Sociolinguistic aspects of crowdsourcing for a vocal corpus of Alsatian

Pascale Erhart[1], **Lucile Hamm**[1], **Carole Werner**[1]
Malek Yaich[2], **Sam Bigeard**[2], **Slim Ouni**[2]
[1] LiLPa, Université de Strasbourg, Strasbourg, France
[2] Université de Lorraine, CNRS, Inria, LORIA, F-54000 Nancy, France
pascale.erhart@unistra.fr lucilehamm@unistra.fr wernerc@unistra.fr
malek.yaich@inria.fr sam.bigeard@inria.fr slim.ouni@loria.fr

Abstract

Alsatian is a regional low-resource language spoken in a majority-language context. In order to create a voice dataset suited for training automatic speech recognition and speech-to-text models, we launched a crowdsourcing campaign on the platform Mozilla Common Voice. We describe sociolinguistic issues we ran into, such as participants' perception of their own language and its role in the AI landscape, which are vital to address to raise the participation in the crowdsourcing effort. We found that the participants are often confused about NLP and AI tools, and have a strong interested in preserving their language.

Keywords: Alsatian, Regional languages of France, German dialects, Crowdsourcing, Sociolinguistics

1. Introduction

Alsatian is a regional low-resource language spoken in a majority-language context: it is a German dialect spoken in France, where no language other than French benefit from any statutory recognition. Alsatian is one of the languages for which the COLaF project (Corpus et Outils pour les Langues de France, Corpus and Tools for the Languages of France) aims to contribute to the development of free corpora and tools. For Alsatian more specifically, the project aims to develop automatic speech recognition (ASR) and speech to text (TTS) tools, which requires speech and text training data in Alsatian. In order to overcome the usual issues associated with low-resource languages, such as limited and poor-quality data, the creation of a new corpus was undertaken via a crowsdsourcing campaign. For crowdsourcing to be successful, it is vital to take into account the public's perception of the topic. Thus, in this paper, we report on the sociolinguistic aspects and issues we encountered during the launch of the campaign, and what we did to address them.

The main contributions of this article are as follows:

- We give a detailed description of Alsatian and its sociolinguistic context to highlight the specific challenges encountered in developing NLP tools for this language (Section 2).
- We provide a state of the art in the field of Automatic Speech Recognition in Alsatian to show why the creation of a specific corpus is necessary (Section 3)
- We describe and discuss the different challenges we met during the implementation of the crowdsourcing campaign (Section 4).

2. Language context

The name "Alsatian" (als. *Elsässisch* or *Elsassisch*) is frequently used today to refer to the Alemannic and Franconian dialects spoken in Alsace since the 5th century AD. In this section, we describe the main characteristics of this low-resource language as well as its sociolinguistic context.

2.1. Dialect variation

The Alsatian dialect area is structured by phonetic elements which have their origins in the linguistic history of the German dialect area. The Alsatian dialects are mainly part of Upper German and partly part of Central German, which, from a geolinguistic point of view, make up the High German group.

Although it is characterized by variations in pronunciation as well as morphological or lexical variations, the Alsatian dialect area displays features that are shared by nearly all variants and allows its distinction from other German dialect areas. Among these features listed by Huck (2022), we can mention the regular palatalization of Middle High German (MHG) û (*Müs* [mys] 'mouse'), the lowering of MHG ë into [a], except in the northwest, the far north and along the Rhine up to the south of Strasbourg (*Laawe* [la:və] 'life'), the fact that /b, d, g/ consonants are realized as [b̥, d̥, g̊], that is to say as weak voiceless occlusives, or the de-rounding of certain rounded palatal vowels as in *scheen* [ʃeːn] ('beautiful') or *mied* [miːət] ('tired'). This

Proceedings of Workshop on Dialects in NLP — A Resource Perspective (DialRes) , pages 256–264
16 May 2026.

phenomenon affects nearly all the Alsatian area, whereas in Swiss German dialects, for example, labialization is maintained.

2.2. Effects of language contact

Alsace is a border region that was part of the German Empire for many centuries before becoming French at the end of the 17th century. During the last two centuries, it has been a theater for the conflicts between France and Germany and changed its state affiliation four times in less than 100 years (1871, 1919, 1940, 1945). Since 1945, Alsace is part of the French Republic, which officially recognizes French as its sole legitimate and official language. Since 2008, the French Constitution considers regional languages as heritage languages, but they are not subject to any specific policy. [1] This general context explains, at least partially, why the number of Alsatian speakers has declined since 1945. In the most recent survey conducted at the regional level Collectivité européenne d'Alsace (2022), 46 % of respondents stated that they 'speak Alsatian fairly well or very well', but three-quarters of these speakers were over 60 years old. The survey also reveals that understanding Alsatian is more important than speaking it, leading to a loss of the language.

Furthermore, it is important to note that nowadays all Alsatian speakers are also speakers of French, and often of other languages as well. This means that the use of Alsatian can be replaced by another language anytime and that speakers can use the multilingual resources of their repertoire in any spoken or written utterance in Alsatian. Language contact effects are thus numerous even in sentences that could be categorized as « in Alsatian ». Effects of language contact, especially with French, affect the morphosyntactic as well as the lexical level of the Alsatian dialects (Koehler, 2024) and make them diverge from other German dialects like Swiss German (although the High Alemannic dialects spoken in Alsace remain very close to the Swiss German dialects on the phonetic level) Koehler (2024). This explains why we cannot only draw on research on any other German dialect (like Swiss German or Bavarian) and need specific datasets for the finetuning of ASR or TTS models for Alsatian.

2.3. Graphic variation

At the local level, in Alsace, it was primarily a political vision, based on the idea that only a standard language could be taught in schools, that led the then head of the regional education system, Rector Pierre Deyon, to define German as the "regional language" of Alsace. The Académie de Strasbourg's current definition of "regional language" is as follows: « The regional language of Alsace is understood to be German in its standard form and in its dialectal variants (Alemannic and Franconian)» [2]. It is noteworthy that the term "Alsatian" is absent from this definition. Although both variants (dialect and standard German) are officially recognized as the "langue régionale d'Alsace", standard German is almost exclusively taught in public schools, with the exception of experimental trilingual programmes (French-Alsatian-German) introduced in 2023, which involve a very small number of pupils Hamm (2024).

When the opportunity arose in the 1980's to introduce a "bilingual" education in French and regional languages, « it was virtually impossible in Alsace to reach agreement on whether a bilingual education policy meant teaching both "French" and "Alsatian", or whether it should mean teaching "French" and "German", as "Alsatian" was not identified as a "language" » Le Page and Tabouret-Keller (1985). This view of Alsatian as a non-language still prevails and explains why it is still virtually absent from education today. Because of this lack of institutional recognition, Alsatian not being subject to any language policy, no intervention has been made regarding the corpus of Alsatian, so that the written corpus of this language is limited and characterized by graphical variation and the lack of a standard spelling. Its spelling is not normalized and reflects geolinguistic variation. This lack of institutionalisation also explains why Alsatian doesn't appear in language classifications like Glottolog Hammarström et al. (2025), where it generally falls under the umbrella of Swiss Alemannic, even if all Alsatian dialects are not part of the Alemannic area. [3]

Until the beginning of the 21st century, the use of written Alsatian was limited to regional literature, and more specifically to theater and poetry, so that "common/lay" speakers are not used to read or write in Alsatian. This changed with the arrival of digital communication and new writing habits : many people use Alsatian in their digital communication, especially on social media, which allows to observe how speakers deal with language variation in their written use of Alsatian : the absence of standard spelling systems leads to interpersonal variation,

[1] Constitutional Law n° 2008-724 of July 23, 2008 on the modernization of the institutions of the Fifth Republic, article 75 1 : « les langues régionales appartiennent au patrimoine de la France »

[2] Agreement on the regional multilingual policy, period 2015-2030, signed by the representatives of the State, the Région Alsace, the Département du Bas-Rhin and the Département du Haut-Rhin

[3] ISO 639-3 documentation

in which each writer chooses their own spelling convention Erhart (2026) Erhart (2025). The ORTHAL method Zeidler and Crévenat-Werner (2008) that was created to help normalizing the writing Alsatian is not well known by the general public, since it is not supported by any other institution than the local association who developed it. It can nevertheless be useful to normalize the written data gathered and to prepare them for their phonetisation.

3. Automatic Speech Recognition for Alsatian

3.1. Automatic Speech Recognition

Automatic Speech Recognition (ASR) is the task of automatically transcribing audio speech into written text. The field has achieved significant results in recent years with the emergence of multilingual models, such as Hubert Hsu et al. (2021), or Chen et al. (2022). Models such as Keren et al. (2025) claim to be able to transcribe thousands of different languages. Whisper Radford et al. (2023) is able to transcribe over 50 languages out of the box, although with significant differences in quality. These new multilingual models are able to produce good results on languages with smaller training datasets, by taking advantage of larger training data in other languages. They also allow finetuning : re-training a model with a small amount of data from a new language. Our previous experiments on Whipser [4] have found that as little as 30 hours can be enough to drop the Word Error Rate (WER) from around 75% to around 10%, depending on source and target language. However, these experiments have also shown that absence of an orthographic standard, which is the case in Alsatian, is an important limiting factor.

3.2. Existing datasets

Currently available voice datasets for Alsatian come from either broadcast recordings, or ethno-linguistic studies.

The **Linguistic atlas of Alsace** Huck et al. (2014) is an ethno-linguistic corpus from 1965 - 1980 whose goal was to document the language of the time in its geographic variability. Its limitation are the age of the recordings, that might not represent the current language accurately, and be of lower quality ; and its large variability, an obstacle for machine learning.

FLARS is another ethno-linguistic corpus, from the project « Effects of the national border on the linguistic situation in the Upper Rhine area », gathered between 2012 and 2014 Erhart (2017). These recordings are more recent and thus better represent the current language, but still are focused on variability.

Broadcast archives are a potentially large source of data. However, getting access to this kind of data can be an arduous process, as they are typically under strong authorship protections. This type of recording also presents difficulties for ASR training : There can be background music, overlap of voices, and it can be difficult to identify all participants to get consent. They also rarely come with transcription, although seven of them were fully transcripted in Erhart (2012). For these reasons, this type of data is not included in this study.

None of those sources of data are made for machine learning, and share this same limitation : Audio files are several minutes long, if not hours, and if there is a transcription, it is not aligned at a smaller granularity. These datasets must go through the preprocessing step of chunking them into smaller parts, and align the transcription to these chunks, which is costly.

3.3. ASR experiments

These experiments were performed with Whisper, on whisper-large-v3 model constrained to German. We chose German because it is the closest available language to Alsatian, and is one of the languages where Whisper performs the best, with a WER of 5.7%.

Whisper at finetuning requires segments of 30 seconds at most. To chunk both audio and transcription, we used Montreal Forced Aligner McAuliffe et al. (2017).

The results of the fine-tuning experiment are presented in table 1. As expected, results on the German ASR model with no finetuning are very low. Finetuning with the few hours of data at our disposal raises significantly the quality of the model, showing that even a small increase in available data can be useful. But current available data is not enough to bring the results to an acceptable level.

Following these results, we decided to launch a crowdsourcing campaign. This will allow to generate a new dataset that will be bigger, already chunked into small segments, transcribed, and devoid of music or other noise.

Name	Dataset size	WER	CER
No finenuning	-	90.66	57.22
FLARS	9h 11min	48.89	27.47
Atlas	2h 42min	58.70	33.48

Table 1: Results of fine-tuning. Results are expressed in Word Error Rate and Character Error Rate as a percentage. Results are evaluated on FLARS and Atlas data.

[4] citation redacted for anonymisation

4. Crowdsourcing campaign

4.1. Preparation steps

Common Voice Ardila et al. (2020) is a platform that uses crowdsourcing to create voice data in a large panel of languages. It became one of the standard benchmarks for automatic speech recognition. In July 2025, it launched a new mode : Spontaneous Speech. Instead of reading a pre-written sentence, participants can freely answer a given question or prompt. This new mode reduces the amount of work needed to launch a new language : instead of entering hundred of written sentences for participants to read aloud (which might be challenging to gather in a predominantly oral language), only 60 prompts are needed to launch a language. Since participants are free to answer those prompts in their own words, the resulting dataset contains a larger variety of sentences. It also allows participants to contribute in their own varieties of their language, as long as the inter-comprehension is sufficient to understand the prompt.

We created 60 prompts trying to strike a balance in representing the different varieties of Alsatian, while keeping the sentences understandable for a speaker of any variety. The Alsatian adaptation could have been restricted to that, given that users would have the option to use the platform in their preferred language (the platform being already available in French, German or English). However, to avoid confusing participants and creating language interferences, and given that one of the campaign's objectives was to demonstrate to Alsatian speakers that their dialect can be used in any situation, regardless of its written form, it was decided to adapt the entire website into Alsatian. As Alsatian is not included in international classifications and does not have its own ISO code, it is listed under the abbreviation "gsw", which actually stands for Swiss German.

The translation was undertaken by three authors of this paper. All three are linguists and have Alsatian as their first language. The translation process was challenging due the characteristics of Alsatian described above. First, since only one Alsatian version of Common Voice could be offered, the decision had to be made as to in which Alsatian variety the site would be translated. The Low Alemannic dialect was selected given its prevalence (still with many variations) in 90 per cent of the Alsatian region. To make the platform as inclusive as possible of all dialect varieties, we made exceptions in cases where the lexical differences between varieties were too significant. For example, the verb "listen" can have two different forms: "horiche" in the northern and main part of Alsace and "loose" in the extreme south. We therefore decided to include both forms in the category appearing on the platform, i.e. "horiche/loose".

In addition, the issue of language contact effects had to be addressed. As a dialect of the High German group, Alsatian is spoken by individuals who also speak French and also have contact, at different levels, with German and/or English. A parallel lexicon of the four languages was established based on the three existing versions of the Common Voice website (English, German, French). The Alsatian form for each entry was chosen on the following criteria: In cases where a specific term existed in Alsatian, it was retained systematically, provided that it appeared sufficiently widespread and transferable within the context of the platform's usage. Otherwise, it was necessary to assess to what extent a transposition from Standard German could be understandable/acceptable, and whether it would be better to borrow the word from French or English, and if so, the extent to which it should be integrated into Alsatian. For instance, the Alsatian verb "fùrtschìcke" was chosen over the Standard German "absenden" (which could have been adapted into "àbsände"), and "Gschirrkischt" was selected over the English "toolbox", which was chosen for the German version of the website. Although the referent of the Alsatian word is usually a real object, we believe that it can also be used in an abstract sense specific to the digital world, and that this unconventional use may appeal to users. In comparison, we decided not to translate the English term "cookies", despite the existence of an equivalent Alsatian term ("bredle"), as the latter is too specific. It refers to Alsatian Christmas cakes and has entered regional French usage, meaning any other use would undoubtedly seem inappropriate.

The management of internationalisms and terminology specific to digital usage has also led to specific terminological choices in Alsatian relating to neology Erhart (2023a). The term "avatar" is used in French, German and English, for example, but since its use is not attested in Alsatian, it was decided to propose a term that would be more meaningful to users, namely "Profilbild," literally "profile picture". While many English terms in use in the digital world were not translated into German, it was decided to translate them directly into Alsatian, basing on the historical relationship between the two languages stemming from common Old Germanic : for example, the verbs to "upload" and "to download" were translated into the neologisms "hochlàde" and "runterlàde". The decision was also taken to translate the English word "webbrowser" by creating a new term, "Webbrüser". This was motivated by the wish

to illustrate the possibility of lexical innovation in Alsatian, and was based on a play with the sounds of the English word ”browser” and the Alsatian noun ”brüser”, which is derived from the verb ”brüse” (German ”brausen”) meaning ”to produce a steady, uninterrupted noise”.

Finally, the question of the graphic choices made for this translation arose with particular sharpness: the Alsatian version of the site had to be clearly distinct from both the German and French versions, while remaining readable for the majority of users who speak the different varieties of Alsatian and are not used to read or write it. Consequently, a relatively flexible approach to the ORTHAL method (see above) was adopted, using diacritical marks only for vowels characterised by specific pronunciation in Alsatian: à, ì, ù.

For all of the translation choices made, an attempt was made to achieve a balance between predictability, acceptability, and the creativity of the speakers ; these three criteria had previously been suggested by LePage and Tabouret-Keller (1985) to characterize linguistically heterogeneous situations, such as the Alsatian one. From this point of view, it would be interesting to know whether participants in the collection use the interface in Alsatian or in other languages and, if so, whether or not they adopt the terminology and/or spellings proposed in the Alsatian version of the site.

4.2. Results and feedback from the public

Events : A crowdsourcing campaign is only as strong as its communication plan. To generate interest, we held two in-person events : A press conference, presented as the launch of the campaign in October 2025, and a conference for the general public, in January 2026. Many of the exchanges we report below happened during one of those events, or in conversations immediately following them.

Reception and collection results : The press conference resulted in 10 press articles in national regional newspapers and radios. These included articles in two major daily regional newspapers, *L'Alsace* and *Dernières Nouvelles d'Alsace*, as well as the morning programme on the main regional public radio station, *Ici Alsace*. As of the writing of this paper, we collected 5 hours and 34 minutes of voice data. However, only 35 minutes have been transcribed.

Reactions regarding Alsatian and AI The way in which the Common Voice Spontaneous Speech data collection for Alsatian was presented in the press provides an initial indication of how it was perceived by journalists. For example, one of them admitted that he didn't really understand the purpose of the data collection. In his article, he wrote that « the language used by researchers can sometimes be obscure », yet he still invited Alsatian speakers to « donate their voice » and « help researchers develop regional languages, such as Alsatian, in the digital sphere ». Another article was published in the ”Insolite” (bizarre, quirky) section of the *Ici Alsace* radio station's website. This may suggest that the project was not taken seriously, or at least that it was perceived as an attempt to bring together two seemingly incompatible worlds: the Alsatian language and computational sciences.

This brings us to what appears to be the main obstacle to encouraging people to participate in this spectific data collection: the ideologies and attitudes of speakers towards their languages, and towards language technologies in general. Indeed, the reactions observed during the organised events, the messages received as well as the comments on press articles shared on social media generally confirm the common perception of Alsatian as a language linked to the past and with very low social prestige. These perceptions also suggest that its typical form is spoken by elderly people who are rural and less mobile, who have less contact with other languages and who are the usual witnesses of traditional dialect geolinguistics Bothorel-Witz (2007).

Consequently, the latter are regarded by the general public as the ideal candidates for the collection (one student told us « I will tell my grandpa about it » ; a colleague suggested going to a mountain cabin he visited on vacation where he heard the owners speaking Alsatian), whereas the project aims, in fact, to reach all speakers of Alsatian and to collect and document Alsatian as it is spoken today, encompassing its entire diversity. However, elderly speakers are less likely to be comfortable with digital tools and with participating in such a data collection exercise without the presence of an interviewer who would ask them questions in person, like in traditional dialect data gatherings.

Furthermore, variations in the language are subject to contradictory representations; they are perceived both as a defining characteristic of Alsatian and as an obstacle to mutual understanding between its different varieties. It has been repeatedly expressed as a concern that the Common Voice platform and the questions asked for the data collection do not correspond to the ”Alsatian” spoken by the participants.

Finally, even though the platform is primarily aimed at collecting voices in Alsatian, it confronts its users with written Alsatian, which they are not used to. Indeed, according to Huck (2002), in the perception of society, written Alsatian belongs to the « realm of the anecdotal and incidental, a realm

it has never really left since its emergence in the nineteenth century ». Despite its well-established use in digital communication Erhart (2025) Erhart (2026), written Alsatian remains largely used for informal communication. More formal uses, such as the tasks to be completed on Common Voice, i.e. answering questions formulated in writing in different varieties of the language or transcribing audio recordings in Alsatian into writing, are not yet widespread. While reading the questions does not appear to be an insurmountable task, it is noteworthy that most of the participants, probably concerned about their ability to transcribe Alsatian correctly, give up on the transcription activity. This would explain the low volume of transcribed data collected to date.

Reactions regarding technology and the campaign's goal

Eventually, to understand why people do or do not decide to participate in the collection, it is necessary to integrate the analysis of their attitudes towards the Alsatian language with that of their attitudes towards technology and their comprehension of the campaign's objective.

The press campaign's association of Alsatian with Artificial Intelligence (AI) may have resulted in potential candidates experiencing confusion, as many of them tend to associate AI with generative AI and anticipate (or fear) the emergence of an "Alsatian ChatGPT." Concerns were also expressed regarding the utilisation of the collected data and the ethical and personal data protection issues that would arise. (« So *anyone* can use my recordings to do *anything*? »)

There was also confusion between the activities of researchers and activities of the industry. If we explain that the campaign will help develop ASR technologies, people tend to understand that we will personally develop the application that will appear by itself on their phone. If we say we won't, they become confused about what it is we do, exactly, which might even erode trust into public research.

All of these examples are obstacles to public interest and involvement. To circumvent these obstacles, effective communication is imperative. In our experience, the most effective method of overcoming such potential resistances is through personalised, face-to-face or at least interpersonal explanations.

4.3. Technical difficulties

Finally, in part because Alsatian has been pigeonholed into a topic of interest for the older generation, we had many users confused about how to use the platform. Common Voice is an old and large project. Users had trouble navigating the interface of the large website to find the Spontaneous Speech section. Some participants did not understand that the task consisted in answering the questions and recorded themselves reading the questions out loud. In some cases, their browser automatically attempted to translate the Alsatian prompts into French, a functionality that the users didn't know existed nor how to turn it off.

We mitigated these issues by creating a homepage giving out simple instructions and direct links to the relevant pages, and asked the press to share a link to this homepage, rather than directly Common Voice. We gave a contact e-mail in all our communications, and were reactive in answering many requests for support, with kind help from Mozilla when we couldn't solve an issue ourselves. However, it was difficult to help people navigate technical difficulties remotely. This is why we organised the second, public, event, and included time for the participants to set up their account while we were physically in the room and able to assist them.

From this, we conclude that there are advantages and issues with using a large, well-recognized platform such as Common Voice, rather than creating your own platform for your usage : The platform is more complex, less easy to navigate for your users. But the advantage is the technical maturity of the platform (capturing voice from many different devices and browsers is not trivial) and the visibility and perenity of the dataset.

4.4. Feedback from the data collection on participants motivations

To gain insight into participants' motivations, the following specific question was included in the first 60 Alsatian prompts of the Spontaneous Speech mode of Common Voice : « Wàrùm màche-n-r dann àn Common Voice mìt? » (Why do you participate in Common Voice ?). At the time of writing this article, 19 out of the 47 participants in the collection had recorded a response to this question.

The decline in the use of Alsatian is explicitly stated by several of them, who also express their regret about it (« ùn dìss ìsch Schàd » : and it's a shame). While most of the respondents consider Alsatian and/or its preservation to be « important » (the adjective « wichtig » was used seven times), it is interesting to observe that their attitudes are complex and ambivalent, and may diverge in some respects. The motivation behind their participation appears to be twofold: firstly, the desire to circumvent a potential loss, and secondly, the aspiration to perpetuate a linguistic tradition that has endured for several centuries, Alsatian having served as the primary medium of intergenerational communication in the region from the Middle Ages until the advent of the 20th century.

The choice of verbs used in the responses

reveals slightly different issues: the use of the verbs *verschwinde* (to disappear), *vergasse* (to forget) and *verliere* (to lose), preceded by the negative adverb *nitt*, indicates above all a desire to preserve the language before it disappears, while the use of verbs such as *bhàlte* (to keep), *preserviera* (to preserve) and *bliewe* (to stay) indicates more of a desire to preserve what one participant compares to a treasure (*Schàtz*). But what seems to be at stake in most of the responses is the transmission of Alsatian to future generations (*nächste Generatione*), as reflected by the frequency of verbs formed with the adverb *widderscht* (further) such as *widderscht gehn* (to go on), *widderscht redde* (to keep speaking) and *widderscht gan* (to pass on). Some participants even consider that the survival of the Alsatian language is at stake, as shown by the use of compound verbs based on the verb *lawe* (to live) : one uses the verb *widderschtlabe* (to keep leaving) while others use the verb *ìwwerläwe* (to survive), as in the following response : « duich de Common Voice kànn s Elsässische vielleicht iwwerläwe » (thanks to Common Voice, the Alsatian language can perhaps survive). However, it is also worth noting the recurring use of modalizers such as *vielleicht* (perhaps), which serve to relativise the hopes expressed.

The importance of the participants' emotional responses and personal commitment is emphasised by the frequent use of possessive adjectives. Three participants use a possessive adjective in the first person singular before the term "mother tongue", which they use to refer to Alsatian (e.g. « ich will zeje wie mini Muetersprooch isch »: I want to show what my mother tongue is), while six other participants use a possessive adjective in the plural before another name for the language (e.g. « fer dàss mr unser Dialekt euh bhàlte kànn »: so that we can keep our dialect), the referent to which the plural refers remaining implicit. With regard to the way they name the Alsatian language, it is noteworthy that only one participant used the term "dialect" to refer to Alsatian, with the majority using *Elsassischi Sproch* or *Elsassisch*. While one participant expressed a strong personal commitment to the preservation of their language, by saying « ich màch, àlles wàs ich kann » (I do everything I can), several other express their willingness to contribute in their own humble way (« wenn ich e bìssel helfe kànn »: if I can help a little). One participant's involvement in the collection appears to serve as a form of compensation for a sense of guilt surrounding the non-transmission of Alsatian, as articulated in his testimony (« do sin mir Eltere d erschte d Schùld dràn »: we, the parents, are responsible for it).

Regardless of the positions expressed by each individual participant, it appears clearly that the motivation behind their participation in the collection was the preservation of the language, rather than a interest in new technologies. This is a crucial aspect that must be considered in future communications related to the Alsatian Spontaneaous Speech data collection.

5. Conclusion

In this study, we launched an Alsatian version of Common Voice to address a lack of voice data in sufficient quantity and quality. Our campaign gathered, at the time of writing, 5 hours and 34 minutes of data, although only 35 minutes have been transcribed. Future work will be focused on transcribing this data, and evaluating the impact of this new data on finetuning an ASR model.

By launching an Alsatian version of Common Voice, we gave the opportunity to Alsatian speakers to play an active role in the preservation of their linguistic heritage within the context of the modern digital world. However, the fact that less than 50 participants took part in the crowdsourcing campaign seems to indicate a lack of interest in the general public in this issue. As has already been shown in previous studies Erhart (2023b), a proportion of Alsatian speakers appear to be resigned to its potential extinction. This suggests a discrepancy between society's evident interest in this language and the actual willingness of speakers to revive its use. It is conceivable that Alsatian has become a post-vernacular language Shandler (2006) in the sense that its importance lies now more in its representation than in its actual use.

In any case, the implementation of this campaign revealed sociolinguistic issues that we had not sufficiently anticipated, in particular a significant gap between the current uses and representations of Alsatian and the use of language technologies, leading to what could be called a « digital code-switching ». This makes us question our role as researchers in the potential revitalization of the languages we work on. If we want to influence speakers' practices and ideologies about languages, and encourage participation in campaigns like the one we launched, effective communication about the research issues is key. We consider that the preservation of linguistic diversity on a global scale is at stake, along with « the possibility to think differently, to produce worlds that do not merely reproduce, in impoverished forms, the codes of a hegemonic culture, but which embody the vitality of multiple cosmologies, to speak languages that are not mere means, but universes, to design technologies that do not standardise the world, but expand it » Baraldi and

Rico Menge (2025).

6. Bibliographical References

Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M. Tyers, and Gregor Weber. 2020. Common voice: A massively-multilingual speech corpus.

Luca Baraldi and Ma. del Carmen Rico Menge. 2025. Vers une nouvelle épistémologie : repenser le pouvoir, la technologie et le langage. *Cahiers du plurilinguisme européen*, 17.

Arlette Bothorel-Witz. 2007. L'Alsace et ses langues. Eléments de description d'une situation sociolinguistique en zone frontalière. Variétés en contact et représentations sociolinguistique [sic]. In *Aspects of Multilingualism in European Border Regions. Insights and Views from Alsace, Eastern Macedonia and Thrace, Lublin Voivodeship and South Tyrol*, pages 39–56. EURAC Research (Europäische Akademie) .

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, Jian Wu, Long Zhou, Shuo Ren, Yanmin Qian, Yao Qian, Jian Wu, Michael Zeng, Xiangzhan Yu, and Furu Wei. 2022. WavLM: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518.

Collectivité européenne d'Alsace. 2022. Étude sociolinguistique sur l'alsacien et l'allemand. Rapport de présentation, mai 2022 .

Pascale Erhart. 2012. *Les dialectes dans les médias : quelle image de l'Alsace véhiculent-ils dans les émissions de la télévision régionale ?* Theses, Université de Strasbourg.

Pascale Erhart. 2017. Les effets de la frontière sur les pratiques linguistiques dans le Rhin Supérieur . *Cahiers du plurilinguisme européen*, (9).

Pascale Erhart. 2023a. « Hett dis kenn nàmme uf elsässisch ? » La néologie en alsacien par le prisme des émissions de la télévision régionale. *Neologica : revue internationale de la néologie*, (17):47–72.

Pascale Erhart. 2023b. L'alsacien ne vaut rien, mais rien ne vaut l'alsacien ! In *Appartenances, marchés et mobilités : penser la valeur des langues*, Sociolinguistique, pages 71–80. L'Harmattan.

Pascale Erhart. 2025. « One does not simply bstell e Flammküeche ohne Ziwwle »: Sociolinguistic Issues of Multilingual Computer Mediated Communication in Alsace. In *Social Media Current Issues in Romance Linguistics and Foreign Language Education*, pages 117–145. AVM.edition.

Pascale Erhart. 2026. Elsässisch 2.0. Dialektgebrauch in der digitalen Kommunikation am Anfang des 21. Jahrhunderts im Elsass. In *Dialekt in Gesellschaft und Schule. Variation und Wandel in Gebrauch und Wahrnehmung des Alemannischen*, pages 103–124. Franz Steiner Verlag.

Lucile Hamm. 2024. Enseignement de l'alsacien / en alsacien à l'école maternelle publique : (nouvelles) perspectives de recherche sur le terrain alsacien ? *Synergies Pays germanophones*, (17).

Harald Hammarström, Robert Forkel, Martin Haspelmath, and Sebastian Bank. 2025. glottolog/glottolog: Glottolog database 5.2.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. HuBERT: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460.

Dominique Huck. 2002. Les dialectes en Alsace : fonctions et statut de l'écrit dialectal et normes graphiques. Etat d'un non-débat. In *Codification des langues de France. Actes du colloque « Les langues de France et leur codification »*, pages 99–110, Paris, Inalco.

Dominique Huck, Arlette Bothorel-Witz, Sylviane Spindler, Ernest Beyer, Raymond Matzen, and Marthe Philipp. 2014. Atlas Linguistique et ethnographique de l'Alsace.

Gil Keren, Artyom Kozhevnikov, Yen Meng, Christophe Ropers, Matthew Setzler, Skyler Wang, Ife Adebara, Michael Auli, Can Balioglu, Kevin Chan, Chierh Cheng, Joe Chuang, Caley Droof, Mark Duppenthaler, Paul-Ambroise Duquenne, Alexander Erben, Cynthia Gao, Gabriel Mejia Gonzalez, Kehan Lyu, Sagar Miglani, Vineel Pratap, Kaushik Ram Sadagopan, Safiyyah Saleem, Arina Turkatenko, Albert Ventayol-Boada, Zheng-Xin Yong, Yu-An Chung, Jean Maillard, Rashel Moritz, Alexandre Mourachko, Mary Williamson, and Shireen Yates. 2025. Omnilingual ASR: Open-source multilingual speech recognition for 1600+ languages. ArXiv preprint arXiv:2511.09690.

Anaïs Koehler. 2024. *L'impact du français sur les parlers dialectaux des locuteurs de l'Alsace contemporaine*. Theses, Université de Strasbourg.

Robert B. Le Page and Andrée Tabouret-Keller. 1985. *Acts of identity: Creole-based approaches to language and ethnicity*. Cambridge University Press.

Michael McAuliffe, Michaela Socolof, Sarah Mihuc, Michael Wagner, and Morgan Sonderegger. 2017. Montreal forced aligner: Trainable text-speech alignment using kaldi. In *Proc. Interspeech 2017*, pages 498–502.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *40th International Conference on Machine Learning*, pages 28492–28518.

Jeffrey Shandler. 2006. *Adventures in Yiddishland. Postvernacular Language and Culture*. University of California Press.

Edgar Zeidler and Danielle Crévenat-Werner. 2008. *Orthographe alsacienne : bien écrire l'alsacien de Wissembourg à Ferrette*. Do Bentzinger.