Vers une solution légère de production de données pour le TAL : création d'un *tagger* de l'alsacien par *crowdsourcing* bénévole

Alice Millour[1] Karën Fort[1] Delphine Bernhard[2] Lucie Steiblé[2]
(1) STIH - EA 4509, Université Paris-Sorbonne, France
(2) LiLPa - EA 1339, Université de Strasbourg, France
prenom.nom@paris-sorbonne.fr, {dbernhard,lucie.steible}@unistra.fr

RÉSUMÉ

Nous présentons ici les résultats d'une expérience menée sur l'annotation en parties du discours d'un corpus d'une langue régionale encore peu dotée, l'alsacien, *via* une plateforme de myriadisation (*crowdsourcing*) bénévole développée spécifiquement à cette fin : Bisame[1]. La plateforme, mise en ligne en mai 2016, nous a permis de recueillir 15 846 annotations grâce à 42 participants. L'évaluation des annotations, réalisée sur un corpus de référence, montre que la F-mesure des annotations volontaires est de $0,93$. Le *tagger* entraîné sur le corpus annoté atteint lui 82 % d'exactitude. Il s'agit du premier *tagger* spécifique à l'alsacien. Cette méthode de développement de ressources langagières est donc efficace et prometteuse pour certaines langues peu dotées, dont un nombre suffisant de locuteurs est connecté et actif sur le Web. Le code de la plateforme, le corpus annoté et le *tagger* sont librement disponibles.

ABSTRACT

Toward a lightweight solution to the language resources bottleneck issue: creating a POS tagger for Alsatian using voluntary crowdsourcing

We present here the results of an experiment on part-of-speech annotation of a corpus in a low-resourced regional language, Alsatian, using a specifically-developed voluntary crowdsourcing platform: Bisame.[1] It has been online since May 2016 and has allowed to gather 15,846 annotations, thanks to 42 participants. An evaluation performed on a reference corpus shows a F-measure of 0.93 of the produced annotations. The tagger trained on these annotations is accurate in 82% of the cases. This is the first POS tagger developed for Alsatian. This language resources development method proved to be efficient and promising for some low-resourced languages, for which a significant number of speakers have access to the Internet. The platform code, the annotated corpus and the tagger are all freely available.

MOTS-CLÉS : Annotation en parties du discours, *crowdsourcing*, alsacien, langues peu dotées.

KEYWORDS: POS-Tagging, crowdsourcing, Alsatian, low-resourced languages.

[1] Voir : http://bisame.herokuapp.com.

1 Introduction

La production de ressources langagières est notoirement coûteuse[1] et représente de ce fait un goulot d'étranglement (*language resources bottleneck*), qui limite le développement d'outils de traitement automatique des langues (TAL). Ce phénomène est d'autant plus prégnant pour des langues dont l'intérêt économique ou politique n'est pas immédiat ou dont le nombre de locuteurs est peu élevé.

Notre hypothèse de départ est qu'une simple plateforme Web légèrement ludifiée et permettant la formation et l'évaluation des participants pourrait suffire à recueillir des annotations, en particulier pour des langues présentant une communauté de locuteurs motivée et présente sur le Web, comme l'alsacien. En effet, le rapport réalisé à la demande la DGLFLF[2] sur la place des langues de France sur Internet (Pimienta & Prado, 2014) a montré que l'alsacien jouit d'une bonne présence sur Internet, grâce notamment au milieu associatif et aux initiatives de particuliers. On trouve ainsi des pages en alsacien sur la `Wikipédia` alémanique[3].

L'expérience que nous présentons ici concerne des annotations en parties du discours, que nous avons utilisées pour entraîner un étiqueteur en parties du discours (*tagger*) état de l'art pour le français, `MElt` (Denis & Sagot, 2010). Nous avons mené une double évaluation : celle des annotations des participants d'une part et celle de la qualité de l'outil entraîné avec ces annotations d'autre part.

Cette méthode d'annotation de corpus pourrait représenter une solution au moins partielle pour d'autres langues régionales comme le breton ou l'occitan, ou pour toute autre langue appartenant au groupe hétéroclite des langues peu dotées[4], en permettant de produire une brique de base indispensable à de nombreuses applications de TAL.

2 État de l'art

2.1 Existant pour l'alsacien

L'alsacien est en réalité un terme englobant, qui permet de regrouper les nombreuses variétés dialectales présentes en Alsace et une partie de la Moselle. Il s'agit de langues germaniques, majoritairement alémaniques, bien que des parlers franciques soient en usage au nord de la zone. Le bas-alémanique est la variété très largement majoritaire sur le territoire alsacien, et cette variété peut elle-même être découpée en deux sous-ensembles : le bas alémanique du nord, et celui du sud, grossièrement correspondant aux département du Bas et du Haut-Rhin, respectivement. Les changements ont lieu de proche en proche, les limites entre les variétés n'étant pas précises, mais relevant plutôt du continuum. Ces langues, contrairement à l'allemand standard, n'ont jamais été lissées par l'usage d'une forme normative écrite. De plus, la langue régionale étant utilisée simultanément au français, les locuteurs ont tendance à remplir leurs éventuels trous lexicaux par des termes français. Malgré les conséquences de la coexistence entre les alsaciens et le français, une étude de l'INSEE en 1999 recensait encore 550 000 locuteurs (Barre & Vanderschelden, 2004).

[1] Peu d'articles évaluent ce coût avec précision, mais la construction du `Prague Treebank` a été évaluée à 600 000 $ dans (Böhmová *et al.*, 2001).

[2] Délégation générale à la la langue française et aux langues de France, Ministère de la culture et de la communication.

[3] Voir : `http://als.wikipedia.org`.

[4] Pour une définition plus précise de ces langues, nous conseillons la lecture de la thèse de Berment (2004).

Les travaux existant pour l'heure sur l'étiquetage en parties du discours de l'alsacien sont très exploratoires. À notre connaissance, la seule expérience dans ce sens a été réalisée par Bernhard & Ligozat (2013). La méthode proposée consiste à remplacer les « mots-outils » alsaciens par leurs équivalents allemands, à l'aide d'un lexique bilingue de petite taille, pour ensuite appliquer des `taggers` existant pour l'allemand (en l'occurrence, `TreeTagger` (Schmid, 1997) et `Stanford POS Tagger` (Toutanova *et al.*, 2003)). Les mesures d'exactitude obtenues sur les données de test vont de 79 à 89 % en fonction du texte et du `tagger` utilisé.

2.2 *POS-tagging* pour les langues peu dotées

Outre l'exploitation de la parenté avec une langue étymologiquement proche et mieux dotée (voir par exemple (Scherrer & Sagot, 2013), (Bernhard & Ligozat, 2013)), plusieurs approches semi ou non supervisées tirent parti de l'existence de ressources additionnelles (voir par exemple (Li *et al.*, 2012) exploitant le `Wiktionnaire`), ou de corpus bilingues (*via* l'utilisation de réseaux neuronaux, voir par exemple (Zennaki *et al.*, 2016), ou *via* la projection d'annotation, voir par exemple (Agić *et al.*, 2016)) pour développer de nouveaux outils.

Néanmoins, la nécessité de pouvoir évaluer effectivement les outils développés sur chacune des langues considérées, ainsi que les caractéristiques spécifiques de l'alsacien, pour lequel il n'existe ni corpus bilingue exploitable, ni ressources lexicales suffisamment riches à ce jour, nous poussent à nous focaliser sur une méthodologie de construction de corpus annotés pouvant s'appliquer de manière indifférenciée à des langues pour lesquelles l'existant disponible est extrêmement restreint.

2.3 Myriadisation d'annotations en parties du discours

La myriadisation (*crowdsourcing*) consiste en un appel ouvert visant à faire produire des données (des entrées encyclopédiques, un dessin, un vote, etc.) à une masse de gens, aujourd'hui principalement *via* Internet. Selon Fort (2016), cette activité peut être considérée selon deux axes : la rémunération et la transparence de la tâche (le participant est-il immédiatement conscient de ce qu'il produit ?). Cette typologie permet de distinguer grossièrement la myriadisation bénévole (par exemple, `Wikipédia`), des plateformes de travail parcellisé (*microworking*, comme `Amazon Mechanical Turk`) et des jeux ayant un but (comme `JeuxDeMots`[5] (Lafourcade & Joubert, 2008) ou `ZombiLingo`[6] (Guillaume *et al.*, 2016)).

`Amazon Mechanical Turk` a été utilisé par de nombreux chercheurs, directement ou *via* `Crowdflower`, pour faire produire des annotations en parties du discours (voir, notamment, (Hovy *et al.*, 2014)). Outre les problèmes éthiques qu'il pose (Sagot *et al.*, 2011), ce type de plateforme n'est pas adapté aux langues que nous visons, du fait du manque de travailleurs les maîtrisant. Par ailleurs, les plateformes de travail parcellisé ne permettent pas de former les annotateurs, uniquement de les tester.

Les jeux ayant un but ont fait leurs preuves comme moyen d'obtenir des données langagières de qualité satisfaisante, à un coût moindre que les méthodes traditionnelles (Chamberlain *et al.*, 2013). Cependant, le développement d'un véritable jeu est une entreprise de longue haleine, qui nécessite des compétences variées (développement Web, mécanismes de jeux, design, etc.) et doit être rentabilisée

[5] Voir : `http://www.jeuxdemots.org`.
[6] Voir : `http://www.zombilingo.org`.

sur la durée[7]. Il est à noter que JeuxDeMots propose l'annotation en parties du discours, mais uniquement comme « bonus » et sans formation spécifique des joueurs.

Si nous n'avons pas connaissance d'une application de myriadisation bénévole pour l'annotation en parties du discours, le domaine est en plein essor et il n'est pas impossible qu'une plateforme, plus ou moins ludique, existe sans que nous le sachions[8]. Des tâches d'un type proche ont pu être réalisées avec succès par des bénévoles, notamment l'annotation de lettres de suicidés par des volontaires (Pestian *et al.*, 2012) ou la traduction de textos en situation d'urgence humanitaire (Munro, 2013)[9]. Enfin, il existe des plateformes génériques de sciences participatives telles que Crowd4U[10] ou Zooniverse[11], mais aucune ne propose encore d'application linguistique.

3 Méthodologie

3.1 Choix du jeu d'étiquettes

Dans un souci d'évaluer une méthodologie pouvant facilement être adaptée à différentes langues, nous avons choisi de travailler avec le jeu d'étiquettes universel pour les parties du discours (*Universal POS tagset*) présenté dans (Petrov *et al.*, 2012) synthétisant les jeux d'étiquettes de 22 langues et facilement adaptable aux spécificités de chaque langue[12]. Ce jeu d'étiquettes est présenté dans le tableau 1,

À ce jour, l'unique modification que nous avons apportée à ce jeu d'étiquettes de 17 catégories a été de faire correspondre à la catégorie X, habituellement associée à « Autre » (catégorie fourre-tout), uniquement les mots en langues étrangères.

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

Table 1: Liste des étiquettes utilisées selon le classement des créateurs du jeu d'étiquettes[13].

[7] Pour plus de détails sur ces sujets, voir (Lafourcade *et al.*, 2015).

[8] Une des raisons à cela est que ces plateformes ne font pas toujours l'objet de publications scientifiques, à l'image de LanguageQuiz : http://quiz.ucomp.eu/.

[9] Le projet a fait également usage de *microworking*.

[10] Voir : https://crowd4u.org.

[11] Voir : https://www.zooniverse.org.

[12] Voir : http://universaldependencies.org/u/pos/all.html.

[13] Nous reproduisons leur classification telle quelle, sans pour autant valider leur choix : certaines « classes fermées », comme les prépositions (ADP), étant très productives.

3.2 Présentation du corpus et des lexiques utilisés

Notre corpus de référence (C_{Ref}) est constitué de quatre textes totalisant 1 262 tokens (102 phrases) annotés manuellement par deux des autrices. Ce corpus, comme tous ceux utilisés dans notre expérience, est librement disponible sous licence CC BY-NC-SA[14]. La proportion des étiquettes présentes dans C_{Ref} est donnée dans le tableau 2.

ADJ	ADV	INTJ	NOUN	PROPN	VERB	ADP	AUX	CONJ	DET	NUM	PRON	PART	SCONJ	SYM	X
4 %	6 %	<1 %	18 %	7 %	13 %	12 %	5 %	5 %	11 %	4 %	6 %	<1 %	1 %	<1 %	7 %

Table 2: Proportion de chaque étiquette dans le corpus de référence.

Le corpus brut (C_{Brut}) à annoter contient, au moment de la rédaction de cet article, 578 phrases, soit 6 288 tokens, et est constitué de six articles de la `Wikipédia` alémanique. Tous les articles de C_{Brut} sont rédigés en bas alémanique du sud (haut-rhinois, noté ici HR). Le corpus de test est quant à lui composé pour moitié de textes en bas alémanique du nord (bas-rhinois, noté ici BR), et du sud. Le tableau 3 détaille le contenu des corpus.

Étant donné le peu de corpus disponibles dans cette langue, nous avons privilégié une approche pragmatique (McEnery & Hardie (2011) parlent de « corpus opportuniste ») et avons intégré ce dont nous pouvions disposer librement, afin de pouvoir le redistribuer. Ce choix n'est pas neutre et a entraîné un déséquilibre dans le corpus, principalement composé de textes issus de la `Wikipédia` alémanique. Celle-ci comprend environ 50 000 mots (les articles présentant des lieux, nombreux et très uniformes entre eux, ayant été exclus). Ce corpus est par ailleurs hétérogène en terme de variantes orthographiques et conventions graphiques utilisées par les contributeurs.

	Source	Nb. phrases		Nb. tokens	
C_{Ref}	Wikipédia [15] (BR)	47	102	875	1 468
	Recettes de cuisine [16] (HR)	29		362	
	Pièce de théâtre [17](HR)	26		231	
C_{Brut}	`Wikipédia` [18](BR)	578		6 288 (4 105 annotés)	

Table 3: Description des corpus.

Nous avons par ailleurs intégré deux lexiques lors de l'entraînement de MElt. Le premier (L_{MO}) comprend uniquement des « mots-outils » (déterminants, pronoms, prépositions, conjonctions, particules, adverbes et verbes fréquents totalisant 322 entrées) et a été compilé par Bernhard & Ligozat (2013). Le second (L_{gsw}) est plus grand (plus de 40 000 entrées) et contient des entrées issues de divers lexiques bilingues : lexiques thématiques du site Web de l'OLCA (Office pour la Langue et la Culture d'Alsace)[19], dictionnaire bilingue du site de l'Association Culture et Patrimoine d'Alsace (ACPA)[20] et dictionnaire multilingue français-allemand-anglais-alsacien (Adolf, 2006). L'utilisation de différentes sources permet de couvrir des formes de scripturalisation variées. Ainsi, on trouve les graphies suivantes pour le mot « coude » : *Elleböje* (OLCA), *Ellaboja* (OLCA), *Elleboje* (ACPA), *Ällabooga* (ACPA).

[14] Voir : `https://bisame.herokuapp.com/corpora`.
[19] Voir : `http://www.olcalsace.org/`.
[20] Le site n'est malheureusement plus en ligne.

3.3 Préparation des données

Les textes ont été tokénisés grâce à un script développé en langage Python. Outre les signes de ponctuation classiques permettant de délimiter les tokens, (« ? », « « » , « & » etc.), les séparateurs spécifiques de l'alsacien sont pris en compte. Par exemple, dans le contexte « *ich mach's* » (« je le fais »), « *mach* » (« fais ») est considéré comme un token car « *'s* » (pronom « le ») est considéré comme un séparateur. Les séparateurs sont par ailleurs considérés comme des tokens à part entière.

Le corpus complet est ensuite pré-annoté avec deux *taggers* : i) `TreeTagger` (Schmid, 1997) appliqué aux textes alsaciens après transposition des « mots-outils » en allemand, selon la méthodologie proposée par (Bernhard & Ligozat, 2013)[21], et ii) `MElt` (Denis & Sagot, 2010) entraîné au fur et à mesure de l'enrichissement du corpus d'entraînement.

3.4 Conditions de la participation bénévole

3.4.1 Recrutement

Outre la communication réalisée par l'OLCA [22] et les diffusions par diverses pages `Facebook` regroupant des dialectophones, un temps conséquent a dû être investi pour rentrer en contact avec des participants potentiels, leur expliquer le projet et répondre à leurs questions. Nous avons entrepris de contacter directement *via* `Facebook` les utilisateurs déclarant parler alsacien et s'exprimant dans des groupes tels que le « Centre Culturel Alsacien / Elsässisches Kulturzentrum » ou « Alsace Bilingue » . Cela nous a permis d'atteindre une nouvelle communauté à partir de laquelle l'information s'est propagée de proche en proche de manière plus efficace que grâce à la communication d'un organisme public.

En définitive, depuis la mise en ligne de la plateforme en mai 2016, 180 personnes ont créé un compte, 64 ont complété la phase de formation et 42 ont effectivement produit des annotations.

Trois périodes d'annotation en mai 2016, novembre 2016 et janvier 2017, impulsées par des communications sur les réseaux sociaux et des mails de relance auprès des participants déjà inscrits, ont conduit à la production de 15 846 annotations. Les 42 participants effectifs se répartissent par intervalle de nombre d'annotations produites, comme illustré dans le tableau 4[23] et seuls neuf d'entre eux ont produit plus de 250 annotations. Notre expérience confirme donc le phénomène déjà décrit, notamment dans (Chamberlain *et al.*, 2013) : peu de personnes participent (et produisent) beaucoup.

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

Table 4: Répartition des participants par intervalle de nombre d'annotations produites.

Il est cependant intéressant de noter qu'au cours du seul mois de janvier 2017, 6 200 annotations ont été produites sur C_{Ref} (soit 75 %) par 12 participants. Parmi eux, six ont produit entre 50 et

[21] `Treetagger` a été remplacé au cours de la rédaction de cet article par le `Stanford POS Tagger` (Toutanova *et al.*, 2003), qui présente de meilleures performances.

[22] Office pour la Langue et la Culture d'Alsace, voir `https://www.olcalsace.org/`.

[23] Ces intervalles ont été choisis de manière empirique.

150 annotations, quatre entre 400 et 500 et deux respectivement 831 et 3 057. Ces derniers résultats témoignent de la dynamique positive dont jouit la plateforme au moment de la rédaction de cet article.

3.4.2 Formation

Pour chaque étiquette, un aide-mémoire constitué d'exemples et des cas problématiques ayant été identifiés est consultable sous la forme d'un menu déroulant présenté dans la figure 1. Le contenu de cette documentation s'inspire très largement du guide d'annotation rédigé par Bernhard *et al.* (2016).

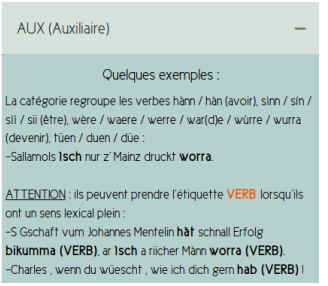

Figure 1: Extrait du menu déroulant de l'aide-mémoire relatif à chaque catégorie.

Avant de pouvoir produire des annotations, les participants doivent passer par une phase de formation au cours de laquelle leur sont présentées quatre phrases de C_{Ref} à annoter complètement. Le participant ne peut pas passer à la phrase suivante tant que toutes les étiquettes ne sont pas correctes. En cas d'erreur, un rappel reprenant les informations de l'aide-mémoire pour chaque catégorie erronée choisie est affiché. Les annotations produites au cours de cette phase ne sont pas enregistrées.

3.4.3 Production d'annotations

Les participants corrigent des pré-annotations, nous parlerons cependant d'annotation et non de correction, comme il est d'usage de le faire dans le domaine. Une séquence d'annotation comprend quatre phrases, dont trois sont tirées aléatoirement de C_{Brut} et une provient de C_{Ref}. Cette dernière nous permet d'évaluer le participant à l'issue de chaque séquence annotée.

Selon les résultats produits par les outils de pré-annotation, l'annotation $Ann_{T,U,C}$ d'un token T par un participant U avec la catégorie C, peut-être réalisée de deux manières :

- par attribution d'une étiquette à partir de la suggestion des deux propositions données par les *taggers* lorsqu'ils sont en désaccord (voir figure 2.),
- par validation ou rejet de l'étiquette proposée par les deux *taggers* lorsqu'ils sont d'accord (voir figure 3.). En cas de rejet de l'étiquette, les deux catégories ayant le meilleur score de confiance (défini ci-après) sont proposées au participant.

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

Figure 2: Annotation directe.

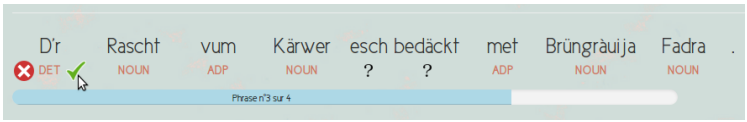

Figure 3: Annotation par validation de l'étiquette suggérée.

Les phrases de référence servant à la formation et à l'évaluation du participant sont proposées sans suggestion de catégorie. Ce point devra être corrigé afin que le participant ne puisse pas identifier durant la phase de production d'annotation quelles phrases sont issues de C_{Ref}.

Le score de confiance $Confiance_U$ attribué au participant U ayant produit $NbAnn_{Ref}$ annotations sur des phrases issues de C_{Ref} correspond au taux de catégorisations effectuées correctement, soit :

$$Confiance_U = \frac{NbAnn_{Ref,Correcte}}{NbAnn_{Ref}}$$

Ce score est recalculé à l'issue de chaque séquence annotée.

Nous associons à une annotation créée sur un token T un score de confiance égal au score de confiance du participant au moment de l'annotation, soit : $ScoreAnn_{T,U,C} = Confiance_U$. Ce score correspond à la probabilité que l'attribution d'une étiquette C à un token T par le participant U soit correcte.

Pour chaque token T, nous déterminons une catégorie unique C_T grâce aux scores de confiance attribués aux annotations[24]. Ainsi, dans la phrase : *Dr Mentelin hàt sina Stroßburger Drukaréi grinda.* (« Mentelin a fondé son imprimerie strasbourgeoise »), le token $T =$ *Stroßburger* a été annoté de trois manières $\{C_{i=1..3}\}$ par cinq participants différents, $\{U_{j=1..5}\}$. Le score de confiance $Score_{T,C_i}$ associé à chaque catégorie C_i est obtenu en calculant l'opposé de la probabilité jointe d'erreur pour les événements indépendants que constituent les annotations Ann_{T,U_j,C_i} produites, soit :

$$Score_{T,C_i} = 1 - \prod_j (1 - ScoreAnn_{T,U_j,C_i})$$

[24]Ce mode de calcul, préféré à un simple vote majoritaire, a fait ses preuves dans `ZombiLingo` (Guillaume *et al.*, 2016).

et

$$C_T = \arg\max_i(Score_{T,C_i})$$

On obtient dans notre cas $C_{\text{Stroßburger}}$ =ADJ.

Ce cas est illustré dans le tableau 5.

Catégorie C_i choisie	Score de l'annotation $ScoreAnn_{T,U_j,C_i}$	Score de l'étiquette $Score_{T,C_i}$
PROPN	$0,935$	$0,935$
ADJ	$0,875$	**$0,991$**
	$0,846$	
	$0,938$	
NOUN	$0,25$	$0,25$

Table 5: Calcul de l'étiquette la plus probable pour un token donné.

4 Résultats obtenus

4.1 Corpus annoté

Parmi les annotations produites depuis mai 2016, 7 750 ont été ajoutées sur C_{Ref} (qui comprend 1 468 tokens) lors de l'évaluation des participants. Les 8 096 annotations (de 4 336 tokens, soit 253 phrases) produites sur C_{Brut} (qui comprend 6 288 tokens), ont été utilisées pour entraîner `MElt`. Le nombre important d'annotations réalisées sur C_{Ref} s'explique par la répétition des phrases servant à l'évaluation, une phrase de C_{Ref} étant annotée lors de chaque séquence de quatre phrases.

Nous évaluons la qualité du corpus annoté en calculant, par catégorie, la F-mesure des annotations produites par les participants par rapport à la référence. Les résultats ainsi obtenus sont présentés en figure 4.

La F-mesure moyenne arithmétique calculée en janvier est de $0,85$. En revanche, la moyenne pondérée par les effectifs des F-mesures par catégorie atteint $0,93$. En effet, les trois catégories PART, SYM et INTJ, pour lesquelles les résultats sont inférieurs à $0,5$, représentent chacune moins de 1 % du corpus et sont par conséquent peu annotées. Il est à noter que Hovy *et al.* (2014) obtiennent dans leur expérience d'annotation par travail parcellisé (de *tweets* en anglais) une exactitude d'environ 80 %[25], bien inférieure, donc, à celle obtenue dans notre expérience (93 %).

Nous observons dans les deux cas que la qualité de l'annotation augmente avec le nombre de participations, confirmant les résultats obtenus par Guillaume *et al.* (2016) : le nombre d'annotations ayant doublé, nous avons constaté entre juin 2016 et janvier 2017 un gain sur la F-mesure moyenne de plus de 40 %.

[25]La complexité de la tâche d'annotation en parties du discours, évaluée dans (Fort *et al.*, 2012), est pourtant relativement faible lorsque le texte est pré-annoté.

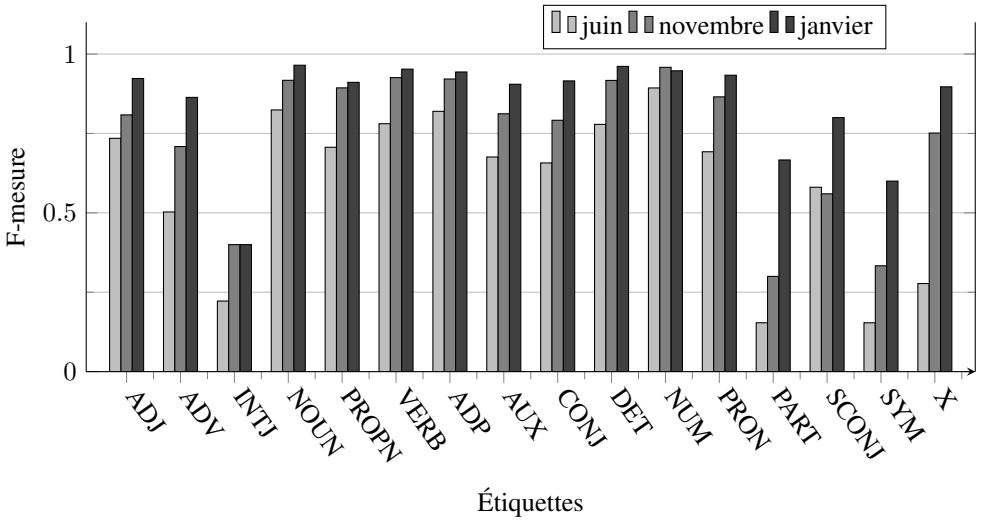

Figure 4: Comparaison par étiquette de la F-mesure des annotations produites par les participants sur C_{Ref} en juin 2016 (3 436 annotations), novembre 2016 (5 888 annotations) et janvier 2017 (7 750 annotations).

Au final, 25 % des erreurs concernent la catégorie ADV, confondue dans un tiers des cas avec la catégorie ADJ. Les erreurs concernant la catégorie VERB (19 % du total) sont à 75 % dues à la confusion avec la catégorie AUX. Par ailleurs, la catégorie X a entraîné une confusion entre le cas de l'alternance codique, par exemple: *Toi* /X *tais-toi* /X, et le cas des emprunts, par exemple : *bis an de plafond* /NOUN (« jusqu'au plafond »). Ces observations révèlent des difficultés récurrentes que la documentation proposée doit traiter plus efficacement, afin de s'adapter aux conditions d'annotation, par exemple en proposant des tests simples pour désambiguïser les catégories.

Le corpus ainsi produit est le premier corpus de l'alsacien annoté en parties du discours disponible librement sous licence CC BY-NC-SA[26]. Si sa taille est encore réduite, il continue à grossir et nous espérons atteindre rapidement les 20 000 tokens annotés, moyennant quelques modifications dans la plateforme que nous détaillons en conclusion.

4.2 Performances du *tagger*

Afin d'évaluer les performances du *tagger* entraîné avec les annotations des participants, nous avons procédé par validation croisée sur dix blocs. Pour chaque bloc, nous entraînons le *tagger* en extrayant de 10 à 250 phrases du corpus annoté et testons son exactitude sur un bloc non-contigu de C_{Ref} représentant 20 % du corpus global (environ 80 phrases). Les résultats moyennés sont présentés en figure 5. À la rédaction de cet article, l'exactitude de l'outil est de 82 %[27].

L'influence des lexiques intégrés à `MElt` (décrits en partie 3.2) est visible sur la figure 5.

[26] Voir : `https://bisame.herokuapp.com/corpora`.

[27] Le modèle entraîné est disponible à l'adresse `https://bisame.herokuapp.com/melt`.

L'ajout du lexique L_{MO} à l'entraînement de `MElt` apporte un gain allant de 2 % à 7 % en fonction de la taille du corpus d'entraînement, tandis que le gain additionnel apporté par L_{gsw} est inférieur à 1 %. On observe par ailleurs que le gain en pourcentage de mots connus (du corpus d'entraînement et du lexique additionnel), s'il est de 25 % lorsque la taille du corpus d'entraînement passe de 50 à 100 phrases, tombe à 5 % lorsqu'elle passe de 200 à 250 phrases. Cela peut s'expliquer par la faible couverture du lexique, due aux nombreuses variantes graphiques pouvant exister pour chaque mot, ainsi que par la faible diversité des textes présents dans le corpus d'entraînement.

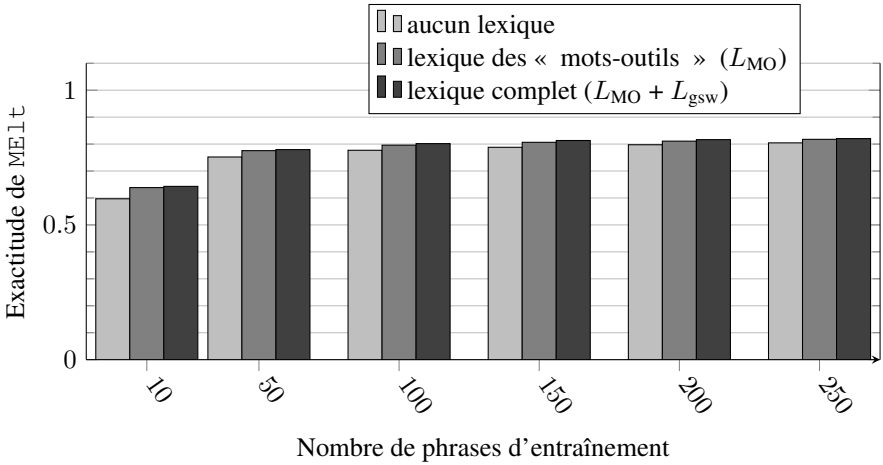

Figure 5: Performances de `MElt` (exactitude) selon la taille en nombre de phrases du corpus d'entraînement et le lexique additionel intégré.

Nous avons pu comparer l'exactitude obtenue par MElt aux meilleurs résultats obtenus par Bernhard & Ligozat (2013) sur deux textes issus de C_{Ref} : pour le texte extrait d'une pièce de théâtre, l'exactitude obtenue (66 %) est bien en deçà des 83 % obtenus par le `Stanford Tagger`. Néanmoins, pour l'article *Elsassisch Museum (Stroßburri)* de la `Wikipédia` alémanique, nous obtenons une exactitude de 84 %, plus proche des performances du `Stanford Tagger` (85 %).

Bien que les faibles tailles de ces deux corpus (respectivement 230 et 396 tokens) ne permettent pas de conclure sur les raisons de cette différence, nous proposons deux hypothèses : le texte théâtral est le seul texte du corpus constitué de dialogues, registre absent du corpus d'entraînement. Il est par ailleurs rédigé en bas-rhinois alors que C_{Ref} est constitué exclusivement de textes rédigés en haut-rhinois.

Si la méthode proposée par Bernhard & Ligozat (2013) semble être davantage insensible aux variantes dialectales, nous observons que le `MElt` atteint 81 % d'exactitude sur le sous-corpus de référence des deux textes rédigés en haut-rhinois (soit la variante du corpus d'entraînement), contre 72 % sur le corpus rédigé en bas-rhinois. Cela nous encourage à prendre en considération les variantes dialectales dans notre collecte de corpus, d'autant qu'il est plus aisé pour un participant d'annoter un texte rédigé dans sa variante maternelle.

5 Discussion

5.1 Choix du jeu d'étiquettes

L'utilisation stricte de l'*Universal POS Tagset* pose certains problèmes, notamment dans le cas fréquent des combinaisons ADP (préposition) + DET (déterminant), par exemple *vum* (*von* + *dem*), retrouvées également en allemand standard. L'alsacien autorise par ailleurs la formation de combinaisons n'existant pas en allemand standard, telles que *nonet*, contraction de *noch*/ADV et *net*/PART (« pas encore »). Ces combinaisons inattendues sont susceptibles d'apparaître du fait de l'absence de norme orthographique établie.

À la manière de Hollenstein & Aepli (2014) qui ont adapté le Stuttgart-Tübingen-TagSet (STTS) (Schiller *et al.*, 1995), standard pour l'allemand, aux caractéristiques du suisse allemand, nous envisageons d'étendre ce jeu d'étiquettes en ajoutant la possibilité de signaler la concaténation de catégories grammaticales lors de l'annotation d'un token (par exemple ADP+ signalerait un token combinant une préposition et un token d'un autre type).

5.2 Taille des corpus et évaluation

L'évaluation de `MElt` au fur et à mesure des annotations produites nous a permis de mettre en avant que notre méthode d'évaluation des annotations produites par report du score de confiance accordé au participant est insuffisante (98 % des annotations produites présentent un score de confiance supérieur à $0,8$). En effet, un participant au score de confiance élevé (supérieur à $0,9$) a produit un grand nombre d'annotations de mauvaise qualité (en commettant des erreurs récurrentes et du fait de la traduction automatique de son navigateur faussant sa lecture des tokens) sans que cela ne soit détecté. La faible taille de C_{Ref} ne nous permet d'évaluer le participant que sur un échantillon extrêmement réduit de cas possibles. Travailler à la construction d'un corpus de référence pensé spécifiquement pour l'évaluation de la tâche de catégorisation en parties du discours est donc indispensable.

Pour les mêmes raisons, nous n'avons pas pu évaluer `MElt` sur un corpus nouveau n'ayant pas déjà été exploité pour former les participants ou les évaluer. Il existe donc un biais dans l'évaluation du *tagger*.

Par ailleurs, nous n'avons pas réalisé d'évaluation directe de l'impact de la pré-annotation sur la qualité de l'annotation. Fort & Sagot (2010) ont démontré que ce biais existe et qu'il est plus marqué chez les annotateurs les moins formés, il est donc probable qu'il affecte les annotations obtenues, malgré la phase de formation obligatoire. Une évaluation de ce point précis doit devra donc être prévue.

5.3 Attractivité de la plateforme et dimension ludique

Un élément clé de la réussite d'une telle campagne est d'identifier la « bonne » communauté de locuteurs (connectée et motivée) et de l'atteindre[28]. Nous fondions, en début de projet, beaucoup

[28] Voir (Cosquer *et al.*, 2012) pour plus de détails sur les communautés de participants aux sciences participatives.

d'espoir sur le rayonnement de structures et de médias officiels (OLCA[29], France Bleu Elsass [30]) mais ceux-ci se sont révélés peu mobilisants. Il est à noter néanmoins qu'une contributrice majeure (troisième en terme de nombre d'annotations produites) a eu connaissance du projet *via* l'émission diffusée sur France Bleu Elsass.

Il semblerait que les locuteurs de l'alsacien présents sur le Web soient répartis dans des micro-communautés qu'il est difficile d'atteindre grâce à une communication unique. Les augmentations notables du nombre de participants coïncident en réalité avec des prises de contact directes. Notamment, le relais effectué par mail par des membres de diverses associations comme Le FILAL[31] ou d'entreprises telles que la Marque Alsace[32], ainsi que la campagne de recrutement réalisée *via* `Facebook` grâce à l'identification de groupes de dialectophones relayant la page dédiée au projet se sont révélées efficaces. Il est également probable que le cumul des diffusions ait contribué à convaincre des participants de se rendre sur la plateforme.

Ces efforts à mener pour attirer les participants (motivation) s'accompagnent d'une difficulté à les faire revenir (volition (Fenouillet *et al.*, 2009)) pour faire vivre la ressource. À deux exceptions près, les participants ne sont revenus sur la plateforme que lorsque des relances par mail ont été effectuées. La plateforme n'est pas un jeu et ne propose à l'origine qu'une seule fonctionnalité ludique, un classement par points (égal au nombre d'annotations multiplié par le score de confiance du participant). Nous avons par ailleurs noté que l'ajout d'une barre de progression indiquant l'état d'avancement de l'annotation du corpus en cours en pourcentage de tokens annotés a permis d'augmenter le nombre de séquences annotées lors d'une session active.

En moyenne, les dix participants ayant produit le plus grand nombre d'annotations se sont connectés quatre jours et ont annoté une quinzaine de séquences de quatre phrases depuis la mise en ligne de la plateforme[33]. Ces observations, ajoutées à la dynamique positive discutée en section 3.4.1 montrent que, si nous commençons à trouver « notre » communauté, l'application n'est pas encore assez attractive et ludique.

L'équilibre entre fonctionnalités ludiques, formation et évaluation des participants, et légèreté de la plateforme est donc encore perfectible.

6 Conclusion et perspectives

La plateforme que nous avons créée a permis de recueillir 15 846 annotations auprès de 42 participants et de construire ainsi le premier corpus librement disponible de l'alsacien annoté en parties du discours, de 4 105 tokens. Ce corpus a servi à entraîner le premier *tagger* spécifique à l'alsacien.

La qualité de l'annotation produite par les participants ($0,93$ de F-mesure), ainsi que les performances du premier *tagger* produit (82 % d'exactitude) montrent la validité de notre démarche.

Le système doit cependant être amélioré, afin de palier les problèmes que nous avons identifiés : i) manque de variété des corpus disponibles, ii) évaluation incomplète des participants et iii) ludification

[29] Office pour la Langue et la Culture d'Alsace, voir `https://www.olcalsace.org/`.

[30] Voir : `https://www.francebleu.fr/elsass`.

[31] Fonds international pour la langue alsacienne, voir `https://filalsace.net/`.

[32] Voir : `http://www.marque-alsace.fr/`.

[33] Afin de ne pas fausser ces moyennes, le participant le plus productif (96 séquences annotées pour neuf jours de connexion) a été exclu de ces observations.

insuffisante.

En ce qui concerne les corpus, nous envisageons de suivre la suggestion de Liberman (2016) et de faire créer des corpus par les participants eux-mêmes, dans leur variante de la langue considérée. Certains participants nous ont déjà proposé des textes, que nous avons dû refuser pour la plupart pour des questions de droit, mais nous allons mettre en place une procédure d'information et de recueil de textes qui permettra de profiter de leur élan en ce sens.

Afin d'améliorer la qualité des annotations produites, nous prévoyons de repenser notre guide d'annotation et notre stratégie d'évaluation des participants, au regard des difficultés que nous avons pu constater. L'ajout d'une interface d'administration intégrant le suivi des performances des participants serait également précieux.

Enfin, nous réfléchissons à des moyens de développer la dimension ludique sans pour autant alourdir la plateforme, par exemple grâce à des affichages de motivation (du type « X est en avance de Y points sur vous, rattrapez-le ! ») ou des mails de relance automatique (paramétrables par le participant).

Nous allons, en parallèle, tester l'utilisation de la plateforme sur au moins une autre langue, par exemple le créole guadeloupéen. Par ailleurs, le code de la plateforme, librement disponible sur `GitHub`[34] sous licence CeCILL v2.1[35] peut être adapté en un temps réduit à toute autre langue pour laquelle il existe au minimum : i) un corpus libre de droit dans la langue, ii) un guide d'annotation pour la tâche d'annotation en partie du discours, iii) une référence annotée minimale, et, si possible, iv) un *tagger* pour une langue proche.

Remerciements

Nous remercions vivement les participants de `Bisame` pour leur implication, encouragements et commentaires, ainsi que Nicolas Lefèbvre (Inria Nancy Grand Est) et Émilie Paillous, pour leurs conseils et leur aide quant au développement de la plate-forme. Nous remercions également Clément Dorffer pour avoir participé à l'annotation. La création de la référence et du lexique utilisés dans cet article ont bénéficié du soutien partiel de l'ANR (projet RESTAURE - référence ANR-14-CE24-0003).

References

ADOLF P. (2006). *Dictionnaire comparatif multilingue: français-allemand-alsacien-anglais*. Strasbourg, France: Midgard.

AGIĆ Ž., JOHANNSEN A., PLANK B., MARTÍNEZ H. A., SCHLUTER N. & SØGAARD A. (2016). Multilingual projection for parsing truly low-resource languages. *Transactions of the Association for Computational Linguistics*, **4**, 301–312.

BARRE C. & VANDERSCHELDEN M. (2004). *L'enquête "étude de l'histoire familiale" de 1999 - Résultats détaillés*. Paris: INSEE.

[34] Voir : `https://github.com/alicemillour/Bisame`.
[35] Voir : `http://www.cecill.info/`.

BERMENT V. (2004). *Méthodes pour informatiser les langues et les groupes de langues "peu dotées"*. Thèse, Université Joseph-Fourier - Grenoble I.

BERNHARD D., ERHART P., HUCK D. & STEIBLÉ L. (2016). *Guide d'annotation morphosyntaxique pour les dialectes alsaciens*. Guide d'annotation, LiLPa, Université de Strasbourg.

BERNHARD D. & LIGOZAT A.-L. (2013). Es esch fàscht wie Ditsch, oder net? étiquetage morphosyntaxique de l'alsacien en passant par l'allemand. In *Actes de TALARE 2013 : Traitement Automatique des Langues Régionales de France et d'Europe*, p. 209–220, Les Sables d'Olonne, France.

BÖHMOVÁ A., HAJIČ J., HAJIČOVÁ E. & HLADKÁ B. (2001). The prague dependency treebank: Three-level annotation scenario. In A. ABEILLÉ, Ed., *Treebanks: Building and Using Syntactically Annotated Corpora*. Kluwer Academic Publishers.

CHAMBERLAIN J., FORT K., KRUSCHWITZ U., LAFOURCADE M. & POESIO M. (2013). 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, p. 3–44. Springer Berlin Heidelberg.

COSQUER A., RAYMOND R. & PREVOT-JULLIARD A.-C. (2012). Observations of everyday biodiversity: A new perspective for conservation? *Ecology and Society*, **17**.

DENIS P. & SAGOT B. (2010). Exploitation d'une ressource lexicale pour la construction d'un étiqueteur morphosyntaxique état-de-l'art du français. In *Traitement Automatique des Langues Naturelles : TALN 2010*, Montréal, Canada.

FENOUILLET F., KAPLAN J. & YENNEK N. (2009). Serious games et motivation. In *4ème Conférence francophone sur les Environnements Informatiques pour l'Apprentissage Humain (EIAH'09), vol. Actes de l'Atelier "Jeux Sérieux: conception et usages"*, p. 41–52, Le Mans, France.

FORT K. (2016). *Collaborative Annotation for Reliable Natural Language Processing*. Focus series. ISTE Wiley.

FORT K., NAZARENKO A. & ROSSET S. (2012). Modeling the complexity of manual annotation tasks: a grid of analysis. In *Actes de International Conference on Computational Linguistics (COLING)*, p. 895–910, Mumbaï, Inde.

FORT K. & SAGOT B. (2010). Influence of pre-annotation on POS-tagged corpus development. In *Actes du ACL Linguistic Annotation Workshop*, p. 56–63, Uppsala, Suède.

GUILLAUME B., FORT K. & LEFEBVRE N. (2016). Crowdsourcing complex language resources: Playing to annotate dependency syntax. In *Actes de International Conference on Computational Linguistics (COLING)*, Osaka, Japon.

HOLLENSTEIN N. & AEPLI N. (2014). Compilation of a swiss german dialect corpus and its application to pos tagging. In *Proceedings of the First Workshop on Applying NLP Tools to Similar Languages, Varieties and Dialects*, p. 85–94, Dublin, Irlande.

HOVY D., PLANK B. & SØGAARD A. (2014). Experiments with crowdsourced re-annotation of a pos tagging data set. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, p. 377–382, Baltimore, USA.

LAFOURCADE M. & JOUBERT A. (2008). JeuxDeMots : un prototype ludique pour l'émergence de relations entre termes. In *Journées internationales d'Analyse statistique des Données Textuelles (JADT)*, Lyon, France.

LAFOURCADE M., LEBRUN N. & JOUBERT A. (2015). *Jeux et intelligence collective: résolution de problèmes et acquisition de données sur le web*. Collection science cognitive et management des connaissances. ISTE.

LI S., GRAÇA J. A. V. & TASKAR B. (2012). Wiki-ly supervised part-of-speech tagging. In *Actes de the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, p. 1389–1398, Stroudsburg, USA.

LIBERMAN M. (2016). Oral histories: Linguistic documentation as social media. In *Actes de NIEUW: Novel Incentives and Engineering Unique Workflows*, Philadelphie, USA.

MCENERY T. & HARDIE A. (2011). *Corpus Linguistics: Method, Theory and Practice*. Cambridge Textbooks in Linguistics. Cambridge University Press.

MUNRO R. (2013). Crowdsourcing and the crisis-affected community: lessons learned and looking forward from mission 4636. *Journal of Information Retrieval*, **16**(2).

PESTIAN J. P., MATYKIEWICZ P. & LINN-GUST M. (2012). What's in a note: Construction of a suicide note corpus. *Biomedical Informatics Insights*, **5**, 1–6.

PETROV S., DAS D. & MCDONALD R. (2012). A universal part-of-speech tagset. In *Actes de Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turquie: European Language Resources Association (ELRA).

PIMIENTA D. & PRADO D. (2014). *Etude sur la place des langues de France sur l'Internet*. Langues & recherche. DGLFLF/Maaya.

SAGOT B., FORT K., ADDA G., MARIANI J. & LANG B. (2011). Un turc mécanique pour les ressources linguistiques : critique de la myriadisation du travail parcellisé. In *Actes de Traitement Automatique des Langues Naturelles (TALN)*, Montpellier, France. 12 pages.

SCHERRER Y. & SAGOT B. (2013). Lexicon induction and part-of-speech tagging of non-resourced languages without any bilingual resources. In *RANLP Workshop on Adaptation of language resources and tools for closely related languages and language variants*, Hissar, Bulgarie.

SCHILLER A., TEUFEL S. & THIELEN C. (1995). Guidelines für das tagging deutscher textcorpora mit stts. *Universitäten Stuttgart und Tübingen*.

SCHMID H. (1997). *New Methods in Language Processing, Studies in Computational Linguistics*, chapter Probabilistic part-of-speech tagging using decision trees, p. 154–164. UCL Press, Londres.

TOUTANOVA K., KLEIN D., MANNING C. D. & SINGER Y. (2003). Feature-rich part-of-speech tagging with a cyclic dependency network. In *Actes de Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, p. 173–180, Stroudsburg, USA.

ZENNAKI O., SEMMAR N. & BESACIER L. (2016). Inducing Multilingual Text Analysis Tools Using Bidirectional Recurrent Neural Networks. In *Actes de International Conference on Computational Linguistics (COLING)*, Osaka, Japon.