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 :

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

Membres du jury :

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

Position de thèse

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.

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 impactées par ces 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 terme 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 terme 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

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.

Les trois langues que nous avons étudiées au cours de cette thèse sont l'alsacien, le créole guadeloupéen et le créole mauricien.

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 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.

Conclusions

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, la plateforme de myriadisation d'annotations en parties du discours développée, 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 [1] 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).

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. 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.

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 raffinées : nous espérons que le travail amorcé dans cette thèse pourra être poursuivi et étendu dans ce sens.

Références

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.

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`.

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.

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`.

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.

1. Voir : `https://www.enetcollect.net`, juillet 2020.