

4.4. *Annoter sa variété : mise en place d'un outil d'édition*

Afin de construire un lexique de variantes alignées nous renseignant sur les mécanismes de variations à l'œuvre et pouvant être intégré à l'entraînement d'outils supervisés, nous avons mis en place l'interface « Moi je l'aurais dit comme ça ! » (voir la figure 4). Elle permet à tout participant d'ajouter une variété scripturale ou dialectale d'un mot d'une recette. Les participants ayant la possibilité de placer leur lieu d'apprentissage de l'alsacien sur une carte de l'Alsace découpée en cinq aires dialectales, nous envisageons par ailleurs d'utiliser cette fonctionnalité pour leur proposer les contenus existants dans la variété qu'ils préfèrent afin de faciliter leur participation (voir 3.5.4).



Figure 4. *Ajout de la variante* Kugelhof *pour le mot* Kugelhopf

4.5. *Améliorer la ludification*

La plateforme `Recettes de Grammaire` est plus stylisée et personnalisée que les plateformes précédentes notamment grâce à un profil plus complet, un accès aux profils publics des participants et à leurs recettes, et à 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. Outre les badges s'accumulant à mesure que le participant ajoute recettes, anecdotes, variantes et annotations, nous avons introduit des badges de compétence obtenus à l'issue des formations réalisées sur les catégories *difficiles*. Le système de points a également été rendu plus interactif : deux participants proposant la même étiquette pour un *token* donné voient leurs points doubler. Le participant ayant annoté en premier en est averti *via* une fenêtre *pop-up* à la connexion suivante.

5. Discussions et conclusion

Les résultats de nos premières expériences de production participative d'annotations en parties du discours pour des langues peu dotées sont encourageants. Elles ont en effet abouti à la création d'une plateforme *open source*, d'un corpus annoté de 7 973 *tokens* pour l'alsacien, présentant une F-mesure de 0,93, et d'un premier corpus de référence annoté de 2 439 *tokens* pour le créole guadeloupéen. Les outils entraînés grâce à ces corpus atteignent une exactitude allant de 71 % à 84 % en fonction de la variété du corpus d'évaluation pour l'alsacien, et de 84 % pour le créole guadeloupéen. La mise en ligne des ressources annotées sur ORTOLANG[28], afin d'en assurer la pérennisation, est en cours. La démarche de production participative dans laquelle nous nous inscrivons est un processus cyclique dans lequel le dialogue avec le locuteur fait partie intégrante du développement. Les enseignements tirés de ces premières expériences nous ont donc conduites à développer la plateforme `Recettes de Grammaire` intégrant la collecte de corpus et de lexique de variantes alignées.

Cette nouvelle plateforme, lancée en juin 2018, a permis de recueillir à ce jour de premiers résultats encourageants (9 recettes, 515 annotations, 110 variantes dialectales et scripturales), qui sont en cours de traitement.

La participation à nos plateformes est très respectable si on la compare à d'autres portant sur des langues plus importantes en termes de nombre de locuteurs. Ainsi, la première version de `Phrase Detectives` (pour l'anglais) a attiré 2 000 joueurs en 32 mois (Chamberlain *et al.*, 2013), ce qui est proportionnellement bien inférieur pour une langue parlée par environ 350 millions de personnes[29]. Cependant, l'investissement des communautés concernées doit et peut être amélioré. Il apparaît notamment que les locuteurs militant pour la survie de leur langue n'ont pas conscience du fac-

28. Accessible ici : `https://www.ortolang.fr/`.
29. Selon `Wikipédia` : `https://en.wikipedia.org/wiki/English-speaking_world`.

teur aggravant que constitue l'absence de ressources numériques et d'outils adaptés. Nous pensons que les projets de production participative tels que le nôtre doivent ainsi également servir de vecteur de sensibilisation. Il nous revient donc d'atteindre la communauté des jeunes apprenants, par exemple en facilitant l'utilisation sur mobile des plateformes. Nous espérons enfin que la diversification des tâches présentée en section 4 équilibrera la répartition des participants, en termes de genre notamment.

Par ailleurs, la mobilisation des participants étant coûteuse quant à la communication qu'elle requiert, il est nécessaire que leur investissement soit exploité au mieux. L'utilisation de la préannotation et la fluidité d'utilisation des plateformes sont deux moyens mis en place permettant d'optimiser leur participation. Par ailleurs, l'étude des erreurs qu'ils commettent nous renseigne sur le niveau de difficulté *pour les annotateurs* et nous permet d'améliorer notre méthodologie à cet égard, notamment en proposant des formations ciblées. La comparaison des performances de `MElt` avec celle d'autres *taggers* supervisés nous permettra d'identifier où se situe la difficulté *pour les outils*, et de l'intégrer à la conception de notre plateforme.

Si l'interface `Recettes de Grammaire` ne rencontrait pas le succès espéré, la modularité des composants développés nous permettrait de modifier facilement le mécanisme d'incitation à la production de corpus bruts. Enfin, le code source de la nouvelle plateforme est librement disponible[30], ce qui permettra à tous ceux qui le souhaitent de l'améliorer et de l'adapter à leurs propres besoins.

Remerciements

Nous remercions vivement les participants du projet PLURAL (production ludique de ressources linguistiques pour le TAL) Langues et numérique 2018 (DGLFLF) : Delphine Bernhard, Bruno Guillaume et André Thibault, ainsi que les contributeurs des projets `Bisame` et `Krik` et l'OLCA pour son soutien. Nous remercions également la première relectrice pour ses nombreuses remarques et ses conseils.

6. Bibliographie

Agić Ž., Johannsen A., Plank B., Martínez H. A., Schluter N., Søgaard A., « Multilingual projection for parsing truly low-resource languages », *Transactions of the Association for Computational Linguistics*, vol. 4, p. 301-312, 2016.

Avanzi M., Stark E., « A crowdsourcing approach to the description of regional variation in French object clitic clusters », *Belgian Journal of Linguistics*, 2017.

Barre C., Vanderschelden M., *L'enquête "étude de l'histoire familiale" de 1999 - Résultats détaillés*, INSEE, Paris, 2004.

Barteld F., « Detecting spelling variants in non-standard texts », *Actes de Student Research Workshop (EACL 2017)*, Valence, Espagne, mai, 2017.

30. Voir : `https://github.com/alicemillour/Bisame/tree/recipes`.

Benjamin M., « Hard Numbers : Language Exclusion in Computational Linguistics and Natural Language Processing », *Actes de 11th International Conference on Language Resources and Evaluation (LREC'18)*, Miyazaki, Japon, mai, 2018.

Berment V., Méthodes pour informatiser les langues et les groupes de langues "peu dotées", Thèse, Université Joseph-Fourier - Grenoble I, mai, 2004.

Bernabé J., *La graphie créole*, Ibis Rouge edn, Guides du CAPES de Créole, 2001.

Bernhard D., Ligozat A.-L., « Es esch fàscht wie Ditsch, oder net ? Étiquetage morphosyntaxique de l'alsacien en passant par l'allemand », *Actes de TALARE (Traitement Automatique des Langues Régionales de France et d'Europe) (TALN'13)*, Les Sables d'Olonne, France, p. 209-220, juin, 2013.

Bernhard D., Ligozat A.-L., MARTIN F., Bras M., Magistry P., Vergez-Couret M., Steible L., Erhart P., Hathout N., Huck D., Rey C., Reynés P., Rosset S., Sibille J., Lavergne T., « Corpora with Part-of-Speech Annotations for Three Regional Languages of France : Alsatian, Occitan and Picard », *Actes de 11th edition of the Language Resources and Evaluation Conference (LREC'18)*, Miyazaki, Japon, mai, 2018.

Bernhard D., Todirascu A., MARTIN F., Erhart P., Steible L., Huck D., Rey C., « Problèmes de tokénisation pour deux langues régionales de France, l'alsacien et le picard », *Actes de DiLiTAL (Diversité Linguistique et TAL) (TALN'17)*, Orléans, France, juin, 2017.

Bettinson M., Bird S., « Developing a suite of mobile applications for collaborative language documentation », *Actes de 2nd Workshop on Computational Methods for Endangered Languages*, Honolulu, Hawaï, p. 156-164, mars, 2017.

Blachon D., Gauthier E., Besacier L., Kouarata G.-N., Adda-Decker M., Rialland A., « Parallel Speech Collection for Under-resourced Language Studies Using the Lig-Aikuma Mobile Device App », *Procedia Computer Science*, vol. 81, p. 61-66, 2016.

Böhmová A., Hajič J., Hajičová E., Hladká B., « The Prague Dependency Treebank : Three-Level Annotation Scenario », *in* A. Abeillé (ed.), *Treebanks : Building and Using Syntactically Annotated Corpora*, Kluwer Academic Publishers, 2001.

Boula de Mareüil P., Rilliard A., Frédéric V., « A Speaking Atlas of the Regional Languages of France », *Actes de 11th International Conference on Language Resources and Evaluation (LREC'18)*, Miyazaki, Japon, mai, 2018.

Branco A., « We Are Depleting Our Research Subject as We Are Investigating It : In Language Technology, more Replication and Diversity Are Needed », *Actes de 11th International Conference on Language Resources and Evaluation (LREC'18)*, Miyazaki, Japon, mai, 2018.

Callison-Burch C., Dredze M., « Creating speech and language data with Amazon's Mechanical Turk », *Actes de Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk (CSLDAMT '10) de NAACL HLT 2010*, Association for Computational Linguistics, Los Angeles, CA, États-Unis, juin, 2010.

Calvet L.-J., *Le marché aux langues : Essai de politologie linguistique sur la mondialisation*, Plon, 2002.

Chamberlain J., Fort K., Kruschwitz U., Lafourcade M., Poesio M., « Using Games to Create Language Resources : Successes and Limitations of the Approach », *in* I. Gurevych, J. Kim (eds), *The People's Web Meets NLP*, Theory and Applications of Natural Language Processing, Springer Berlin Heidelberg, p. 3-44, 2013.

Chamberlain J., Poesio M., Kruschwitz U., « A new life for a dead parrot : Incentive structures in the Phrase Detectives game », *Actes de WWW 2009*, Madrid, Espagne, avril, 2009.

Cohen K. B., Xia J., Zweigenbaum P., Callahan T., Hargraves O., Goss F., Ide N., Névéol A., Grouin C., Hunter L. E., « Three Dimensions of Reproducibility in Natural Language Processing », *Actes de 11th International Conference on Language Resources and Evaluation (LREC'18)*, Miyazaki, Japon, mai, 2018.

Colot S., Ludwig R., « Guadeloupean and Martinican Creole », *in* S. M. Michaelis, P. Maurer, M. Haspelmath, M. Huber (eds), *The survey of pidgin and creole languages.*, vol. 2, Oxford University Press, 2013.

Cox C., « Probabilistic tagging of minority language data : a case study using Qtag. », *Language & Computers*, vol. 71, n° 1, p. 213-231, 2010.

Crévenat-Werner D., Zeidler E., *Orthographe alsacienne - Bien écrire l'alsacien de Wissembourg à Ferrette*, Jérôme Do Bentzinger, 2008.

Dandapat S., Biswas P., Choudhury M., Bali K., « Complex Linguistic Annotation — No Easy Way out ! : A Case from Bangla and Hindi POS Labeling Tasks », *Actes de Linguistic Annotation Workshop*, ACL-IJCNLP '09, Stroudsburg, PA, États-Unis, p. 10-18, août, 2009.

Denis P., Sagot B., « Exploitation d'une ressource lexicale pour la construction d'un étiqueteur morphosyntaxique état-de-l'art du français », *Actes de Traitement Automatique des Langues Naturelles (TALN'10)*, Montréal, Canada, juillet, 2010.

Diki-Kidiri M., « Comment assurer la présence d'une langue dans le cyberespace », *UNESCO. Retrieved December*, vol. 31, p. 2007, 2007.

Fišer D., Tavčar A., Erjavec T., « sloWCrowd : a Crowdsourcing Tool for Lexicographic Tasks », *Actes de 9th International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Islande, mai, 2014.

Fort K., Adda G., Cohen K. B., « Amazon Mechanical Turk : Gold Mine or Coal Mine ? », *Computational Linguistics (editorial)*, vol. 37, n° 2, p. 413-420, juin, 2011.

Fort K., Guillaume B., Lefèbvre N., « Who wants to play Zombie ? A survey of the players on ZOMBILINGO », *Actes de Games4NLP 2017 - Using Games and Gamification for Natural Language Processing*, Symposium Games4NLP, Valence, Espagne, p. 2, avril, 2017.

Fort K., Sagot B., « Influence of Pre-annotation on POS-tagged Corpus Development », *Actes de ACL Linguistic Annotation Workshop*, Uppsala, Suède, p. 56-63, juillet, 2010.

Garcia M., Gamallo P., Gayo I., Cruz M. A., « PoS-tagging the web in portuguese. National varieties, text typologies and spelling systems », *Procesamiento de Lenguaje Natural*, vol. 53, p. 95-101, 2014.

Garrette D., Mielens J., Baldridge J., « Real-World Semi-Supervised Learning of POS-Taggers for Low-Resource Languages », *Actes de 51st Annual Meeting of the Association for Computational Linguistics (ACL-2013)*, ACL '13, Sofia, Bulgarie, p. 583-592, août, 2013.

Guillaume B., Fort K., Lefebvre N., « Crowdsourcing Complex Language Resources : Playing to Annotate Dependency Syntax », *Actes de International Conference on Computational Linguistics (COLING)*, Osaka, Japon, décembre, 2016.

Hana J., Feldman A., Brew C., « A Resource-light Approach to Russian Morphology : Tagging Russian using Czech resources », *Actes de Conference on Empirical Methods in Natural Language Processing (EMNLP)*, ACL, Barcelone, Espagne, p. 222-229, juillet, 2004.

Hazaël-Massieux M.-C., *Ecrire en créole : Oralité et écriture aux Antilles*, L'Harmattan, 2000.

Hovy D., Plank B., Søgaard A., « Experiments with crowdsourced re-annotation of a POS tagging data set », *Actes de 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2 : Short Papers)*, Baltimore, MD, États-Unis, p. 377-382, juin, 2014.

Jamatia A., Das A., « Part-of-Speech Tagging System for Indian Social Media Text on Twitter », *Actes de Workshop on Language Technologies For Indian Social Media (SOCIAL-INDIA)*, Goa, Inde, p. 21-28, novembre, 2014.

Klubička F., Ljubešić N., « Using crowdsourcing in building a morphosyntactically annotated and lemmatized silver standard corpus of Croatian », *Actes de 9th Language Technologies Conference*, Ljubljana, Slovénie, octobre, 2014.

Krumm J., Davies N., Narayanaswami C., « User-Generated Content », *IEEE Pervasive Computing*, vol. 7, n° 4, p. 10-11, octobre, 2008.

Lafourcade M., Joubert A., « JeuxDeMots : un prototype ludique pour l'émergence de relations entre termes », *Actes de Journées internationales d'Analyse statistique des Données Textuelles (JADT)*, Lyon, France, mars, 2008.

Leemann A., Kolly M.-J., Goldman J.-P., Dellwo V., Hove I., Almajai I., Grimm S., Robert S., Wanitsch D., « Voice Äpp : a mobile app for crowdsourcing Swiss German dialect data », *Actes de INTERSPEECH 2015*, Dresde, Allemagne, septembre, 2015.

Li S., Graça J. a. V., Taskar B., « Wiki-ly Supervised Part-of-speech Tagging », *Actes de 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, Jeju, Corée du Sud, p. 1389-1398, juillet, 2012.

Ljubešic N., Zupan K., Fišer D., Erjavec T., « Normalising Slovene data : historical texts vs. user-generated content », *Actes de 13th Conference on Natural Language Processing (KONVENS 2016)*, Bochum, Allemagne, p. 146-155, septembre, 2016.

Ludwig R., Montbrand D., Poullet H., Telchid S., « Abrégé de grammaire du créole guadeloupéen », *Dictionnaire créole français (Guadeloupe), avec un abrégé de grammaire créole et un lexique français-créole*, SERVEDIT, p. 17-38, 1990.

Malherbe M., *Les langages de l'humanité (une encyclopédie des 3000 langues parlées dans le monde)*, Collection Bouquins, Laffont, 1983.

McEnery T., Hardie A., *Corpus Linguistics : Method, Theory and Practice*, Cambridge Textbooks in Linguistics, Cambridge University Press, 2011.

Melero M., Costa-Jussà M. R., Domingo J., Marquina M., Quixal M., « Holaaa ! ! writin like u talk is kewl but kinda hard 4 NLP », *Actes de 8th International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turquie, mai, 2012.

Millour A., Fort K., « Why do we Need Games ? Analysis of the Participation on a Crowdsourcing Annotation Platform », *Actes de Games4NLP 2017 - Using Games and Gamification for Natural Language Processing*, Symposium Games4NLP, Valence, Espagne, avril, 2017.

Millour A., Fort K., « Krik : First Steps into Crowdsourcing POS tags for Kréyòl Gwadloupéyen », *Actes de 11th International Conference on Language Resources and Evaluation (LREC'18)*, Workshop CCURL, Miyazaki, Japon, mai, 2018a.

Millour A., Fort K., « Toward a Lightweight Solution for Less-resourced Languages : Creating a POS Tagger for Alsatian Using Voluntary Crowdsourcing », *Actes de 11th International Conference on Language Resources and Evaluation (LREC'18)*, Miyazaki, Japon, mai, 2018b.

Munro R., « Crowdsourcing and the Crisis-Affected Community : lessons learned and looking forward from Mission 4636 », *Journal of Information Retrieval*, 2013.

Petrov S., Das D., McDonald R., « A Universal Part-of-Speech Tagset », *Actes de 8th International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turquie, mai, 2012.

Pingali S., Mortensen D., Littell P., Levin L., Phonetically-Aware Approximate Search for Low-Resource Languages, Technical report, Carnegie Mellon University, Pittsburgh, PA, États-Unis, 2017.

Plank B., « What to do about non-standard (or non-canonical) language in NLP », *Actes de 13th Conference on Natural Language Processing (KONVENS)*, Bochum, Allemagne, p. 13-20, août, 2016.

Prado D., « Présence des langues dans le monde réel et le cyberespace », *Net.lang Réussir le cyberespace multilingue*, c&f édition edn, Vannini, Laurent and Le Crosnier, Hervé, 2012.

Rivron V., « L'usage de Facebook chez les Éton du Cameroun », *Net.lang Réussir le cyberespace multilingue*, c&f édition edn, Vannini, Laurent and Le Crosnier, Hervé, p. 171-178, 2012.

Sagot B., Informatiser le lexique, Habilitation à diriger des recherches en inguistique informatique, Institut national de recherche en informatique et en automatique (Inria), juin, 2018.

Samardzic T., Scherrer Y., Glaser E., « Normalising orthographic and dialectal variants for the automatic processing of Swiss German », *Actes de 7th Language and Technology Conference*, Poznań, Pologne, novembre, 2015.

Scherrer Y., Sagot B., « Lexicon induction and part-of-speech tagging of non-resourced languages without any bilingual resources », *Actes de Workshop on Adaptation of language resources and tools for closely related languages and language variants*, RANLP '13, Hissar, Bulgarie, septembre, 2013.

Steiblé L., Bernhard D., « Pronunciation Dictionaries for the Alsatian Dialects to Analyze Spelling and Phonetic Variation », *Actes de 11th edition of the Language Resources and Evaluation Conference*, Miyazaki, Japon, mai, 2018.

Toutanova K., Klein D., Manning C. D., Singer Y., « Feature-rich Part-of-speech Tagging with a Cyclic Dependency Network », *Actes de Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, Stroudsburg, PA, États-Unis, p. 173-180, mai, 2003.

Tseng H., Jurafsky D., Manning C., « Morphological features help POS tagging of unknown words across language varieties », *Actes de Fourth SIGHAN Workshop on Chinese Language Processing*, Jeju, Corée du Sud, p. 32-39, octobre, 2005.

Tuite K., « GWAPs : Games with a Problem », *Actes de 9th International Conference on the Foundations of Digital Games*, Liberty of the Seas, Caraïbes, avril, 2014.

Zaghouani W., Dukes K., « Can Crowdsourcing be used for Effective Annotation of Arabic ? », *Actes de 9th International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Islande, mai, 2014.