Pilot survey on Common Voice and the sociolinguistic perspectives of techno-vehicular languages

Grégoire ANDREO-RAYNAUD(1),
Alexandre LISSY(2), Francis TYERS(3)

(1) Dipralang EA739.
(2) Mozilla Corporation.
(3) Department of Linguistics, Indiana University.

Résumé : This paper describes the Common Voice project, a community-based language planning policy (LPP) in the area of techno-vehicular languages, that is those languages used in Human-Machine interaction. There has already been some research in the field of computational linguistics in this area, some of which includes insights from sociolinguistics on the revitalisation of endangered languages (Eisenlohr 2004 ; Soria, 2016 ; Jones, 2020.). Kaplan (2014, 2015) focused on the use of English as "pivot language" in automatic translation and its impact on language practices. Péréa (2016) worked on linguistic interactions with machines, and Larsonneur (2019) looked at the implications of the use of artificial intelligence in the translation industry. In this study, we focus on the sociolinguistic aspects of techno-vehicular languages. The initial findings of our pilot survey show that the empowerment of less-resourced language communities has a positive impact on language diversity in the domain of techno-vehicular languages. We also highlight that communities see contribution to Common Voice as a way to achieve different goals, many of which are related to discourses on language endangerment and political action. We also show that the collection of data used for techno-vehicular languages is sensitive to different sociolinguistic, ideological and economic biases that produce inequality in human-machine interaction through language.

Mots-clés : Language policy ; human-machine interaction ; language diversity ; techno-vehicular languages

1 Introduction

This paper aims to provide a preliminary approach to the field of "techno-vehicular languages", as we propose to call linguistic forms which are used in Human-Machine interaction (such as those involving voice assistants or other Internet-connected devices such as smartphones). This class of languages is created using two main technologies : automatic speech recognition (ASR) and natural language processing (NLP).

These technologies rely training statistical models on large collections of data, or *datasets,* made up of individual correspondences between voice recordings and written transcripts. These datasets may be either publically available for anyone to use, or proprietary, restricted to a single corporation. We aim to highlight the relevance of this matter for sociolinguistics and in particular in language planning policy (LPP) studies. The LPP of techno-vehicular languages is handled by a range of different public and private agents. Their characteristics differ according to how data is gathered, the financial model and the motivations. Private agents, mostly multinational technological corporations such as GAFAMs (Google, Amazon, Facebook, Apple, Microsoft),[1] collect data from users of their products without remuneration *via* conditions according to their terms of service. In addition, they occasionally hire individuals to work on creating proprietary datasets. Their financial model is based on extracting value from the product or the data they sell. Business is their main motivation and they have no explicit language policy strategy. Their choice to support a given language is influenced by its capacity to give access to a profitable market.[2] As a result of this *de facto* language policy, most voice assistants are developed to support only 24 languages when over 7,000 languages exist (according to Ethnologue). All of these are official languages of the richest countries. On the other side, public agencies are founded by governments, state institutions or language communities. They mostly create datasets by mobilising within language communities. They are driven by linguistic and cultural revival and their discourses often value the importance of preserving cultural diversity. Therefore, they work mostly on what have been termed *peripheral* languages at the smallest scale.

We choose the Common Voice project of Mozilla as a case study because it occupies a kind of middle ground. As we will see, it is a private project, but is also driven by individual language communities according to the principles of the open-source movement[3] (Carillon and Okoli, 2008). After the description of our methodological approach in the first section, the second section will be dedicated to the history, functioning, of the project and the profile and motivations of the actors. In the third section, we will focus on the sociolinguistic aspects of the project.

2 Methodological notes

2.1 Research questions

In this study we focus on three main questions : What is Common Voice ? Why are people contributing to it ? How are people contributing to it ? Then we discuss the implications of the Common Voice language policy.

[1] There are other companies that sell NLP and ASR technologies but they are much smaller :

[2] For instance, a metric such as "potential euros from advertising per capita" appears to correlate with the set of language supported companies whose profits are derived principally from advertising such as Google and Facebook.

[3] The open-source movement refers in this paper to the movement launched by R. Stallman in 1983 in the GNU manifesto. Its main idea is to give free access of source code of computer programs to everyone in behalf of freedom (for more details, see the Website of GNU : https://www.gnu.org/philosophy/free-software-even-more-important.en.html).

Published in : Andreo-Raynaud, Grégoire ; Douidi, Nadhir (dir.), *Langue(s) en mondialisation. Libre(s) échange(s) à l'heure néolibérale ?*, Editions des archives contemporaines, ISBN : 9782813004284, doi : 10.17184/eac.9782813004284

2.1.1 Data

Due to the *online* nature of Common Voice, and the COVID-19 crisis, our field work has mostly been carried out remotely. The first set of data we use are statistical information freely-available on the project's Website. Those information concern languages, contributors (number, self-identification of speakers in terms of gender, accent and age), recordings and their metadata, set of transcripts.[4] Thereafter, between February and October 2020 we conducted seven semi-directed interviews with members of the Italian, Kabyle, Welsh, Taiwanese, Votic and Occitan language communities. We also interviewed two Common Voice staff members who gave us some insight into the history of the project and contribution dynamics. Based on these interviews we launched a survey questionnaire for Common Voice contributors for which we received 54 responses. The questionnaire was advertised using the Common Voice forum and we acknowledge that this will entail some sampling bias including over-representation of technologists (30% of the sample), young people (60% of the sample is less than 40) and males (84% of the sample). The questionnaire was more likely to be filled by active contributors and less likely to be completed by less regular contributors. We completed these data with qualitative data coming from topic threads on the Common Voice discussion forum which gives a good insight of linguistic debates in the platform. It is worth noting that the Mozilla employees who participated in our study, were interested in getting a better understanding of community dynamics as this is an area that they have not actively investigated in the Common Voice project.

2.1.2 Research field

Our first connection with the institutional actors in the Common Voice project was provided by a co-author of this work, Alexandre Lissy, former engineer on the DeepSpeech team and one of the contributors to Common Voice in French. This relationship helped us reach other actors in the project. They did not support our study any further than answering our questions. Their interest in our work does not go beyond collecting data about the project.

3 The Common Voice project

3.1 History

The Mozilla Project is founded on the premise of defending the access to an Open Web, as laid out in its Manifesto. With the expansion of voice technology in the technological market, Mozilla had a need for a speech-to-text system that would align with both their commercial goals and their social ideals. They needed speech-to-text to be local[5] for privacy and efficiency reasons. The first version of their speech recognition system was shipped in November 2017, using English as a demonstration language. This was followed by the first release of a dataset at the end of November 2017. The positive reception of the project and multiple requests from other language communities paved the way for localisation. This was carried out in the way Mozilla

[4] This set of data includes only what people agree to share with Common Voice, so it is not complete.

[5] By "local" we mean software that runs on a particular person's computer as opposed in the *cloud*. The majority of offerings by major participants in this area are cloud-based.

Published in : Andreo-Raynaud, Grégoire ; Douidi, Nadhir (dir.), *Langue(s) en mondialisation. Libre(s) échange(s) à l'heure néolibérale ?*, Editions des archives contemporaines, ISBN : 9782813004284, doi : 10.17184/eac.9782813004284

usually conducts localisation : It was open and collaborative ; that is, it is not up to Mozilla to handle a new language, but to the contributors of that language to do so. The first multi-locale version was shipped with English, German, Welsh and French enabled in June 2018, and soon, many more languages were added. Several releases of the Common Voice dataset have been produced. As of the date of this study, the latest one was in June 2020 and features several thousands of hours of audio in 54 languages (5671 validated, 7226 recorded in total).

3.2 Why is it interesting to study ?

Common Voice is just one example of LPP in techno-vehicular languages among many others. It can be qualified as an "hybrid project" as it occupies an intermediary position that will help us understand the multiple aspects and layering of LPP (Horneberger & Ricento, 1996). It stands between private and public agents, as Mozilla is a private company with both commercial and social aims. It has both top-down and bottom-up characteristics as it is driven both by institutional staff and communities of speakers. Finally, the project provides a lot of data about ideological aspects as it involves many different persons sharing diverse ideological views, beliefs about language, and motivations.

3.3 How does it work ?

Common Voice is a crowd-sourcing project, aiming at collecting pairs of data consisting of audio clips and their matching text transcriptions. At a very basic level, it is a Website where people record themselves reading a sentence written on their screen. It is both driven by Mozilla employees on the infrastructure (Website, hosting) part, and by Mozilla contributors on the data collection part, with some employees also interacting depending on their own goals : personal involvement in a language, partnership with a university, or a specific project. It was built as an open project from the beginning, and every language is welcome as long as it has an official existence (ISO 639-3 code).

The process is rather straightforward : contributors first translate the interface, they then coordinate themselves to collect text data (i.e. sentences) matching the requirements of the project – both legal and technical ; once sufficient text has been collected and validated, it is then presented on the Website for people to both record themselves speaking, and verify other people's contributions.

3.4 The actors of Common Voice

Common Voice can be characterised as a bottom-up language policy according to the definition of Nekvapil and Sherman (2015, 2) :

> The "top-down" direction tends to be associated with actors who possess significant power, while the "bottom-up" direction is associated with actors who do not have such a degree of power.

Indeed, it involves non-governmental agents and language speakers working together. Even if this "bottom-up" categorisation hides other complex differences and disparity

Published in : Andreo-Raynaud, Grégoire ; Douidi, Nadhir (dir.), *Langue(s) en mondialisation. Libre(s) échange(s) à l'heure néolibérale ?*, Editions des archives contemporaines, ISBN : 9782813004284, doi : 10.17184/eac.9782813004284

between actors,[6] it points to the self-organisation of the project and its dependence on the individual efforts of volunteers from each language community. This project is organised according to basic rules set up by Mozilla which carries out a regulatory role. These rules concern each language's status as recognised or unrecognised, dialect management (accents are based on the region of origin of the speakers). Other issues are discussed among different contributor communities and regulated by a Mozilla moderator. Common Voice can be seen as an online "*community of practice*" in which participants mutually engage in joint enterprise (Wenger, 1998, 73).

3.5 Overview of languages on Common Voice

Between June 2017 and August 2020 Common Voice was able to mobilise 147,819 speakers in 60 languages (we only count languages that are open to recording). We can distinguish between the types of languages in this overview using the global language system of De Swaan (2001) and Calvet (1999). This leads to the following classification :

- 1 hyper-central language which is official in many central states and whose native speakers tend to be monolingual : English ;
- 9 super-central languages which are official in several central states and whose native speakers tend to be monolingual or bilingual in other super-central and hyper-central languages : Spanish, French, Portuguese, Russian, Mandarin (China and Taiwan), Arabic, Hindi, Indonesian, German ;
- 25 central languages which are official in a least one central state and whose native speakers tend to be bilingual with a hyper-central, super-central language or other central languages : Dutch, Italian, Japanese, Hungarian, Greek, Turkish, Estonian, Vietnamese, Georgian, Finnish, Slovenian, Romanian, Latvian, Polish, Persian, Mongolian, Dhivehi, Ukrainian, Kinyarwanda, Thai, Tamil, Maltese, Swedish, Lithuanian, Czech ;
- 22 peripheral languages which occupy a subordinate position in the central states where they are spoken and whose native speakers tend to be bilingual with another central, super-central or hyper-central language : Catalan, Chinese in Hong-Kong (equal to Cantonese), Basque, Frisian, Hakha Chin, Tatar, Breton, Kyrgyz[7], Irish,[8] Chuvash, Romansh Sursilvan, Romansh Vallander[9], Luganda, Abkhaz, Assamese, Upper Sorbian, Votic, Odia, Punjabi, Kabyle, Welsh, Sakha (Yakut)
- 2 international auxiliary languages : Interlingua[10] and Esperanto.

[6] As we will see later, Common Voice contributors appear to be a very specific part of the speech community.

[7] Kyrgyz is the official language of Kyrgyzstan but in a diglossic relation with Russian

[8] Although Irish is an official language in Ireland, it has a very low number of speakers and it is classified as "definitely endangered" by UNESCO. These two elements justify its classification here as a peripheral language.

[9] Romansh is official in Switzerland but does not have the same legal status than Italian, French and German.

[10] Language created by the International Auxiliary Language Association based on Romance languages.

Published in : Andreo-Raynaud, Grégoire ; Douidi, Nadhir (dir.), *Langue(s) en mondialisation. Libre(s) échange(s) à l'heure néolibérale ?*, Editions des archives contemporaines, ISBN : 9782813004284, doi : 10.17184/eac.9782813004284

In large part this panorama reflects the sociolinguistic situation in the wider world. Nevertheless, we note a larger diversity both typologically and in the demographic situations of languages other than in techno-vehicular languages developed so far by other private companies. Among the 23 languages classified as peripheral, we count five endangered languages according to the UNESCO Atlas. One of them, Votic, is even classified as "critically endangered" and has about 15 speakers. We identify 15 language families (9 in techno-vehicular languages so far) plus two international auxiliary languages : Interlingua and Esperanto.

Table 1 presents a ranking of these languages according to the number of contributors. The amount of recorded hours for each language is indicated in parentheses (as a decimal number).

Table 1. This table gives the ranking of languages present in Common Voice by number of contributors, starting with English in first place with 64.995 contributors. The total number of hours collected is in parentheses after this number, for English 1600 hours.

Language ranking per contributors				
1. English 64.995 (1600)	2. Spanish 19.350 (306)	3. French 12.592 (591)	4. German 12.409 (742)	5. Italian 5.596 (139)
6. Catalan 5.133 (555)	7. Persian[12] 3.595 (285)	8. Chinese 3.452 (55)	9. Polish 2.605 (107)	10.Chinese Hong-Kong[13]. 2.471(54)
11. Chinese Taiwan 1.426 (54)	*12. Welsh 1.318 (88)*	13 Russian 1.305 (110)	14. Kabyle 1.282 (572)	15 Portuguese 1.033 (50)
16. Basque 1.025 (89)	17. Dutch 970(47)	18. Turkish 665 (20)	19. Esperanto 560 (87)	20. Estonian 541 (23)
21. Arabic[14] 479 (13)	22. Ukrainian 429 (27)	23. Kinyarwanda 409 (1200)	24. Mongolian 367 (11)	25. Frisian 356 (12)
26. Hakha Chin 297 (2,4)	27. Czech 296 (30)	28. Japanese 255(9,10)	29. Tamil 246 (14)	30. Swedish 211 (10)
31. Tatar 179 (26)	32. Indonesian 178 (8,75)	33. Maltese 169 (7,47)	34. Dhivehi 166 (12)	35. Thai 153 (6,85)
36. BRETON 151 (7,07)	37. Kyrgyz 134 (11)	38. Greek 109 (6,78)	39. Romanian 107 (5,73)	40. Irish 101 (3,17)
41. Latvian 97 (6,00)	42. Chuvash 84 (5,18)	43.Romansh Sursilvan 78 (5,75)	44. Slovenian 74 (5,10)	45. Luganda 48 (0,22)
46. Vietnamese 47 (0,55)	47. Sakha 41 (4,25)	48. Georgian 40 (3,38)	49.Romansh Vallader 37 (2,03)	50. Hungarian 35 (5,20)
51. Odia 34 (0,83)	52. Interlingua 28 6,05)	53. Lithuanian 25 (1,47)	54. Punjabi 24 (0,30)	55. Finnish 19 (1,07)
56. Upper Sorbian 18 (2,30)	57. Assamese 16 (0,73)	58. Abkhaz 14 (0,05)	**59. Votic 3 (0,001)**	60. Hindi 3 (0)
Techno-vehicular languages / *Vulnerable languages* / Definitely endangered languages / SEVERELY ENDANGERED LANGUAGES / **Critically endangered**[15]				

Ten of these exceed one thousand contributors. This is certainly due to their super-central or hyper-central status. In the case of English, French and German they also benefit from the help of some Mozilla staff who have been involved to start the project and who speak these languages. We must also notice that this ranking does not include the relative ranking of regional variants of these languages (e.g. Canadian French or New-Zealand English).

Published in : Andreo-Raynaud, Grégoire ; Douidi, Nadhir (dir.), *Langue(s) en mondialisation. Libre(s) échange(s) à l'heure néolibérale ?*, Editions des archives contemporaines, ISBN : 9782813004284, doi : 10.17184/eac.9782813004284

In order to have an idea of the activity rate of contributor communities, we must consider another ranking : the ratio of hours recorded to the number of speakers (Table 2). The goal is to show the level of activity in each community of contributors.

Table 2. Ratio hours recorded to the number of speakers

Ratio hours on contributor				
1.Kinyarwanda 2.93	2.Kabyle 0.45	3.Interlingua 0.22	4.Esperanto 0.15	4.Tatar 0.15
4.Hungarian 0.15	5.Sorbian, Upper 0.13	6.Catalan 0.11	7.Czech 0.10	7.Sakha 0.10
8.Basque 0.9	9.Georgian 0.08	9.Kyrgyz 0.08	9.Russian 0.08	9.Persian 0.8
10.Dhivehi 0.07	10.Romansh Sursilvan 0.07	10.Slovenian 0.07	*10.Welsh 0.07*	11.German 0.06
11.Ukrainian 0.06	11.Tamil 0.06	11.Greek 0.06	11.Latvian 0.06	11.Chuvash 0.06
11.Lithuanian 0.06	12.French 0.05	12.Portuguese 0.05	12.Dutch 0.05	12.Swedish 0.05
12.Indonesian 0.05	12.BRETON 0.05	12.Romanian 0.05	12.Romansh Vallander 0.05	12.Finnish 0.05
13.Polish 0.04	13.Chinese Taiwan 0.04	13.Estonian 0.04	13.Maltese 0.04	13.Thai 0.04
13.Japanese 0.04	13.Assamese 0.04	14. Irish 0.03	14.Frisian 0.03	14.Mongolian 0.03
14. Turkish 0.03	14. Arabic 0.03	15.Italian 0.02	15.Chinese (china) 0.02	15.Chinese Hong-Kong 0.02
15.Odia 0.02	15.English 0.02	15. Spanish 0.02	16.Hacka Chin 0.01	16.Vietnamese 0.01
16.Punjabi 0.01	**17.Votic <0.01**	17.Luganda <0.01	17.Abkhaz <0.01	
Techno-vehicular languages / *Vulnerable languages* / Definitely endangered languages / SEVERELY ENDANGERED LANGUAGES / **Critically endangered**[16]				

In this ranking, hyper-central and super-central languages are out of the top-ten and peripheral languages are in a higher position. It seems that contributors to peripheral languages are much more motivated and active than those from hyper-central and super-central languages. This is particularly noticeable in the case of Kinyarwanda with a total of 1,200 recorder hours for 409 contributors or Kabyle with 1,282 hours for 597 contributors. This outlier is partly motivated by the fact that Common Voice is participating in a development programme for language technology in Rwanda with the participation of non-governmental organisations.

3.6 Aspects of the Common Voice contributor communities

3.6.1 Contributor profile

Contributors to Common Voice participate in a variety of different ways in the project. They can be split into core contributors and more peripheral contributors. The core contributors are typically people involved with the Mozilla localisation[11] community ;

[11] Localisation refers to the process of translating and adapting text content to a locale (combination of language and region).

Published in : Andreo-Raynaud, Grégoire ; Douidi, Nadhir (dir.), *Langue(s) en mondialisation. Libre(s) échange(s) à l'heure néolibérale ?*, Editions des archives contemporaines, ISBN : 9782813004284, doi : 10.17184/eac.9782813004284

and generally are much more heavily invested in the project and in Mozilla's other projects.

These participants invest a lot of time to collect and record sentences and to mobilise other contributors. They are the core of each language community on Common Voice. Aside from these core participants, there are other contributors. Their involvement varies, but they do not often contribute with a leading role.

Contributors have a similar profile in terms of gender, age, education, and language. They are also typically already involved in other Mozilla projects : most of the recruitment is thus endogenic. Most of them are highly-educated and speak more than one language. According to our research, a large proportion of them are employed in teaching-related or computer-related professions. They are not necessarily linguists and do not consider themselves as linguistic experts but they are "language curious" or they have a personal interest in the particular language they are contributing to.

We speculate that most of them share the "open-source philosophy", a belief in open access to intellectual products (see footnote 4). They also often have a critical view of GAFAMs.[12] An Italian contributor answered the question "*What could make you contribute more [to Common Voice] ?*" saying "*Stopping the use of GAFAMs tools (like AWS, Google Fonts, etc.) on the Website.*". This comment indicates a very critical view of employing non-open platforms and engaging with companies that are seen to be not following the principles of the open-source movement. This specific profile can explain the presence of Esperanto and Interlingua in the Common Voice dataset. Indeed, a lot of speakers and activists of international auxiliary languages are concerned with globalisation issues and can identify within values that Common Voice promotes (Heller and McElhinny, 2017, 128).

A large part of the primary contributors (we will refer to these also as "leaders") are also language activists. The Kabyle[13] community, for instance, is promoted and supported by a computer scientist working on an open-source project who also appears to be a language activist. This situation is strongly favourable because he quickly heard of Common Voice and he enjoys the technological skills and the network to involve the actual speech community over the Common Voice project. He was able to collect a lot of sentences to record in Kabyle with the help of a writer who agreed to allow his books to be used under an open-source licence. He also gives conferences gathering people who are interested in the defence and the spread of Kabyle. At the same time, he directly recorded the voices of people who do not have the technological skills to do it themselves. The community leaders of Welsh language are also both computer scientists who are open-source enthusiasts, language scholars and activists. Therefore, those two languages have reached 1,000 contributors and are in the top-ten in terms of ranking by number of contributions. This situation affects mostly peripheral or central languages which benefit from a community of activists or institutional support.

[12] This acronym is in wide use in the technical community and refers to Google, Amazon, Facebook and Microsoft.

[13] Kabyle (ISO-639-3 : kab) is a Berber language spoken in Algeria and in a wide diaspora.

Published in : Andreo-Raynaud, Grégoire ; Douidi, Nadhir (dir.), *Langue(s) en mondialisation. Libre(s) échange(s) à l'heure néolibérale ?*, Editions des archives contemporaines, ISBN : 9782813004284, doi : 10.17184/eac.9782813004284

3.6.2 Community dynamics

Leaders with linguistic and technical skills and connection with language activists create active communities. Others are substantially less dynamic. This is the case for Occitan for example, a language spoken by about 1,000,000 people in France, Spain, and Italy.[14] The collection of sentences in Occitan is slowed by the lack of connection between the person doing the interface translation (the localiser) and the broader activist community. The existence of other projects for data collection is also an explanation, since the linguistic institutions for this language have launched their own project. Another barrier can be the characteristics of the speech community itself. In the case of Votic, an endangered Uralic language, closely related to Finnish and Estonian, the main translator is an Estonian computational linguist who got interested in endangered Uralic languages during his studies. He has been teaching Votic to children in Russian villages every summer for a decade. He has the skills and some access to a network of activists, but Votic has only about 15 living native speakers and the recordings are done based only on his fieldwork and educational work.

When a community leader has both the skills and the network to develop a language on Common Voice, it still requires convincing people to give their time to record sentences and/or validate recordings. In the next section we will analyse what the motivations of contributors are.

3.6.3 Motivations of contributors

Based on our survey, the motivations of contributors are quite diverse.y. We propose to analyse them as two different categories. The first one would be the motivations linked to discourses of endangerment (Duchêne & Heller, 2007). The answers we consider for this category are : "*I want to protect language diversity in technologies" ; "I want to collect datas* [sic] about a endangered language" ; "*I want to struggle against linguistic and cultural assimilation*" ; "*I do it for political reasons*".

In different but related ways, contributors perceive Common Voice as a way to protect their languages against a "threat" which can be precisely connected to political issues or just to discourses of language preservation (Duchêne & Heller, 2008). This set of motivations is overrepresented in our survey with 61.7% of the given responses. They seem to invest their contribution in either a symbolic value to their contribution or a practical way to "do" something for a language they care about. 7.8% of respondents state that they contribute in order to collect data about an endangered language. As Eisenlohr (2004, 24) pointed out, new technologies are "*an inexpensive and effective way of recording linguistic practice in lesser-used languages*". In an interview, the Votic specialist, explained his use of Common Voice like this. This contributor also mentioned that Common Voice is a nice opportunity for young people who learn Votic language to practice the language which is becoming more and more difficult in other situations. We can imagine the possibilities for language revitalisation of critically endangered languages. A system with artificial intelligence that was fluent

[14] Based on an extrapolation from the last sociolinguistic study by the Ofici Public de la Lenga Occitana (Office Public *de le Langue Occitane*) in 2020.

Published in : Andreo-Raynaud, Grégoire ; Douidi, Nadhir (dir.), *Langue(s) en mondialisation. Libre(s) échange(s) à l'heure néolibérale ?*, Editions des archives contemporaines, ISBN : 9782813004284, doi : 10.17184/eac.9782813004284

in Votic (for example) to speak with could be a way to compensate for the loss of last speakers. Some projects are following this path like *Idiomata* or *Macsen* for Welsh (Jones, 2020). We must mention that ideological stand for diversity is one of the arguments used by Common Voice staff to convince people to participate like in this video designed for a communication campaign (never released) which said : "*we want everyone to be understood*". It takes stand on the spread of language preservation institutional discourses (such as those from UNESCO), and the common beliefs that languages must be preserved in behalf of world's cultural diversity. This examination of motivations is consistent with the connection between motivation rate and language status we highlight in Table 2.

Another set of motivations is more linked to the will to help implement language technologies in a specific language ("*I want voice or text recognition based technology in my language*"). It represents 26.7% of given responses. It can either mean that they want their language being used to communicate with machines or they want alternatives to existing tools that allow it. In the second case, it is connected to the open-source movement and the philosophy that underlies it. For instance, an Italian contributor commented : "I want to have an alternative to Google, Microsoft and Amazon's monopoly on voice recognition software".

A third set of motivation, much more narrow, is the use of Common Voice as a tool for language learning. Indeed, the platform gives them the possibility to get feedback from other speakers of their target language. This motivation is quite underrepresented in our data (2,5%) and we can wonder what is the real weight of this kind of use, which can be considered a "side effect" as it is not at all the original point of Common Voice. According to the language spoken by respondents who give this answer, we can speculate that it can concern the learning of central and super-central languages like English but also the learning of peripheral languages as such as heritage languages (Polinsky and Kagan, 2007).

4 Sociolinguistic bias impacting Common Voice dataset

4.1 Sociology, language status and language norms

As a community-based policy with low regulation, it is not protected from the same inequalities that have an impact on the rest of society. In particular, the fact that it is a technology-based language policy tends to reproduce technological inequalities, as has already been noted by Eisenlohr (2004 :26).

This sociology of technology usage which is reflected in Common Voice may impact the sociology of techno-vehicular languages. We saw in the second section that even though Common Voice was designed to encourage access to technology for peripheral languages, super-central languages are still more present presumably because most of the potential contributors live in countries where these languages are spoken and these languages have greater demographic weight. This explains the hegemony of English both in terms of contributors and contributions.

The risks do not only concern the over-representation of some languages, but the over-representation of dominant language norms in each language. We can mention

Published in : Andreo-Raynaud, Grégoire ; Douidi, Nadhir (dir.), *Langue(s) en mondialisation. Libre(s) échange(s) à l'heure néolibérale ?*, Editions des archives contemporaines, ISBN : 9782813004284, doi : 10.17184/eac.9782813004284

the under-representation of oral speech in many languages that tends to favourise written norms usually used by dominant groups. In the case of endangered languages, the sociology of Common Voice favourises standard forms used by second-language speakers over native speaker varieties (mostly old and rural). The "Hong-Kong Chinese" case also shows some struggle over language denominations. Indeed, some Cantonese speakers from China claimed that the "Hong-Kong Chinese" was not a proper name because it prejudiced many Cantonese speakers from China. We can also mention the case to demand the implementation of Bosnian as a separate language from Serbian and Croatian which was considered as "exclusive" by a contributor and recurrent requests to separate Brazilian Portuguese and Iberian Portuguese.

This situation is interesting, even if it has its origin in technical issues (and maybe a lack of linguistic knowledge from the first localiser). So far, Cantonese speakers from China were not able to implement Cantonese on Common Voice and they can only contribute to "Hong-Kong Chinese".[15]

4.2 Language standardisation

The social and political situation of a language also impacts its standardisation and therefore, the availability of text corpora. This is the case for Mandarin in Taiwan, which is used in Common Voice over Taiwanese or indigenous languages because of a lack of CC-0 texts corpora in these languages due to the diglossic situation there. Indeed, Mandarin is the official language and main education language in Taiwan. It is also linked to the degree of standardisation of a language. A language with official status is more likely to have a writing system, ISO-639 language code and unified grammatical norms that makes written uses of it more normalised, and, in the end, its presence in Common Voice more likely.

4.3 Language inclusivity

As we have seen before, there is a lack of language diversity in techno-vehicular languages. This situation is not limited to minoritised languages and many other kinds of voices are muted. Not surprisingly, like in television or radio, the same minoritised groups of people are concerned : women, the elderly, the economically disenfranchised, etc. In the case of Common Voice, the contributor profile we draw based on information is consistent : a highly-educated male between 19 and 40.[16] The gender and age gaps are particularly noticeable according to available data. Most of the datasets are composed by male voices. The same thing is seen in age distribution : most of the contributors are between 19 and 40. This situation must lead the relevant actors to question how they will go about ensuring inclusivity in voice technologies and more specifically in techno-vehicular languages. This is the question of who machines will sound like and who they will understand better.

[15] Since this article was written, Cantonese has been added to Common Voice.

[16] Though, when we analyse the precedent datasets released, we see a diminution of the gender and age gaps. In the French dataset for instance, the gap decreased by 11 points between February 2019 and June 2020. This is link, we think, with a campaign in the mainstream media lead by A. Lissy during this period).

Published in : Andreo-Raynaud, Grégoire ; Douidi, Nadhir (dir.), *Langue(s) en mondialisation. Libre(s) échange(s) à l'heure néolibérale ?*, Editions des archives contemporaines, ISBN : 9782813004284, doi : 10.17184/eac.9782813004284

4.4 Bias in the language economy

Companies are making efforts to develop technology for as yet unsupported languages. Still, we can question the impact of this situation on the sociolinguistic situations of multilingual countries where co-official languages are systematically left aside. In diglossic situations, this reinforces the position of the dominant vehicular languages, often the former colonial languages, over others. This applies even if they are spoken by millions of people and have official status like Urdu in Pakistan. It obliges speakers to use the dominant language, even at home, and it affects the social status by giving them more prestige by relating the use of language in technology with modernity (Esenlohr, 2004). Certainly, those companies also try to stick to "official languages" to avoid making choices that could be seen to favour one party in a political dispute over another. Regardless, from our perspective, we could characterise this situation as "sociolinguistic" bias or more precisely "language policy bias" as regarding the action about language practices by specific actors.

The sustainability of the Common Voice project is dependent on the support of the Mozilla Corporation and communities of contributors. In August 2020, the Mozilla Corporation started to restructure due to financial problems which lead to removing most of its staff from working on Common Voice. The project was largely unmaintained for a number of months before being moved to the Mozilla Foundation, a non-profit organisation. The development team was then hired by the NVIDIA corporation and currently the community and planning side of the project is run by Common Voice staff at the Mozilla foundation with engineering support from NVIDIA. This situation shows the limits of a grass-roots project with no short-term profitability (Streiter *et al.*, 2006, 5-6) and also highlights the importance of each language's economic weight.

4.5 Language policies and open source

A final sociolinguistic bias impacting Common Voice lies in open-source practice. All linguistic data collected in Common Voice are free to use by everyone. This situation is accepted by many language activists but some of them, like activists for the Māori language, feel that putting their language data in the public domain would further disenfranchise their already marginalised communities. This open-source strategy, they argue, risks commodifying Māori language and culture. They do not want to see companies freely develop projects for Māori without input from Māori speakers and the Māori-speaking community. The profit that could be made from treating their language as a resource for "extraction" was reminiscent of the process of colonisation (from personal communication). According to Eisenlohr (2004 : 27), the same kind of criticism over the use of technology in language revitalisation was made by Native American groups in the United States. It is likely that many activist communities have different sorts of criticism about open source and prefer to create their own dataset under other licences.[17]

Specific open-source requirements also have an impact on the Common Voice dataset in other ways that replicate biases in wider society. The necessity to use CC-0 (public domain) sentences presents a material disadvantage to language communities with

[17] This is linked to a wider debate about different open-source licences.

Published in : Andreo-Raynaud, Grégoire ; Douidi, Nadhir (dir.), *Langue(s) en mondialisation. Libre(s) échange(s) à l'heure néolibérale ?*, Editions des archives contemporaines, ISBN : 9782813004284, doi : 10.17184/eac.9782813004284

a recent tradition of writing or with a more recent linguistic norm. For example, the orthographic reform of Turkish happened in 1928, giving 22 years for public domain texts, compared to probably 200 years for French or English. Many language communities who wish to participate may have essentially zero years of public domain text available.

5 Conclusion

Common Voice is just one example among many of language policy planning in techno-vehicular languages. In between private and public agents, its intermediary way of founding, functioning as well as the data it provides, constitute a starting point for approaching the field. Indeed, it elucidates some of the technical, financial and ideological aspects of techno-vehicular languages and has uncovered issues that can influence the existence of linguistic datasets and therefore, the existence of a techno-vehicular language.

The subject of techno-vehicular languages can be investigated by a number of approaches from sociolinguistics. Taking a critical approach Heller (2011 : 39) raises the question of new forms of commodification of language practices and the competition that is provoked in competing claims for resources. We could also see the questions surrounding techno-vehicular languages as central in language-planning policy for endangered languages due to the impact that choices of language in technology have on linguistic norms and language prestige (Eisenlohr, 2004). There are also questions relating to techno-vehicular languages to be answered in other fields such as language learning (Eskenazi, 2009), such as the increasing use of chatbots to replace native speakers in conversation training.[18] And as Kaplan (2015) pointed out, technical biases are some of the most important questions in language practices.[19]

We argue that the issues surrounding the management of techno-vehicular languages will become increasingly relevant in the years to come due to the increasing role that speech interfaces are playing in mediating communication with "artificial intelligences", which have become part of our daily environment. This study of how, and by whom, techno-vehicular languages are created is a first step toward a sociolinguistic approach to the question in this field.

References

Ardila Rosana, Branson, Megan *et a*l, 2020, « Common Voice : A Massively-Multilingual Speech Corpus », *Proceedings of the 12th Conference on Language Resources and Evaluation* (LREC 2020), p.4211-4215.

Calvet Louis-Jean, 1999, *Pour une écologie des langues du monde*, Paris : Plon.

Carillo Kevin and Okoli Chitu, 2008, « The Open Source Movement : A Revolution in Software Development », *Journal of Computer Information Systems*, 492(2), p. 1-9.

Duchêne Alexandre, et Heller Monica, 2008, « Discourses of endangerment : Sociolinguistics, Glob-alization and social order », dans Duchêne, A. (éd.), Discourses of Endangerment : Ideolo-gy and Interest in the Defence of Languages, London : Paperback ed. Continuum, p. 1-13.

[18] See example with Duolingo : . Project "Idiomata" for less resourced languages.

[19] These kinds of technological bias are linked with sociolinguistics studies when we start asking why is English used as "pivot language" ?

Published in : Andreo-Raynaud, Grégoire ; Douidi, Nadhir (dir.), *Langue(s) en mondialisation. Libre(s) échange(s) à l'heure néolibérale ?*, Editions des archives contemporaines, ISBN : 9782813004284, doi : 10.17184/eac.9782813004284

Eisenlohr, Patrick, 2004, « Language revitalization and new technologies : Cultures and electronic mediation and the refiguring of communities », *Annual Review of Anthropology*, 18(3), p.339–361.

Eskenazi Maxine, 2009, « An Overview of Spoken Language Technology for Education », *Speech Communication* 51, p. 832-44.

Heller Monica and McElhinny Bonnie, 2017, *Language, capitalism, colonialism : toward a critical history*, Toronto, Ontario : University of Toronto Press.

Hornberger Nancy, 2006, « Frameworks and Models in Language Policy and Planning Research », in Thomas Ricento, ed. *An Introduction to Language Policy : Theory and Method.* p.24-41, Oxford : Blackwell.

Jernudd Björn H. and Das Gupta Jyotirindra, 1971, Towards a theory of language planning in : *Sociolinguistic theory and practice for developing nations*, The Hague : Mouton.

Jones Dewy B., 2020, « Macsen : A Voice Assistant for Speakers of a Lesser Resourced Language_ »_ *Paper in LREC 2020 SLTU-CCURL Workshop*, p. 194-201.

Kaplan Frederic, 2014, Linguistic Capitalism and Algorithmic Mediation, *Representations. 127(1)*, p.57-63.

Larsonneur Claire, 2019, « The disruptions of neural machine translation ». *Sphères, n° 5.* [online], 20th of November 2019.

Nekvapil Jiří and Sherman Tamah, 2015, « An Introduction : Language Management Theory in Language Policy and Planning », *International Journal of the Sociology of Language* 2015(232), p.1-12.

Polinsky Maria and Kagan Olga, 2007, « Heritage Languages : In the ?Wild ? And in the Classroom », *Language and Linguistics Compass* 1(5), p.368-95.

Perea François, 2016, « Nature et technologie langagière dans les dialogues oraux homme-machine », *Communication* [En ligne], 34/1, Accessed : 16th October 2020.

Ricento Thomas K. and Hornberger Nancy, 1996, « Unpeeling the Onion : Language Planning and Policy and the ELT Professional », *TESOL Quarterly* 30(3), p.401.

Streiter Oliver, Scannell Kevin P. and Stuflesser Mathias, 2006, « Implementing NLP Projects for Noncentral Languages : Instructions for Funding Bodies, Strategies for Developers », *Machine Translation* 20(4), p.267-89.

Swaan Abraam, 2001, *Words of the world : the global language system.* Cambridge, UK ; Malden, MA : Polity.

Wenger Etienne, 1998, *Communities of Practice : Learning, Meaning and Identity.* Cambridge New York Port Melbourne : Cambridge University Press.

Published in : Andreo-Raynaud, Grégoire ; Douidi, Nadhir (dir.), *Langue(s) en mondialisation. Libre(s) échange(s) à l'heure néolibérale ?*, Editions des archives contemporaines, ISBN : 9782813004284, doi : 10.17184/eac.9782813004284