REVUE ILES D IMESLI, VOLUME 14, N°1 (2022), PP. 67-81

Common Voice Kabyle : brève présentation du projet suivie d'une enquête sur les contributeurs kabyles qui y travaillent

Common Voice Kabyle: brief overview of the project followed by a survey about the Kabyle contributors who are working on it

Youva Djouadi
Université Abderrahmane Mira de Bejaïa, Algérie / CY Cergy Université, France
youva.djouadi@univ-bejaia.dz

Article information

History of the article- Historique de l'article		
Received: 27/01/2022	Accepted : 19/07/2022	Published: 31/12/2022

Abstract
This article provides an overview of the Kabyle version of the *Common Voice* project, launched by the *Mozilla Foundation*, including a study to sample the Kabyle contributors who are working on it. Before that, we briefly listed the different works related to the automatic processing of the Amazigh language.

Keywords: Digital corpus; Kabyle language; Common Voice; Automatic speech recognition,;Automatic translation.

Résumé
Cet article propose une vue d'ensemble sur la version kabylisée du projet *Common Voice*, lancé par la *Fondation Mozilla*, ainsi qu'une enquête visant à échantillonner les contributeurs kabyles qui y travaillent. Avant cela, il a été question d'inventorier brièvement les différents travaux portant sur le traitement automatique de la langue amazighe.

Mots-clés : Corpus numérique ; Langue kabyle ; Common Voice ; Reconnaissance automatique de la parole ; Traduction automatique.

Agzul
Amagrad-a yefka-d tamuɣli tamatut ɣef ulqem aqbayli n usenfar n *Common Voice*, i d-tesnulfa *Mozilla Foundation*. Yefka-d daɣen yiwet n tsastant i d-yewwin ɣef wid d tid i yettikkin deg-s. Send anect-a, yezzi wawal deg-s ɣef tezrawin i d-yewwin ɣef taɣult n usekker awurman n tutlayt tamaziɣt.

Awalen-tisura : Asagem umḍin; Taqbaylit; Common Voice; Aɛqal awurman n tmeslayt; Tasuqqilt tawurmant.

Auteur correspondant : Youva Djouadi, youva.djouadi@univ-bejaia.dz
ISSN: 2170-113X, E-ISSN: 2602-6449, (cc) BY-SA
Published by: Mouloud Mammeri University of Tizi-Ouzou, Algeria

Introduction

La langue kabyle – glossonyme utilisé par les initiateurs du projet que nous présenterons dans cet article –, à l'instar des langues peu dotées en ressources numériques, se retrouve dans l'obligation de s'accommoder et de recourir aux différents outils destinés au traitement automatique du langage naturel (TALN) qu'offrent les industries de la langue (Fuchs et al. 1993 ; 13). En effet, face à l'ampleur que prennent les nouvelles technologies dans la gestion du quotidien humain, dans nos sociétés d'aujourd'hui, vouloir initier la promotion d'une langue sans une stratégie incluant les outils numériques reviendrait à faire de l'utopisme linguistique.

Toutefois, cette dernière décennie est marquée par une dynamique prometteuse qui compte quelques dizaines de travaux de recherche portant sur le traitement automatique de la langue amazighe standard du Maroc et de la langue amazighe (variété kabyle), en Algérie. C'est le cas, d'une part, de l'Institut Royal de la Culture Amazighe (IRCAM) qui a élaboré plusieurs travaux portant sur divers axes du TALN comme le recensent Ataa Allah et Miftah (2018 : 6) : « le codage des caractères ; la reconnaissance optique des caractères ; les concordanciers et les convertisseurs d'alphabets ; création de corpus annotés ; traitement automatique de la parole. » En effet, il suffit de se rendre sur le site officiel de l'IRCAM[1] pour constater les différents outillages, liés au traitement automatique de la langue amazighe, rendus publics et utilisables gratuitement tantôt par les chercheurs, tantôt par les locuteurs de cette langue. Ainsi, on y retrouve, sur le site internet en question, des appareillages numériques dont la liste, bien que non-exhaustive, est la suivante : un transcodeur ANSI – Unicode qui permet de changer le schéma de codage de texte ; un translittérateur arabe-latin-tifinagh qui sert à convertir des écrits dans les trois graphies usitées en tamazight ; etc. D'autre part, et c'est valable aujourd'hui encore, les travaux initiés dans ce sens au niveau des institutions algériennes sont de moindre importance et ne foisonnent point. Néanmoins, on y recense quelques travaux à l'instar de ceux qui ont été le résultat d'une collaboration entre deux chercheurs du DLCA de Béjaïa avec un ingénieur en informatique en vue de réaliser un correcteur orthographique de la langue kabyle pour le cas du travail fait par Kamal Bouamara et Paul Anderson[2], et la modélisation numérique d'un dictionnaire regroupant Abdelaziz Berkaï et le même ingénieur informatique (Berkaï et Anderson, 2017 : 213-231). Aussi, et pour ne citer que cela, on y retrouve des travaux réalisés avec le logiciel NooJ comme ceux de Hamid Annouz, Kaci Ferroudja et Kamal Naït-Zerrad (2013 : 341-349) ; ceux de Farida Aoughlis (2012 : 229-244) et ceux de Farida Yamouni (2016 : 27-37). En partant de ce constat, la majorité des travaux initiés en faveur de l'outillage de la langue amazighe, en Algérie, est une conception *in vitro* destinée à répondre aux besoins des linguistes. Hormis les travaux de Kamal Bouamara et Paul Anderson, qui ont conceptualisé un correcteur d'orthographe dont la disponibilité est publique et gratuite sur Internet, il n'existe aucun autre outil numérique élaboré à cette fin par les institutions

[1] https://tal.ircam.ma/talam/ref.php (consulté le 13/03/2021)

[2] Site de l'ingénieur informatique cité http://www.akufi.org/en/index.html (consulté le 13/03/2021)

algériennes chargée de la langue amazighe. En somme, contrairement aux initiatives prises en faveur de l'outillage numérique de la langue amazighe par les institutions marocaines, ce champ d'étude reste à encourager du côté de l'Algérie. Par ailleurs, ce manque de production, qui se traduit par l'absence d'une collaboration académique affichée et soutenue entre linguistes et informaticiens, en Algérie, a poussé certains acteurs volontaristes et locuteurs soucieux du statut du kabyle à initier des résolutions en faveur de sa numérisation.

C'est dans ce sens que nous avons souhaité, à travers ce présent article, décrire un projet, baptisé par des locuteurs kabylophones, de numérisation de grande envergure au profit de la langue kabyle : le projet *Common Voice* (CV) de la fondation américaine *Mozilla*. Ce dernier est sous forme d'un grand corpus numérique conçu pour la langue kabyle dans toutes ses variétés linguistiques. Nous mettrons donc en avant cette source de données, largement investie par des locuteurs kabyles, qui fait office d'un large corpus numérique. Ainsi, pour ne pas résumé son fonctionnement en une phrase, le concept du projet CV consiste en l'ajout manuel d'un corpus aligné avec la possibilité de l'oraliser en toute synchronisation avec le texte. Dans ce travail, nous nous intéresserons au fonctionnement de l'interface CV, le rôle joué par les contributeurs, ainsi que leurs profiles, et, d'une manière succincte, les perspectives qu'offre un tel corpus quant à son exploitation dans les domaines du traitement du signal qu'il soit graphique ou oral.

1. Le projet *Common Voice*[3] de *Mozilla Foundation*[4]

Dans cette première section, il s'agira d'un bref exposé portant sur la genèse du projet *Common Voice* dans sa version anglaise et dans sa version kabyle. En outre, nous essayerons de présenter les raisons qui ont conduit à son lancement, et à son ouverture pour d'autres langues, par les équipes de *Mozilla* et les différentes conditions fixées par celles-ci en vue de structurer les demandes d'ajout de langues.

1.1. Brève présentation du projet CV

Avant de s'immiscer dans la présentation du projet CV dans sa version kabyle, il conviendrait d'évoquer, par le biais d'un bref survol, quelques détails essentiels concernant la *Fondation Mozilla*. D'abord, la *Mozilla Foundation* est un organisme à but non lucratif créé au début des années 2000. Ses initiateurs se présentent comme étant une communauté de militants d'un Internet libre et ouvert, une communauté soucieuse d'un monde numérique sain ou le respect de la vie privée et la diversité des usagers est primordial. En effet, à lire le manifeste[5] de l'entreprise, qui est

[3] Site officiel du projet *Common Voice* : https://commonvoice.mozilla.org/ (consulté le 29/01/2021)

[4] Site officiel de la *Mozilla Foundation* : https://foundation.mozilla.org/ (consulté le 29/01/2021)

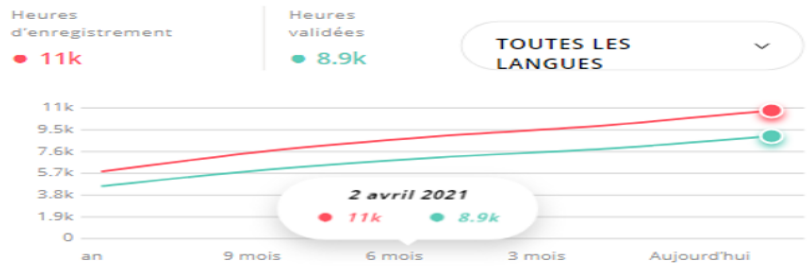
[5] Lien vers le Manifeste Mozilla : https://www.mozilla.org/fr/about/manifesto/ (consulté le 29/01/2021)

d'ailleurs présenté comme étant le guide sur lequel se basent toutes les activités de la société, on y retrouve des engagements très explicites et pour ne pas citer qu'un parmi les autres : « Nous nous engageons en faveur d'un Internet inclusif pour tous les peuples de la planète : un environnement dans lequel les caractéristiques démographiques ne nuisent pas à l'accès, aux possibilités offertes ou à la qualité de l'expérience en ligne. » Ensuite, il conviendrait de rappeler également que la *Mozilla Foundation* travaille sur la conception de logiciels et d'outils informatiques, gratuits et en libre accès, destinés au marché des technologies modernes par le biais de sa deuxième organisation, à savoir : la *Mozilla Corporation*. Enfin, de par ce qui a été dit ci-dessus, la présence du kabyle sur les différents produits développés par *Mozilla*, à l'instar de *Common Voice* que nous présenterons ci-dessous, confirme les thèses avancées par les équipes de *Mozilla* quant à l'application du principe d'équité linguistique entre toutes les langues parlées dans le monde. En d'autres termes, une langue minorée et de moindre diffusion comme le kabyle se voit ouvrir les portes du numérique sans passer par la logique des subventions étatiques et/ou du principe de la demande du marché.

Par ailleurs, le lancement officiel du projet *Common Voice* par les équipes *Mozilla* remonte à l'année 2017 et il a été basé sur le principe de la production participative (*Crowdsourciung*). Ce principe, encourage donc le travail collaboratif et vise à faire travailler le plus grand nombre de locuteurs de chaque langue donnée. Ainsi, l'enjeu premier de l'initiative a été d'élaborer une base de données - de corpus écrits oralisés - libre et ouverte destinée exclusivement à la langue anglaise. Or, la *Mozilla Foundation* n'a pas tardé à élargir son projet à d'autres langues qui sollicitent une intégration comme le signale Ardila et al. (2019 : 1-5) dans un article collectif publié sur le projet CV en avançant que « Le projet a été lancé en se concentrant initialement sur l'anglais en juillet 2017, puis en juin 2018, il a été rendu disponible pour toutes les langues. »[6]

En effet, des langues peu dotées, à l'image de la langue « Hakha Chin », ont pris part au projet et ont commencé à établir des grands corpus informatisés (Berkson et al, 2019). En somme, il y'a au total 63 langues – ce chiffre n'est pas figé – qui bénéficient d'une production foisonnante de corpus écrits oralisés sur *Common Voice*. D'ailleurs, actuellement, le volume de corpus enregistré dépasse les 11.000 heures pour 63 langues à travers le monde (voir figure 1)[7].

Figure 1. Volume total des corpus oralisés sur Common Voice



[6] Texte en langue-source : « The project was started with an initial focus on English in July 2017 and then in June 2018 was made available for any language. »

[7] Graphique récupéré sur : https://commonvoice.mozilla.org/fr/datasets (avril 2021)

En ce qui concerne le traitement des requêtes d'ajout de langues et leur intégration au projet, *Mozilla* a confectionné quelques recommandations sous forme de conditions bien définies façonnées dans un livret[8] publié sur leur site officiel. En effet, les locuteurs-contributeurs, qui souhaitent ajouter leurs langues, doivent satisfaire les conditions conçues par *Mozilla* afin d'intégrer le projet participatif. Ainsi, avant d'avoir un accès au projet, les futurs contributeurs sont astreints de s'organiser en communauté afin d'être en mesure de soumettre une requête d'ajout de leur langue. Par la suite, les équipes CV, chargées du traitement des demandes, procèdent à l'étude du dossier en interne et remettent une réponse sous peu. Si cette dernière est favorable, une structuration tripartite autour de la langue demandée est exigée pour la communauté demandeuse. De ce fait, les membres de la communauté linguistique doivent octroyer des rôles à chaque participant au projet, à savoir : les localisateurs (contributeurs) ; les examinateurs des traductions (phrases) ; les organisateurs du projet global (administrateurs).

Après la phase d'organisation et de gestion vient donc le premier travail sur la langue. En outre, les contributeurs sont invités à accomplir les deux dernières tâches avant d'avoir accès à toutes les fonctionnalités du projet, c'est-à-dire : traduire l'interface CV de l'anglais vers leur langue et implanter un premier corpus de 5000 phrases. En bref, pour récapituler, l'écosystème CV est organisé principalement de la façon suivante :

1. Faire une requête d'ajout de langue auprès des équipes officielles CV ;
2. Localiser et traduire l'interface CV de l'anglais vers la langue cible demandée ;
3. Insérer un corpus libre de droit de 5000 phrases via la plateforme *Sentence Collector*[9] ;
4. Oraliser le corpus écrit qui défile sur l'interface CV ;
5. Générer un jeu de données (Dataset) et entrainer les machines à la reconnaissance automatique de la parole.

1.2. Présentation de l'interface CV attribuée à la langue kabyle

L'ajout de la langue kabyle au sein du projet CV est l'initiative d'un groupe volontariste kabylophone créé par Mohand Oubelkacem[10]. Il regroupe des informaticiens, des enseignants, des étudiants et des militants kabyles de tout bord. En outre, la soumission du dossier kabyle, auprès des équipes *Mozilla*, a été précédée par une large diffusion de l'idée sur les réseaux sociaux sous forme d'appels à participer à la traduction de l'interface CV. Conséquemment, cette action a suscité un véritable engouement parmi la

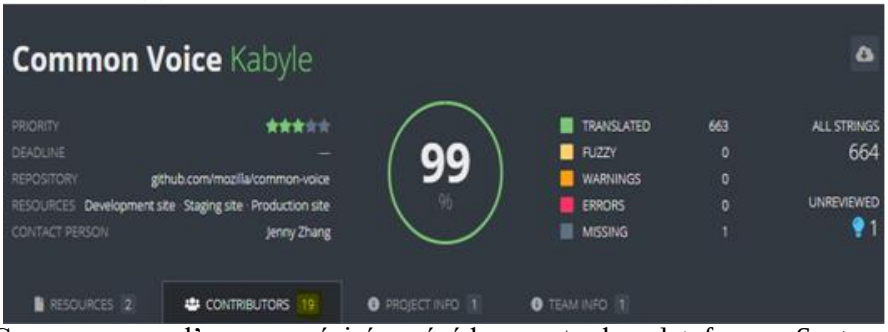
[8] Livret des conditions de participation au projet CV : https://common-voice.github.io/community-playbook/ (consulté le 29/01/2021)

[9] Plateforme *Sentence Collector* : https://commonvoice.mozilla.org/sentence-collector/ (consultée le 29/01/2021)

[10] Est un développeur Web et chef de projet, consultant dans plusieurs sociétés informatiques, consultant technico-fonctionnel en APPLICATIONS IT.

communauté kabyle. En tout, 19 contributeurs ont collaboré pour traduire 663 entrées de l'anglais vers le kabyle (voir figure 2[11]).

Figure 2. Plateforme de traduction Pontoon du projet CV



Comme nous l'avons précisé précédemment, la plateforme *Sentence Collector* (SC) est l'outil qui permet l'implantation du corpus écrit sur le projet CV. On y retrouve une liste de recommandations à respecter comme : la ponctuation, la longueur maximale des phrases, etc. De surcroît, la plateforme SC contient un menu qu'on retrouve au niveau de la page principale du site, sous forme d'une barre latérale fixée à gauche, qui permet aux utilisateurs de réaliser un bon nombre de requêtes comme il est montré en figure 3[12].

1. Se connecter à l'aide de ses coordonnées de connexion en cliquant sur *Login* ;
2. Choisir sa langue en cliquant sur *Profile* ;
3. Lire les recommandations et les règles, qui portent sur la transcription etc., en cliquant sur *How-to* ;
4. Ajouter des phrases en cliquant sur *Add ;*
5. Vérifier les phrases ajoutées par d'autres contributeurs afin de les valider ou les rejeter en cliquant sur *Review* ;
6. Accéder aux phrases, ajoutées par le biais du compte de l'utilisateur, rejetées par d'autres contributeurs en cliquant sur *Rejected Sentences* ;
7. Consulter les statistiques globales qui concernent chaque langue en cliquant sur *Statistics*.

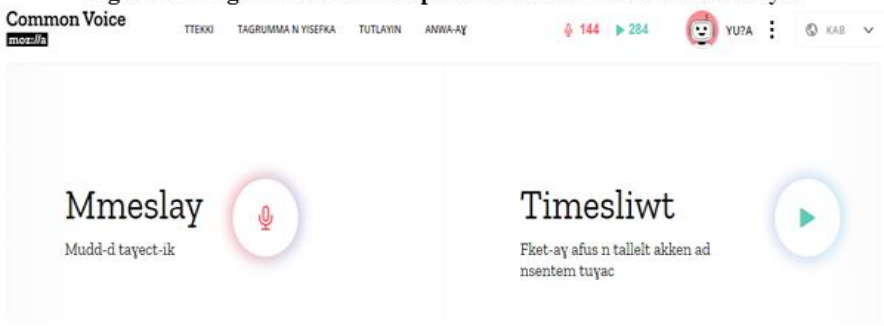
Figure 3. Menu principal de la plateforme Sentence Collector

[11] Plateforme de traduction *Pontoon* du projet CV : https://pontoon.mozilla.org/kab/common-voice/project-info/ (consultée le 29/01/2021)

[12] Figure récupérée sur https://commonvoice.mozilla.org/sentence-collector/#/en

Une fois le corpus intégré sur *Sentence Collector* est validé, il est transféré directement et automatiquement vers l'interface *Common Voice*. En outre, c'est dans cette dernière que les contributeurs font la plus grande partie du travail, c'est-à-dire : enregistrer leurs voix sur le corpus écrit de façon à l'oraliser en parfaite synchronisation son/texte. En somme, la page d'accueil de CV a été façonnée d'une manière à choisir entre deux options principales : *Enregistrer/parler (Mmeslay)* et/ou *Ecouter (Timesliwt)* les enregistrements des autres contributeurs en vue de les valider ou de les refuser (voir figure 4).

Figure 4 : Page d'accueil de la plateforme Common Voice Kabyle



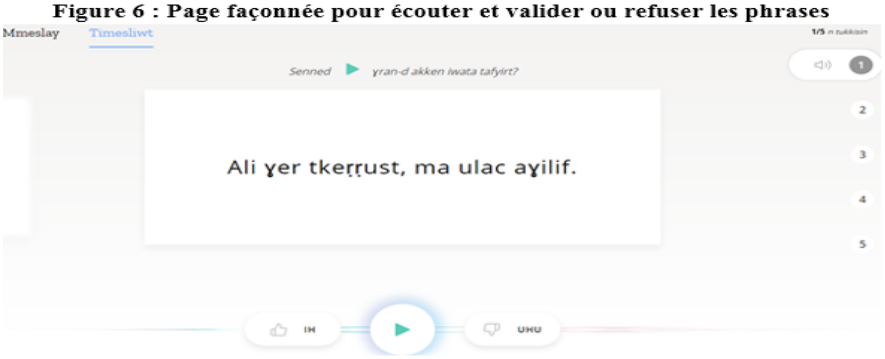
Par ailleurs, en choisissant l'option *Enregistrer*, le contributeur est basculé vers une nouvelle page conçue pour l'enregistrement à l'aide d'un microphone. Une phrase est alors affichée au milieu de l'interface pour être lue et enregistrée en cliquant sur l'icône fixée en bas de la page (voir figure 5). Il faut savoir que la liste numérotée qui se trouve à droite de la figure 5 consiste à aider le contributeur dans ses enregistrements – l'enregistrement se fait par 5 phrases –. Ainsi, en cas de problèmes techniques rencontrés, l'utilisateur peut refaire sa tâche tant qu'il ne quitte pas la page ouverte.

Figure 5 : Page façonnée pour l'oralisation du corpus

En optant pour l'option *Ecouter*, l'utilisateur est basculé vers une page qui diffuse les échantillons de voix synchronisés avec le texte afin d'endosser le rôle d'un examinateur des énoncés enregistrés par les autres contributeurs. Il faut savoir que la phrase en question est, ici aussi, affichée au centre de l'interface. Ainsi, le contributeur dispose d'un pictogramme,

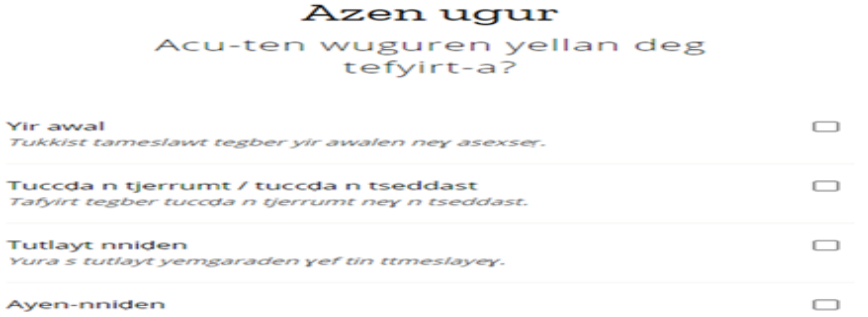
fixé en bas de la page, pour lancer l'échantillon de voix à l'aide d'un clic. Ce dernier se trouve au milieu de deux autres icônes qui permettent de valider ou de rejeter le corpus oralisé par un autre contributeur, l'utilisateur l'actionne en cliquant sur le bouton *Oui* (Ih) ou *Non* (Uhu) (voir figure 6).

Figure 6 : Page façonnée pour écouter et valider ou refuser les phrases



Il est à signaler également que l'interface CV dispose d'un système qui permet aux contributeurs de signaler d'éventuels manquements qui pourraient être jugés comme inadaptés à la politique d'utilisation érigée par *Mozilla*. Ceci dit, ces manquements vont de la simple faute d'orthographe aux propos haineux ou insultants. De ce fait, un bouton est disponible pour effectuer des signalements et des rapports en bas à gauche de la page afin d'alerter les administrateurs du projet (voir figure 7).

Figure 7 : Menu réservé aux différents signalements



2. Esquisse d'une étude qualitative sur les contributeurs kabyles sur CV

Dans cette deuxième section, il sera question de présenter les résultats d'une enquête réalisée sur le volet qualitatif des contributeurs kabylophones qui travaillent sur la version kabyle de *Common Voice*. Cela dit, nous présenterons, à travers la première sous-section (voir la sous-section 2.1), la méthodologie appliquée en vue de réaliser notre investigation en reproduisant les questions posées aux 36 contributeurs qui contribuent sur

CV. Ensuite, nous exposerons, dans la seconde sous-section (voir la sous-section 2.2), les résultats de notre enquête en les illustrant à l'aide de 5 figures qui représentent les réponses avancées par les enquêtés. Enfin, nous proposerons, dans le dernier sous-titre de cette deuxième section (voir la sous-section 2.3), une interprétation des résultats obtenus.

2.1. Méthodologie de l'enquête

Autre que la présentation technique du fonctionnement du projet CV appliqué à la langue kabyle, nous avons souhaité orienter notre intérêt vers le profil des contributeurs kabyles qui ont participé – et qui participent toujours d'ailleurs – au projet d'oralisation du corpus écrit implémenté sur CV. Pour ce faire, nous avons publié un questionnaire – un *Google Doc* que nous reproduirons ci-dessous – dans un groupe Facebook[13] destiné aux locuteurs qui coopèrent et qui travaillent sur les questions liées à la localisation de contenus vers la langue kabyle, la numérisation de cette dernière, etc. Ainsi, les principaux axes des questions choisies pour élaborer notre questionnaire sont les suivants :

1. Le sexe ;
2. La catégorie d'âge ;
3. Le niveau d'instruction ;
4. Le groupe linguistique de naissance ;
5. Les particularités phonétiques dans le processus d'oralisation du corpus.

Pour ce qui est des questions portant sur les axes 1, 2 et 3, nous avons posé des questions à choix multiples afin d'en savoir plus sur leurs sexes, leurs âges et leur niveau d'instruction. Les questionnés ont eu donc le choix de réponse entre les propositions suivantes :

1. Tuzzuft :
- Awtem ☐
- Tawtemt ☐
- War tuzzuft ☐

2. Tameddurt-ik/im (Acḥal i teseiḍ deg leɛmeṛ) :
- -18 ☐
- 19-29 ☐
- 30-50 ☐
- 51- ☐

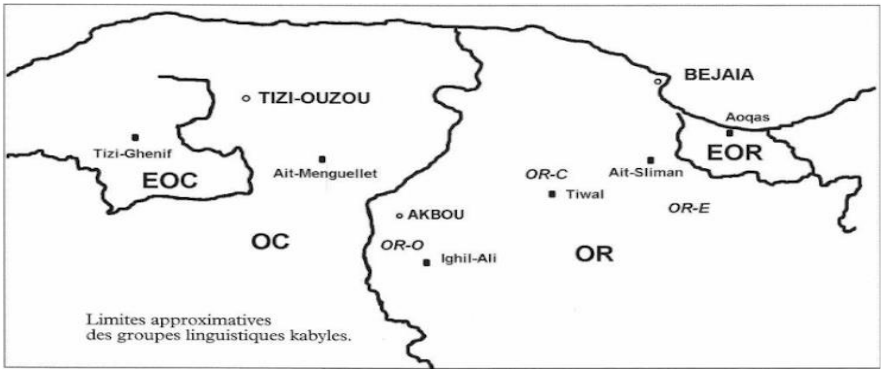
3. Aswir ussnan :

- Aselmed adday 1 ☐
- Aselmed adday 2 ☐
- Aselmed adday 3 ☐
- Aselmed unnig ☐

[13] Groupe Facebook regroupant les localisateurs kabyles « Imsidag Iqbayliyen » https://www.facebook.com/groups/165557820784383

Du côté des deux derniers axes de nos questions, nous avons opté pour les questions à choix multiples également. En l'occurrence, quatre réponses possibles pour la question liée aux groupes linguistiques kabyles – une carte a été jointe à la question en guise d'illustration – tels qu'ils ont été décrits et classés par Nait-Zerrad (2004 : 2). Pour finir, deux réponses possibles pour la question liée aux particularismes phonétiques usités dans le processus d'oralisation du corpus écrit :

1. Wali takarḍa-ya, tiniḍ-d anita i d taqbaylit-ik/im ?



- Taqbaylit n umalu imṭeṛṛef (EOC) ☐
- Taqbaylit n usammar (OR) ☐
- Taqbaylit n umalu (OC) ☐
- Taqbaylit n usammar imṭeṛṛef (EOR) ☐

2. S wanita taqbaylit i tesseklaseḍ deg CV ?

- S teqbaylit n temnaḍt-iw (S ususru n yal ass) ☐
- S teqbaylit n uɣerbaz (S ususru n yal imeslic) ☐

En somme, la portée de notre enquête a été assez large de par le nombre de participants qui ont accepté de répondre au questionnaire. Ainsi, nous avons pu atteindre 36 contributeurs, qui sont d'ailleurs les plus actifs et les plus investis dans le projet, pour sortir avec un échantillonnage assez représentatif.

2.2. Résultats de l'enquête

Le dénouement de notre investigation nous offre des résultats assez déséquilibrés quant aux questions liées aux axes 1, 2 et 3 de notre questionnaire, tantôt sur le genre et l'âge des contributeurs et contributrices, tantôt sur leur niveau d'instruction.

En effet, la majorité des contributeurs sont de sexe masculin avec un taux de 69% et par conséquent, nous constatons que la participation féminine, dans l'oralisation du corpus sur CV, est assez faible avec un taux de seulement 31% (voir figure 8). Semblablement, la catégorie d'âge la plus

dévouée dans le projet est celle des 30-50 ans avec un taux 67%, tandis que les contributeurs âgés entre 19 et 29 ans arrivent en deuxième position avec un pourcentage qui avoisine les 25% (voir figure 9).

Figure 8. Répartition par sexe des contributeurs kabyles sur CV

Hommes Femmes

Figure 9. Catégorie d'âge des contributeurs kabyles sur CV

Moins de 18 ans Plus de 51 ans
De 19 à 29 ans De 30 à 50 ans

Aussi, comme pour corroborer le déséquilibre constaté, l'écrasante majorité des contributeurs sur CV ont atteint le degré de l'enseignement supérieur, et donc un niveau d'instruction étoffé, avec un taux de 75% sur les 36 questionnés (voir figure 10).

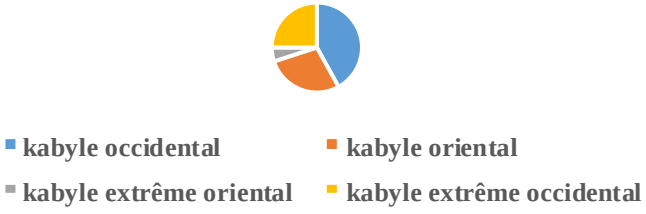
Figure 10. Répartition par niveau d'instruction des contributeurs kabyles sur CV

Enseignement supérieur Enseignement secondaire
Enseignement moyen Enseignement primaire

A l'opposé, les résultats obtenus pour les questions portant sur les groupes linguistiques d'appartenance et les modes phonétiques usitées

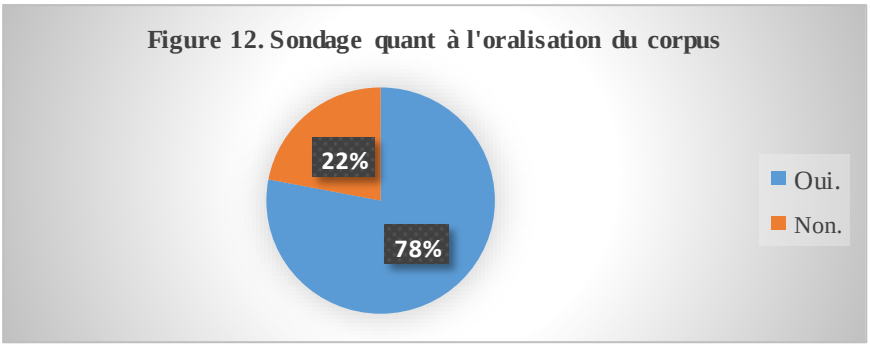
durant l'enregistrement du corpus font figure d'un certain équilibre. Ainsi, les groupes linguistiques auxquels appartiennent les contributeurs sont répartis d'une façon assez équilibrée entre 42% de locuteurs du kabyle occidental, 28% du kabyle oriental, 25% du kabyle extrême occidental et 5% de kabylophones appartenant aux parlers de l'extrême orient de la Kabylie (voir figure 11).

Figure 11. Répartition par groupes linguistiques représentés



Enfin, s'agissant de la question posée sur les particularismes phonétiques, 78% des contributeurs disent pratiquer leurs spécificités locales lorsqu'ils procèdent à l'enregistrement de leurs voix sur le corpus écrit et seulement 22% disent oraliser le corpus sans reproduire leurs spécificités locales (voir figure 12).

Figure 12. Sondage quant à l'oralisation du corpus



2.3. Interprétation des résultats

Pour ne pas se limiter à la présentation du fonctionnement de CV et ne pas s'aventurer précipitamment à analyser le corpus élaboré par la communauté kabyle sans étudier, en amont, le volet qualitatif des femmes et des hommes qui la compose, nous avons préféré commencer par une mise en avant des contributeurs qui numérisent le corpus kabyle, à travers notre enquête (voir les deux sous-sections précédentes), afin d'aboutir à un aperçu intuitif sur la qualité du corpus kabyle forgé sur CV. De ce fait, la finalité de cette démarche est de substantialiser le projet CV appliqué à la langue kabyle, dans les débats académiques amazighisants, d'une manière à permettre aux

éventuels linguistes ou informaticiens, qui s'y intéresseront, d'avoir un préambule sur le fonctionnement de CV ainsi qu'un échantillon sur le profil des locuteurs kabyles qui y travaillent.

En tenant compte des résultats de notre investigation, illustrés par les figures supra, le projet *Common Voice* appliqué à la langue kabyle nécessiterait une meilleure prise en charge ainsi qu'une plus grande diffusion auprès les locuteurs kabylophones afin de réguler le déséquilibre constaté en ce qui concerne notamment la sous-représentation du sexe féminin, l'absence totale de contributeurs pour la tranche d'âge des moins de 18 ans et la surreprésentation des contributeurs qui ont un niveau universitaire.

Premièrement, il faudrait qu'il y'ait un organisme de recherche spécialisé en langue amazighe qui s'investisse dans le projet afin de mieux gérer le corpus écrit en veillant à diversifier le type de phrases à incorporer. Ensuite, inciter un plus grand nombre de locuteurs kabyles à prendre part au projet d'oralisation du corpus écrit serait une stratégie qui permettrait d'assurer une certaine représentativité tantôt sur le plan linguistique – comme le phénomène de la variation linguistique –, tantôt sur le volet social – l'équilibre des genres ; des classes sociales et des catégories d'âges –. Enfin, les résultats de notre enquête démontrent l'existence d'un point de recueil de corpus, qui fait figure d'une certaine représentativité, d'un genre nouveau pour la recherche linguistique amazighe.

3. Enjeux, perspectives et conclusion

La présence du kabyle dans le projet *Common Voice* est en elle-même un défi relevé. En effet, cette langue a toujours été reléguée au rang de langue orale sans un véritable statut de langue pratique et pratiquée dans les domaines formels et encore moins dans les nouveaux secteurs naissants liés aux nouvelles innovations technologiques qui ne cessent de croitre de jour en jour. De surcroît, le corpus linguistique implémenté sur le projet CV kabyle ne cesse de s'étoffer sur le plan quantitatif et c'est justement sur ce volet qu'existe un enjeu de grande ampleur.

En effet, pour que le corpus généré sur CV devienne exploitable en vue de la reconnaissance automatique de la parole, il faudrait qu'il y'ait une quantité colossale de phrases implémentés et oralisés comme le signale Abenaoui dans un article collectif publié sur cette vaste thématique qui est le traitement du signal (cf. Abenaou et al., 2014). En d'autres termes, la mission des informaticiens réside dans la mise en place d'algorithmes adéquats, basés généralement sur des méthodes mathématiques et statistiques, afin d'entrainer les machines avec les données générées. Ces dernière sont, d'ailleurs, librement accessibles sur la plateforme CV et sont, conséquemment, exploitables gratuitement.

Par ailleurs, le corpus *Common Voice* Kabyle est sous forme d'un jeu de données (Dataset) qui, une fois extrait, donne deux dossiers principaux : un dossier contenant le corpus audio sous forme de fichiers audio sous format MP3 ; un autre dossier contenant le corpus écrit sous format TSV[14]. De ce fait, l'exploitation des fichiers donne un corpus aligné et nu qui est

[14] Tabulation-separated values : un type de fichiers ouvrables via Microsoft Office Excel notamment.

parfaitement adéquat pour la mise en place de la traduction automatique à en croire Habert et al. (1997 : 140) qui explique que « le recours aux textes alignés constitue par certains côtés une riposte aux limites rencontrées dans l'automatisation de la traduction automatique. ».

En somme, le projet CV est un outil de type nouveau qui pourrait jouer un rôle de réservoir linguistique pour la langue kabyle, tantôt sur le plan symbolique – en offrant un nouveau statut de langue écrite et numérique au kabyle –, tantôt sur le plan pratique – voir les données exposées infra. Ainsi, il en résulte, de ce projet, plusieurs facettes porteuses de perspectives si les données générées venaient à être exploitées convenablement.

Dans cette contribution, nous avons essayé de démontrer, au-delà de la présentation du projet, le volet qualitatif des contributeurs qui se donnent comme mission de numériser la langue kabyle dans toutes ses variétés. En bref, Le projet *Common Voice* est l'une des plus grandes sources de données électroniques jamais conçue pour la langue kabyle. Son caractère ouvert et libre de droit lui garantit une subsistance à long terme et une présence numérique actée. Nous récapitulons ci-dessous quelques données chiffrées que nous avons pu consulter durant l'année 2021[15] :

- Un corpus aligné et oralisé de plus de 350.000 phrases.
- Plus de 600 heures d'enregistrement (Phrases écrites oralisées).
- Un jeu de données (Dataset) d'une taille de 16 Go.

Il est important de rappeler que toutes les données produites sur *Common Voice* sont sous licence libre et donc exploitable librement. En plus du potentiel qu'elles offrent aux informaticiens dans le processus d'apprentissage pour la reconnaissance automatique de la parole et de la traduction automatique, nous prévoyons de notre côté d'analyser, un peu plus en profondeur, le contenu du corpus afin de critiquer sa représentativité et amorcer son annotation et sa segmentation automatique à l'instar du projet initié par Tigziri (2012 : 261-303).

Bibliographie

Abenaou, Abdenbi, Ataa Allah, Fadoua, Nsiri, Benayad, 2014, « Vers un système de reconnaissance automatique de la parole en amazighe basé sur les transformations orthogonales paramétrables », in *Asinag*, n°09, Rabat, pp. 133-145. https://tal.ircam.ma/talam/Articles/Abnaou-Asinag9.pdf

Annouz, Hamid, Kaci, Ferroudja, Naït Zerrad, Kamal, 2013, « Le logiciel Nooj appliqué au kabyle », in *Iles d Imesli*, n°05, pp. 341-349.

https://www.asjp.cerist.dz/en/article/45814

Aoughlis, Farida, 2012, « Vers un module Tamazight pour le système NooJ », in *Iles d Imesli*, n°04, pp. 229-244.

https://www.asjp.cerist.dz/en/article/45771

[15] Ces données sont régulièrement mises à jour et publiées sur le site de CV : https://commonvoice.mozilla.org/fr/datasets

Ardila, Rosana et al., 2019, « Common Voice : A Massively-Multilingual Speech Corpus », in *arXiv preprint arXiv*, n°*1912.06670*, pp. 1-5.

https://arxiv.org/pdf/1912.06670.pdf

Ataa Allah, Fadoua., Miftah, Nina, 2018, « The First Parallel Multi-lingual Corpus of Amazigh », In *Journal of Engineering Research and Application*, Vol. 8, Issue 6 (Part -V), pp 05-12.

https://journals.indexcopernicus.com/search/article?articleId=1811891

Berkai, Abdelaziz., Anderson, Paul, 2017, « Essai de dictionnaire tasaḥlit (parler kabyle d'Aokas) -français : conception lexicographique et modélisation informatique », in *Iles d Imesli*, n°09, pp. 213-231.

https://www.asjp.cerist.dz/en/downArticle/397/9/1/45018

Berkson, Kelly et al., 2019, « Building a Common Voice Corpus for Laiholh (Hakha Chin) », in *Proceedings of the 3rd Workshop on the Use of Computational Methods in the Study of Endangered Languages*, Vol 2, pp. 05-10. https://journals.colorado.edu/index.php/computel/article/view/443/425

Fuchs, Catherine et al., 1993, *Linguistique et traitements automatiques des langues*. Paris, Hachette Education.

Habert, Benoît et al., 1997, *Les linguistiques de corpus.* Paris, Armand Collin.

Naït Zerrad, Kamal, 2004, « Kabylie : Dialectologie », in *Encyclopédie berbère [En ligne]*, n°26, pp. 1-6.

http://journals.openedition.org/encyclopedieberbere/1433

Tigziri, Nora, (2012), « Corpus oraux : Essai de segmentation automatique », in Youssef Ait Ouguengay et Siham Boulaknadel, *Les ressources langagières : construction et exploitation : le 4ème atelier international sur les technologies d'information et de communication pour l'amazighe. Rabat, 24-25 février 2011*, pp. 261-303.

https://tal.ircam.ma/conference/docs/ticam2011/Corpus%20oraux%20Essai%20de%20segmentation.pdf

Yamouni, Farida, 2016, « Analyse syntaxique automatique en tamazight (kabyle) », in *Actes de conférences : Technologies d'Information et de Communication pour l'Amazighe*, pp. 27-37.

https://biblio.ircam.ma/opac_css/doc_num.php?explnum_id=18

Zhang, Jenny, 2020, End-of-Year Common Voice Dataset Release, rapport publié sur le site officiel de Mozilla Discourse.

https://discourse.mozilla.org/t/2020-end-of-year-common-voice-dataset-release/72287 (20/12/2020)