Sciences participatives et diversité linguistique Retours d'expériences

Alice Millour, Karën Fort

Sciences participatives et diversité linguistique

Retours d'expériences

Alice Millour
Doctorante et attachée temporaire d'enseignement et de recherche
Sorbonne Université

Karën Fort
Maîtresse de conférences
EA 4509 - Sens Texte Informatique Histoire (STIH), Sorbonne Université

Certaines langues pâtissent d'un manque de ressources au sens large, qu'elles soient humaines, linguistiques ou financières, en particulier pour produire les outils de traitement automatique nécessaires à leur intégration numérique. Pour ces langues, dites « peu dotées », la production participative apparaît comme un moyen prometteur de mettre à profit la présence croissante de locuteurs sur Internet.

Les sciences participatives comme moyen de production de ressources linguistiques pour le traitement automatique des langues (TAL) ont fait leurs preuves sur des langues comme le français ou l'anglais. Les projets JeuxDeMots (production d'un réseau lexical), ZombiLingo (production de corpus annotés en syntaxe) ou Phrase Detectives (production de corpus annotés en anaphore)[1] témoignent de l'intérêt de former des locuteurs à des tâches linguistiques pour produire des ressources de qualité.
La plateforme « Bisame » (devenue « Recettes de Grammaire »[2]) a été développée pour tester la faisabilité d'une telle entreprise pour la production de ressources linguistiques pour les langues peu dotées non standardisées, en particulier l'alsacien, continuum dialectal parlé en Alsace et dans une partie de la Moselle.
Nous faisons part ici des obstacles rencontrés et des enseignements tirés de ce projet initié en 2016.

Production participative pour les langues non standardisées

Quelle que soit la langue concernée, une des difficultés majeures de la production participative est de parvenir à mobiliser une communauté d'internautes prêts à consacrer du temps à leurs contributions. Dans le cas particulier des langues peu dotées et non standardisées (sans norme orthographique), des obstacles additionnels doivent être pris en compte.
Considérons par exemple la construction d'un outil d'annotation en morphosyntaxe, l'un des premiers jalons de la chaîne de traitement automatique d'une langue consistant à associer chaque mot d'un texte à sa catégorie grammaticale (nom commun, adjectif, adverbe etc.). Traditionnellement, la mise au point d'un tel outil requiert :
1) La constitution d'un corpus « brut », suffisamment grand pour recouvrir l'ensemble des phénomènes linguistiques présents dans la langue (exemples).
2) L'annotation manuelle de ce corpus, chaque mot se voyant associer sa catégorie grammaticale, éventuellement par production participative.
3) L'entraînement de l'outil à proprement parler, sur la base des exemples présents dans le corpus.

1 www.jeuxdemots.org ; https://zombilingo.org ; https://anawiki.essex.ac.uk/phrasedetectives

2 . https://bisame.paris-sorbonne.fr/recettes/

Pour certaines langues, la constitution d'un corpus « brut » représente en elle-même une difficulté. En effet, les corpus disponibles immédiatement peuvent être de taille réduite (la section alsacienne de la Wikipédia alémanique, par exemple, contient moins de 60 000 mots, alors que la Wikipédia française en contient plus d'un milliard). Ce manque de ressources brutes est aggravé par les variabilités dialectale et scripturale observées dans les langues non standardisées : lorsqu'aucune orthographe ne fait consensus chez les locuteurs, la variabilité dialectale n'est pas « lissée » par une norme et les formes graphiques peuvent se multiplier pour un mot donné. Ainsi, le mot « lait » peut s'écrire sous au moins cinq formes différentes : « Mìlich », « Melech », « Mìlch », « Mélisch », « Melich ». À titre de comparaison, dans le cas du français par exemple, « moins » n'admet qu'une orthographe, que le « s » final soit prononcé ou non.
De ce fait, lors de notre première expérience d'annotation participative, certains participants se sont plaints d'avoir à contribuer sur une variante d'alsacien qui leur était peu familière.
Les locuteurs ne sont pas les seuls à être sensibles aux variétés dialectale et scripturale de leur langue : les outils de traitement automatique des langues le sont également. Ainsi, un outil d'annotation entraîné sur un corpus d'alsacien haut-rhinois verra ses performances chuter lorsqu'il sera évalué sur un texte en alsacien bas-rhinois. Constituer des corpus de taille suffisante pour chacune des variantes possibles étant inenvisageable, nous avons développé une fonctionnalité intitulée « moi j'aurais dit ça comme ça ! » permettant de recueillir des connaissances sur les mécaniques de ces variations. Cette fonctionnalité, à laquelle s'ajoute la possibilité de saisir des textes complets pouvant être annotés dans la foulée par les participants, permet de compenser le manque de ressources brutes disponibles.

Deux hypothèses culturelles fortes

L'alsacien est une langue dont la vitalité est en baisse, mais qui bénéficie tout de même du soutien d'une forte communauté linguistique, engagée dans la défense de sa langue, notamment au travers de cours d'alsacien, à l'Université ou dans le milieu associatif, et *via* l'action de l'Office pour la langue et les cultures d'Alsace et de Moselle (OLCA).
Un des enjeux cruciaux de la production participative quelle qu'elle soit est de parvenir à mobiliser une communauté d'internautes. Différents moyens peuvent être imaginés pour susciter et entretenir la motivation des participants, comme par exemple des micro-paiements, ou la dissimulation de la tâche à accomplir sous une couche de fonctionnalités ludiques rendant la participation plus attractive et plaisante.
Nos travaux portant sur des langues considérées comme vulnérables, nous avons fait l'hypothèse que la communauté de participants serait facile à constituer, les enjeux liés à la survie de leur langue suffisant à les mobiliser. L'existence d'une communauté d'internautes motivés a en effet été confirmée dans le cas de l'alsacien par le succès de notre première expérience d'annotation collaborative, sur une plateforme très peu ludique, au cours de laquelle 53 participants ont produit plus de 26 000 annotations.
Le patrimoine culinaire alsacien étant important et reconnu en France[3], nous avons choisi de focaliser la saisie de textes sur les recettes de cuisine, donnant une couleur plus régionale encore à notre projet.

Des résultats surprenants

Les résultats obtenus à ce jour sont satisfaisants en termes de qualité mais encore en deçà de nos attentes en termes de quantité : 9 recettes saisies (2 081 mots), 26 402 annotations produites, et

3 À titre d'illustration, 490 recettes ont reçu l'étiquette « Alsace » sur le site « 750 g » (www.750g.com/recherche.htm?search=Alsace).

347 variantes scripturales proposées pour 148 mots différents. Nous parvenons à entraîner un outil d'annotation, mais celui-ci sera d'autant plus « robuste » à la variation que nos ressources croîtront. Force est de constater que la participation n'a pas été à la hauteur de l'enthousiasme initialement généré par le lancement de notre plateforme, notamment concernant la production de textes bruts, une tâche *a priori* plus facile à réaliser que l'annotation en morphosyntaxe.
Un sondage que nous avons réalisé en ligne intitulé « L'alsacien, Internet, et moi » permet de proposer plusieurs hypothèses pour expliquer l'insuccès de cette fonctionnalité. Notamment, il apparaît qu'alors même que très peu de locuteurs connaissent l'existence de conventions orthographiques pour leur langue, la grande majorité d'entre eux estime avoir un niveau plus faible à l'écrit qu'à l'oral (si 56 % des répondants estiment avoir un « bon » niveau en production orale, et 47 % un « bon » niveau en compréhension écrite, seuls 17 % évaluent leur compétence de production écrite comme « bonne »). Cela permet d'expliquer le malaise des participants face à la production d'un contenu aussi structuré qu'une recette de cuisine, d'autant plus dans le cadre d'un projet de recherche, alors qu'ils ne sont pas sûrs d'écrire « correctement » (ce qui, dans le cas d'une langue non standardisée, n'a pas de sens). Nous avions sous-estimé cette réticence, conséquence à la fois de l'importance donnée à l'orthographe lors de l'apprentissage du français et de la relégation historique des langues régionales.
Le sondage mené nous enseigne également que la majorité de la production écrite en alsacien concerne des commentaires et conversations sur les réseaux sociaux. Si pour des raisons éthiques et légales, ces contenus ne peuvent pas être exploités, les futures versions de nos plateformes permettront de produire des contenus plus variés.

De la nécessité de rester à l'écoute des locuteurs

Cette expérience de production participative nous a convaincues de l'importance de nous inscrire dans un dialogue avec les locuteurs. Notamment, s'il existe une communauté motivée, celle-ci pourrait être davantage mobilisée grâce à une meilleure pédagogie sur les enjeux de la présence numérique des langues. Ce sondage sur les pratiques linguistiques a également été un moyen efficace d'entrer en contact avec les locuteurs. Enfin, les canaux officiels de communication se sont révélés décevants, il est donc fondamental d'identifier les « influenceurs ». Ainsi, la publication de l'enseigne alsacienne *Geht's in* concernant notre sondage a été partagée plus de 130 fois en quelques jours. *Vielmols* merci !