Getting to Know the Speakers: a Survey of a Non-Standardized Language Digital Use

Alice Millour

Getting to Know the Speakers: a Survey of a Non-Standardized Language Digital Use

Alice Millour

Sorbonne Université, STIH - EA 4509, Paris, France
alice.millour@etu.sorbonne-universite.fr

Abstract

This paper presents the results of an on-line survey regarding the use on the Internet of a less-resourced non-standardized language: Alsatian. The survey, entitled "Alsatian, the Internet, and You" received 1,224 answers in a two months period starting January 2019. The purpose of this survey is twofold. First, we collect generic information on the use of their language by Alsatian speaking Internet users. Second, based on our own experience of crowdsourcing linguistic resources for Alsatian, we use this survey to gather insights on the needs, abilities and expectations of the speakers in order to make the most of their participation.

1. Introduction[1]

Internet practices are evolving fast, and an increasing number of languages are present on the Web. To ensure equity between Internet users and ensure the multilinguality of the cyberspace, language specific resources and processing tools must be developed (Vannini and Le Crosnier, 2012; Rehm and Hegele, 2018). Yet, many languages suffer from a lack of available resources (in a broad sense including linguistic resources, fundings and experts), making such developments challenging.

Since the Internet users community is increasingly multilingual[2], voluntary crowdsourcing (CS) is an interesting option to play upon a valuable source of knowledge: the very speakers of the languages. Not only voluntary CS virtually solves the obstacles mentioned above, it may also enable access to material only the speakers can provide in some cases, for instance when building resources for languages with no consensual spelling system.

The survey we present in this article has been conceived during an ongoing project that aims at empowering speakers to collaboratively produce linguistic resources their language. Two slightly gamified crowdsourcing platforms have been developed: one collects part of speech annotations on existing corpora[3] (Millour and Fort, 2018b), the second aims at overcoming the lack of raw corpus being representative of the language by collecting cooking recipes and dialectal and spelling variants[4] (Millour and Fort, 2018a). These platforms, developed to be easily adapted to any language, have been tested on Alsatian, a French regional language. The results obtained so far are satisfactory in terms of quality, yet insufficient in terms of quantity. This is a common hurdle in CS initiatives, as attracting and retaining participants is known to be challenging (Munro, 2013; Tuite, 2014), and requires good knowledge of the targeted audience needs and expectations.

During a two months period starting January 1st, 2019, we thus interrogated the Alsatian speaking Internet users about their practices of their language on the Internet. The survey, "Alsatian, the Internet, and You", aims at (i) getting a better understanding of how our targeted audience feel about the digitalization of their language and (ii) to understand how we should improve our platforms in terms of specific task design and adequate alternative incentives.

2. Context of the Survey

2.1. Alsatian

Alsatian is a continuum of mostly Alemannic dialects spoken in Alsace and part of Moselle, two Eastern regions of France. It is tagged as "vulnerable" by UNESCO.[5] Although family transmission has known an important decrease in the last decades due to many factors (the most important ones being (i) the French policy to eradicate the so-called "regional languages" after the French Revolution (Perrot, 1997), (ii) the consequently diglossic situation of Alsace, and more specifically (iii) the presence of a non-speaker spouse in the parental couple (OLCA, 2012)), 550,000 speakers were still registered in 2004 by (Barre and Vanderschelden, 2004).

Alsatian presents a great variety of dialectal variants. Additionally, Alsatian does not benefit from a consensual spelling system. Although some interesting flexible orthographies have been developed to match the specific needs of Alsatian (such as the Orthal guidelines (Crévenat-Werner and Zeidler, 2008)), no widespread standard has emerged so far (Erhart, 2018). As has been observed for other languages (see, for instance, (Rivron, 2012; Caulfield, 2013)), this does not prevent some Internet users to freely express themselves in Alsatian on-line, already ensuring a "good Web presence" of the language (Pimienta and Prado, 2014).

[1] We wish to thank our three reviewers for the additional references they provided us with.

[2] See, for instance, the reports provided by `w3tech` such as `https://w3techs.com/technologies/history_overview/content_language/ms/y`.

[3] See: `http://bisame.paris-sorbonne.fr`.

[4] See: `http://bisame.paris-sorbonne.fr/recettes`.

[5] See the group of the Alemannic languages on `http://www.unesco.org/languages-atlas/fr/atlasmap.html`.

2.2. Motivations and Objectives

During our experiments in crowdsourcing linguistic resources for Alsatian, we encountered three main obstacles regarding the participative nature of our projects. They are based on our own perception and on spontaneous feedbacks from some participants:

1. It can be challenging to establish contact with the speakers especially when they are spread across a territory where another language is prevalent.
2. Finding the proper way to advertise on the platform to recruit new participants is hard: being too descriptive is discouraging not to say repulsive, presenting our platforms as playful interfaces is deceiving, leveraging the motivation to participate in a collaborative project that benefits Alsatian is insufficient, and combining these three dimensions is confusing.
3. Even when there exist an *a priori* motivated pool of speakers, keeping their interest and motivation alive in the long run is time consuming and requires the development of adequate incentives.

Additionally, we had no clear idea either of the current use of the Alsatian language on the Internet, or of the expectations of its speakers in terms of digital resources and tools. To understand which actions we should undertake to overcome these obstacles, and to get a better overview of the practices of Alsatian on-line, we thus decided to address directly the speakers through a survey intended to fulfill the following objectives:

- Getting to know the profile of the Alsatian speaking Internet users, with a focus on their relationship with the written form of Alsatian (do they read, write, are at all at ease with written content?).
- Understanding which kind of written content they would like to have at their disposal on-line, and that they would be be eager to share in a context of collaborative corpus production.
- Taking advantage of establishing contact through the survey (more likely to being propagated than information about an academic crowdsourcing platform) to (i) raise awareness among the speakers on the necessity to include their language in the digital world, (ii) advertise about the crowdsourcing platforms existing for their language, (iii) collecting contact e-mails of the speakers that show an interest in the projects we develop.
- Questioning the motivations of the speakers regarding their potential participation to a project aimed at collaboratively building resources for their language.
- Giving the speakers a space to express their opinion about the Internet usability for their language, and making them feel like they are part of our projects development.
- Collecting supplementary feedback on the existing crowdsourcing platforms from the respondents who had already participated.

2.3. Structure of the Survey

The survey has been widely inspired from the studies undertaken by the Digital Language Diversity Project (see for instance (Hicks, 2017), the study carried out for Breton, another French regional language, entitled "Breton — a digital language?"[6]). In order to enable comparison, we kept as such the questions regarding self-evaluation of the language and its digital use. Consequently, the survey is divided in 4 parts: (i) Profile of the respondents, (ii) Self-evaluation of the proficiency in Alsatian, (iii) Opinion on the existing digital tools for Alsatian, (iv) Opinion about crowdsourcing.

2.4. Means of Diffusion

The survey was created on `Framaform`, a free French service respectful of privacy.[7] The responses were thus collected solely on the Web, so we only reached out to respondents that have at least a minimal use of the Internet. We transmitted the survey to the participants of our projects and called upon the official organisms, local radio antennas, and Alsatian speaking Facebook groups to share its link. Even though we did not trace the provenance of the respondents, we can affirm that the most efficient publication was made by a traditional costume shop in Strasbourg[8]; it was shared more than 130 times in a few days. This is interesting to highlight since the official organisms we naturally turned towards to in a first stage appeared to have a far narrower audience.

For obvious reasons, we were not able to conduct a random sampling of the population we targeted. Since we could neither couple our data with social statistics of the Alsatian speaking Internet users, the results we obtain should not be used to infer conclusions about this population.

3. "Alsatian, the Internet, and You"[9]

In this section, we present the analysis of the 1,224 answers we received to our survey. The group of our respondents is unbalanced in terms of gender, 55.1% of them being women. 75% of our respondents have no associative nor professional involvement related to Alsatian preservation. The age repartition, given in Table 1 shows that half our respondents are less than 50 years old. One third of the respondents states their first language is *French*, one third states it is *Alsatian* and the last third states that *both* are[10].

<20	20 to 29	30 to 39	40 to 49	50 to 59	60 to 69	>70	NA
24	176	184	226	278	253	77	3

Table 1: Age repartition of the respondents.

[6] The studies are available at: `http://wp.dldp.eu/reports-on-digital-language-diversity-in-europe/`.

[7] See: `https://framaforms.com`.

[8] The *Geht's in* Shop: see `https://gehts-in.com/`.

[9] The survey was published in French: "*L'Alsacien, Internet, et vous*".

[10] 1.5% chose another language as their first one, for most of them, German.

In the following, we present all the results obtained. For each question, all possible alternatives are presented.

3.1. Debunking Stereotypes

Our efforts on crowdsourcing resources for Alsatian have raised dubious comments about the usefulness of developing NLP tools for such a less-resourced, vulnerable, and non-standardized language. We believe most of these criticisms are based on commonsense questionable ideas, this is why we present part of our results as an answer to the three most widespread stereotypes.

3.1.1. "Alsatian is a Single Patois"

Regional languages tend to be denigrated and considered as unified patois, i.e as substandards of the prevailing language.[11] Yet, to the question "*Which variant(s) do you speak?"*, 41.5% of the respondents answer *Northern low Alemannic*, 25% *Southern low Alemannic*, 19.2% *Strasbourg Alsatian*, 2.5% *Lorraine Franconian, or Plàtt*, 1,7% *Palatine German*, and 1.3% *Other(s)*.[12]

12.5% of the respondents chose at least two answers, and 5% of them answered they did not know the variant they spoke.

Apart from empirically confirming that Alsatian is a generic glottonym for a continuum of dialectal subsystems (Malherbe, 1983), these answers show that the vast majority of speakers identify themselves to their own variant(s) without needing a map to support their answer. This is crucial information that must be taken into account when making use of crowdsourcing for such a multifaceted language. In fact, one should be aware of the necessity to provide content that any member of the targeted community can relate to.

3.1.2. "Alsatian is an Outdated Language Spoken only by Elderly People, not Internet Users"

Self-evaluation of Alsatian was performed by asking the respondents whether they estimate having a *good*, *average*, or *weak* proficiency in (i) *listening*, (ii) *speaking*, (iii) *reading* and (iv) *writing*. We give in Table 2 the proportion of a given age range who evaluated themselves as *good* or *average*: the figures illustrate the drop of language transmission, yet it shows that Alsatian is not absent from the new generations among the respondents.

	Age	<30	30 to 50	50 to 70	>70
Understanding	Good	58%	76%	89%	95%
	Average	31.5%	16%	19%	1%
Speaking	Good	22%	42%	74%	94%
	Average	35.5%	29%	21%	3%
Reading	Good	29%	34%	55%	71%
	Average	31.5%	44%	32%	22%
writing	Good	7%	13%	20%	32%
	Average	16.5%	20%	39%	26%

Table 2: Self-evaluation of the respondents by age range.

Additionally, to the question "*Do you use Alsatian on the Internet (even rarely)?*", 47% of the respondents answer they do, 27.7% of them only to read content, while the majority of them also produces content (articles, publications, comments).

3.1.3. "Alsatian Cannot be Written"

Most regional languages have the reputation of being only spoken. Yet, not only does some literature exist, but the explosion of computer mediated communication has created a fertile ground for the written production of formerly mostly oral languages.

In fact, to the question "*Do you write Alsatian (even rarely)?*", 55.8% of the respondents with either a "good" or "average" speaking proficiency answer they *do*. Among those who do not write it (30.7%), 45% state the reason why is that they would *not know how* to write it, 38.7% state they *do not have the opportunity* to. The rest of them (7% of all the respondents) states Alsatian is a spoken language they *do not want* to write.

The Orthal guidelines (Crévenat-Werner and Zeidler, 2008) provide a comprehensive way to spell Alsatian while respecting its variants. They have existed since 2006. Yet, they are not much used by the speakers who do write Alsatian. In fact, to the question "*When you write, do you follow the Orthal guidelines?*", 8.2% state they *always do*, 10.6% state they *sometimes do*, 7.4% state they would like to, but *do not master it*, 4.6% state they *refuse to use them*, while the majority of the respondents (68,9%) state they had *never heard of it* before.

We thus observe that, although no official guidelines may sanction erroneous writing, part of the Alsatian speaking people (30.7%) repress themselves from writing it. Additionally, these answers showcase that the efforts produced by linguists to compensate this phenomenon remain unknown to the general public.

3.2. Being in Tune with the Speakers

Our previous crowdsourcing experiments, although encouraging in terms of the quality of the collected resources, have not been entirely satisfactory in terms of participation. Yet, in the context of a less-resourced language, which speakers which are aware of the vulnerability and committed to its survival, this lack of interest urged us to engage a dialogue with the on-line community.

Overall, it appears that the respondents have a positive opinion about crowdsourcing. In fact, to the question "*Taking part in the collaborative production of on-line resources for Alsatian seems...*", 6.6% of the respondents answer *...a good idea, I already do!*, while a 28.7% of them states it would be *...too complicated*, either because of their *weak Alsatian proficiency* (75.3%), or because they do *not have the technical computer skills* for it. This leaves us with a 63.5% of the respondents that do not participate despite thinking it is *...a good idea*, either because they *do not know how*, or because they *have no time* to dedicate to it. The rest of the respondents (1.2%) think it is a bad idea.

[11] For a discussion about the term "patois", see (Walter, 2003).

[12] The other variants given by the respondents were: Sundgau Alsatian, high Alemannic, Mulhouse Alsatian, Swiss Alemannic, and village specific variants.

In this section, we present the results of the survey that provide some answers to the difficulties we were confronted to, and that give us an insight on how to make the most of this pool of potential participants.

3.2.1. The Necessary Dialogue

Since crowdsourcing is about involving people into solving a task, a careful attention should be put on its design to ensure its feasibility in the broad sense. Adjusting the design to match the capability of the participants may require engaging a dialogue. We exemplify this point with our experience on crowdsourcing raw corpus production. For a number of languages, the lack of available raw corpus is the very first obstacle for any processing tool development initiative. Alsatian is one of them, this is why one of our crowdsourcing platforms goals is to collect a raw corpus which should be representative of the Alsatian writing practices. Although Pimienta and Prado, 2014 provides a survey of the quantitative presence of Alsatian on-line (e.g. number of blogs, Facebook users etc.), no information about the type of content actually produced was available so far. As a first strategy, we thus decided to crowdsource cooking recipes. This initiative was enthusiastically welcomed by the community, yet in practice, we received a much lower participation than on other *a priori* more complicated tasks such as part of speech tagging.

As a matter of fact, to the question "*When you write Alsatian (off-line or on-line), what do you write (multiple replies are possible)?*", 26% of the respondents answer with *Comments to publications in Alsatian*, 25.2% with *I use it to chat on social networks*, 16% with *Letters or emails*, 14% with *Jokes*, 3.9% with *Literary content*, 3.8% with *Political opinions*, 3.4% with *Informative content (e.g. news, blogs)*, and 2.6% with *cooking recipes*. Apart from showing that the use of Alsatian on-line is mostly conversational, these figures urge us to adapt our platform to match the actual production of the speakers.

Part of the survey was used to collect feedbacks from (i) the respondents who knew about our platforms but did not sign up (9.2%), (ii) the respondents who had signed up to our platforms but did not participate (2.5%), (iii) the respondents who did participate on our platform Bisame (2.9% of the respondents, who represent 60% of the participants). Although this represents a small part of our respondents, we observe that among other reasons such as *I did not want to*, *I did not trust the website*, *It seemed too difficult*, *I could not use the website on my cell phone* etc. the most frequent explanations given by the respondents for either not signing up or participating are *I intended to but I forgot* and *I had no time*.

In fact, the lack of time is the most recurring element when trying to understand what discourages the speakers from participating. On the one hand, this highlights the necessity of (i) designing tasks so that they can be fulfilled in a short amount of time, and (ii) insisting on the participative aspect of the project as a way of distributing the time and the effort among the participants. On the other hand, this is a sign that the incentives we provide are not good enough for the participants to make time for them.

3.2.2. The Right Incentive(s)

All our experiments are based on voluntary crowdsourcing, for ethical reasons and to ensure quality (Fort et al., 2011), and for practical reasons of access to the participants (Callison-Burch and Dredze, 2010).

Although gamification has proved to be an efficient way to incentivize participation for prevalent languages such as English (see, for instance, Phrase Detectives (Chamberlain et al., 2009)) or French (see, for instance, JeuxDeMots (Lafourcade and Joubert, 2008) and ZombiLingo (Guillaume et al., 2016)), our hypothesis was that other types of incentives, maybe more specific to some of the less-resourced languages, deserved to be explored.

To understand which incentive(s) we should (and should not) put efforts in developing, we asked the respondents how they would like to be rewarded for their contribution (multiple replies were possible).

Improving my Alsatian and *Learning things (in general)* were both chosen by more than half of the respondents, followed by *Entertaining myself* (33.5%), *Sharing contents, advices, opinions* (31%), *Bringing out value of my knowledge by participating in a collaborative project* (24.4%), *Reaching out to other Alsatian speaking people* (24%), and eventually *Winning vouchers (e.g. for bookshops, cultural events)* (8.9%). The spontaneous suggestion that were made several times are: *to share my knowledge with learners*, *to promote Alsatian*.

These answers are complemented by the answers given to the question: *Would you like to improve your...* (i) *listening* (38%), (ii) *speaking* (51,4%), (iii) *reading* (45,3%) and (iv) *writing* (57,5%).

Fulfilling this expectation is made difficult by the lack of consensual standard: producing teaching material and evaluation adapted to learners in this context might be challenging. Yet, the high demand from the speakers forces us to explore under which conditions it could be considered as a promising direction.

4. Limitations and Perspectives

We could not take advantage of the part of the survey regarding the opinion of the respondents on the existing digital tools. The aim was to know whether they would like some tools such as "*spell-checkers*" or "*automatic translation tool*" to exist. Yet, the question was formulated in such a way that most of the participants answered that they did not know whether the tools existed or not. Also, we failed at registering how the respondents became aware of the survey, which would been useful to improve our communication strategy.

Finally, the fact that Alsatian is mainly written on-line in a conversational context provides an hypothesis on why participants are reluctant to produce more formally structured content (such as recipes). Nevertheless, the type of content shared in conversational exchanges is not the kind of content we are eager to crowdsource for privacy reasons, so we have yet to figure out how to take advantage of this information.

Our intuition that a survey would be much more widely propagated than our previous communications made through academic or official channels was confirmed, as less than 10% of the respondents had heard of them before they answered the survey. The number of answers we received enabled us to get a credible overview of the practices of Alsatian on-line. Besides, we collected over 500 email addresses from respondents willing to receive future informations on our crowdsourcing projects.

From a crowdsourcing point of view, the answers obtained have highlighted some of the weaknesses of our platforms and have provided interesting insights on the type of features we should put effort on developing in future works. Namely, we should (i) enable the participants to share more diverse types of contents and (ii) reinforce the community feeling within the CS platforms.

The positive additional feedbacks that were provided by close to 10% of the respondents in a free text field encourages us to follow our experiments and to hypothesize the low number of participants we were able to attract so far is partly due to a poor communication strategy. Although we chose to present this work in close relationship with our own experiments in crowdsourcing, we believe part of our conclusions can apply to other less-resourced languages that share some characteristics with Alsatian. We hope our initiative will push the researchers willing to overcome the lack of fundings by using voluntary crowdsourcing to engage fruitful dialogue with their targeted community.

5. References

Barre, Corinne and Mélanie Vanderschelden, 2004. *L'enquête "étude de l'histoire familiale" de 1999 - Résultats détaillés*. Paris: INSEE.

Callison-Burch, Chris and Mark Dredze, 2010. Creating speech and language data with Amazon's Mechanical Turk. In *Proceedings of Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk (CSLDAMT '10) NAACL HLT 2010*. Los Angeles, CA, USA.

Caulfield, John, 2013. *A social network analysis of Irish language use in social media*. Ph.D. thesis, Cardiff University.

Chamberlain, Jonathan, Massimo Poesio, and Udo Kruschwitz, 2009. A new life for a dead parrot: Incentive structures in the phrase detectives game. In *Proceedings of WWW 2009*. Madrid, Spain.

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

Erhart, Pascale, 2018. Les émissions en dialecte de france 3 alsace : des programmes hors normes pour des parlers hors normes ? In *Les Cahiers du GEPE*. Strasbourg : Presses universitaires de Strasbourg.

Fort, Karën, Gilles Adda, and Kevin Bretonnel Cohen, 2011. Amazon Mechanical Turk: Gold mine or coal mine? *Computational Linguistics (editorial)*, 37(2):413–420.

Guillaume, Bruno, Karën Fort, and Nicolas Lefebvre, 2016. Crowdsourcing complex language resources: Playing to annotate dependency syntax. In *Proceedings of International Conference on Computational Linguistics (COLING)*. Osaka, Japan.

Hicks, Davyth, 2017. Breton - a digital language? Technical report, The Digital Language Diversity Project.

Lafourcade, Mathieu and Alain Joubert, 2008. JeuxDeMots : un prototype ludique pour l'émergence de relations entre termes. In *Proceedings of Journées internationales d'Analyse statistique des Données Textuelles (JADT)*. Lyon, France.

Malherbe, Michel, 1983. *Les langages de l'humanité (une encyclopédie des 3000 langues parlées dans le monde)*. Collection Bouquins. Laffont.

Millour, Alice and Karën Fort, 2018a. À l'écoute des locuteurs : production participative de ressources langagières pour des langues non standardisées. In *Revue TAL : numéro spécial sur les langues peu dotées (59-3)*. Association pour le Traitement Automatique des Langues.

Millour, Alice and Karën Fort, 2018b. Toward a Lightweight Solution for Less-resourced Languages: Creating a POS Tagger for Alsatian Using Voluntary Crowdsourcing. In *Proceedings of 11th International Conference on Language Resources and Evaluation (LREC'18)*. Miyazaki, Japan.

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

OLCA, 2012. Etude sur le dialecte alsacien. Technical report, Office pour la Langue et la Culture d'Alsace et de Moselle.

Perrot, Marie-Clémence, 1997. La politique linguistique pendant la révolution française. *Mots. Les langages du politique*, 52(1):158–167.

Pimienta, Daniel and Daniel Prado, 2014. *Étude sur la place des langues de France sur l'Internet*. DGLFLF.

Rehm, Georg and Stefanie Hegele, 2018. Language technology for multilingual europe: An analysis of a large-scale survey regarding challenges, demands, gaps and needs. In *Proceedings of 11th International Conference on Language Resources and Evaluation (LREC)*.

Rivron, Vassili, 2012. L'usage de Facebook chez les Éton du Cameroun. In *Net.lang Réussir le cyberespace multilingue*. Vannini, Laurent and Le Crosnier, Hervé, c&f edition, pages 171–178.

Tuite, Kathleen, 2014. Gwaps: Games with a problem. In *Proceedings of 9th International Conference on the Foundations of Digital Games*. Liberty of the Seas, Caribbean.

Vannini, Laurent and Hervé Le Crosnier, 2012. *Net. Lang: Towards the Multilingual Cyberspace*. C & F Éditions.

Walter, Henriette, 2003. *French inside out: the worldwide development of the French language in the past, the present and the future*. Routledge.