A SPEAKING ATLAS OF MINORITY LANGUAGES OF FRANCE: COLLECTION AND ANALYSES OF DIALECTAL DATA

Philippe Boula de Mareüil, Gilles Adda, Lori Lamel, Albert Rilliard, Frédéric Vernier

LIMSI, CNRS & Univ Paris-Saclay, Orsay, France
{philippe.boula.de.mareuil;gadda;lamel;albert.rilliard;frederic.vernier}@limsi.fr

ABSTRACT

We describe here a speaking atlas that takes the form of a website presenting interactive maps, where it is possible to click on 260 survey points to listen to as many speech samples and read a transcript of what is said, in regional and minority languages of Hexagonal (i.e. Metropolitan) France and its Overseas Territories. We show how an attractive website enables us to collect more data in underresourced and endangered languages and how these data may be used for phonetic analyses and dialectometry purposes. A one-minute story ("The North Wind and the Sun") was used, phonetically transcribed automatically by grapheme-to-phoneme converters and forced aligned with the audio signal: a methodology which can be applied to other languages and dialects.

Keywords: speaking atlas, language documentation, underresourced languages, dialect crowdsourcing.

1. INTRODUCTION

Language documentation is crucial for under-resourced languages which are often endangered. Solutions have been proposed to record, transcribe and translate such languages in India [16], Africa [1], Papua New Guinea and the Amazon [4]. They may be adapted for European minority languages and dialects — the distinction between these two categories being not clear-cut. However, the speakers are often elderly people who may be quite isolated and scattered over a vast territory. We believe that human relationships are essential and that, the interface (web-based or smartphone-or), needs be attractive to collect data that are of interest for phonetic analyses. In this paper, we show that a speaking atlas is well-suited for this purpose.

The first modern linguistic atlas, even if previous attempts at mapping dialects exist, is the Linguistic Atlas of France [9]. Built on the basis of data collected between 1897 and 1900, it did not include audio recordings. Between 1911 and 1914, F. Brunot launched the project of a phonographic linguistic atlas, but it remained in the draft stage, with only three surveys in the Ardennes, Berry and Limousin [12]. Since then, linguistic atlases have been published for various regions of Europe, but with a few exceptions [13][15][3] they are not spoken atlases.

The purpose of this article is twofold. We first describe a speaking atlas which takes the form of a website presenting an interactive map of France, where one can click on more than 260 survey points to listen to speech samples and read a transcript of what is said, in regional or minority languages. The second objective is to show that, elaborating on this base, which follows a traditional approach in dialectology, via the Internet, we were able to collect and analyse new data without necessarily returning to the field.

We recorded the Aesop fable "The north wind and the sun" (one minute of speech), used for over a century by the International Phonetic Association (IPA) to illustrate a number of languages of the world. The main achievements are presented in Section 2. Section 3 shows how they enabled us to acquire more data; it also shows how the recordings gathered can be exploited for phonetic analyses and dialectometry purposes. Section 4 provides a short conclusion.

2. OUTLINE OF THE SPEAKING ATLAS

Launched in 2016, the project *Atlas sonore des langues régionales de France*[i] aimed to highlight the linguistic diversity of France, beginning with the metropolitan area. We recorded the same story (a hundred versions of which can be listened to in different languages and language varieties on the IPA website) in Basque, Breton, Alsatian, Franconian, West Flemish and Romance languages. The atlas has since then been extended to the French Overseas Territories (Caribbean, Pacific and Indian Oceans) as well as non-territorial languages such as Rromani and the French sign language (LSF). With more than 260 survey points represented, around 60 regional and non-territorial languages of France are illustrated, half of them in Oceania.

The homepage of the website, accessible in French and English (https://atlas.limsi.fr), opens with hexagonal France, divided in 25 dialect areas. In addition to the borders of French departments (which delimit the administrative regions), we have included boundaries between linguistic domains. These are eminently more questionable and generally less precise. Any classification being controversial, the

one we propose has no claim to be definitive. It retains the regional languages or primary dialects listed in Figure 1. We have taken up a classical division into Romance languages — *Oïl* (Picard, Gallo, Norman, Mainiot, Angevin, Poitevin-Saintongeais, Berrichon-Bourbonnais, Champenois, Burgundian, Franc-comtois, Lorrain and Walloon), *Oc* (Gascon, Languedocian, Provençal, North-Occitan and *Croissant* 'Crescent'), Catalan, Corsican and Francoprovençal, with particular signage for Ligurian dialects confined to isolated towns like Bonifacian —, Germanic languages (Alsatian, West Flemish, Franconian) and "other languages" (Basque and Breton). The latter three languages, in France, are traditionally subdivided into dialects: Luxembourgish, Mosellan and Rhenish for Franconian, Lapurdian, Lower Navarrese and Souletin for Basque; Trégorois, Léonard, Cornouaillais and Vannetais for Breton. Although each of these ten dialects is represented by at least one survey point, these labels were not shown on the map for scale reasons.

Tabs open maps of the American-Caribbean Zone (Antilles and Guiana), the Indian Ocean (Mayotte and Reunion Island), the Pacific Ocean (New Caledonia and Wallis-and-Futuna, on the one hand, French Polynesia on the other). They can also be accessed directly by clicking inside the rectangles of the world map (https://atlas.limsi.fr/ ?tab=map) which allows navigation from creole to creole. In addition, by checking the appropriate boxes at the bottom of the page, other recordings (and their transcripts) can appear outside of France, in Walloon, Aranese Occitan, Aragonese, Catalan, Asturian, in different Ligurian dialects, Bislama (English-based creole from Vanuatu), Fijian, Latin and even Esperanto. A specific box makes it possible to display, along with the survey points in Belgium, Romand Switzerland and Jersey, the linguistic areas around the corresponding dialects. Another checkbox allows the user to zoom on the Crescentt, in the centre of France, to display survey points that would otherwise be too close to one another at the scale of the French territory, in and around this area — a transition zone between *Oïl* and *Oc* languages which is particularly interesting. In addition, a double orthography has been added for some varieties, in particular (Provençal) Occitan, Berber (in Tifinagh and Latin alphabets) and Arabic dialects. The page "About" enables the visitor to know more about the project with some of our publications [6] and to download the data under a Creative Commons license. Finally, a Search menu allows users to enter a commune name to locate it by a flag on the map.

Whereas, in Hexagonal France and in its periphery (Jersey, Belgium, Switzerland, Italy and Spain), names of localities are displayed, names of languages or language varieties have been reported elsewhere: for example , "Estrian Quebecois" on the world map, "Judeo-Spanish" (in its two Ḥaketía and Djudyó varieties, mapped in Tangier and Thessaloniki, respectively) on the map of non-territorial languages. Judeo-Spanish, which was not mentioned in the Cerquiglini report [8], has since then been added to the list of "non-territorial languages of France", with respect to which the French State acknowledges a patrimonial responsibility. The same applies to LSF, for which, due to its special status, we made an audio-visual recording — from a professional storyteller. The video was "dubbed" in French by a researcher specialised in LSF, who also wrote an explanatory text of a length equivalent to that of the fable (i.e. a hundred words), for educational purposes. In summary, the languages of France we mapped, in addition to the ones listed above for Hexagonal France, are:

- **French-based Creoles**: Guadeloupean, Martinican, Guyanese, Reunion creoles and Tayo (creole of New Caledonia);
- **Nengee languages** (English-based creoles, possibly influenced by Portuguese, of the descendants of slaves taken to Suriname): Aluku, Ndyuka, Pamaka, Saamaka;
- Hmong (an Asian language brought to the French Guiana) and **Indigenous languages of America**: Kali'na, Wayana, Arawak, Palikur, Teko, Wayãpi;
- **Mayotte languages**: Shimaore (Bantu), Kibushi (of Malagasy origin);
- **Kanak languages**: Nyelâyu, Jawe, Nêlêmwa, Zuanga, Pwaamei, Paicî, Ajië, 'orôê, Xârâciùù, Drubea, Numèè, KwényÏ, Iaai, Drehu, Nengone:
- **Polynesian languages**: Faga Uvea, Wallisian, Futunian, Tahitian (including in its Reo Maupiti variety), Pa'umotu (in its Napuka, Tapuhoe, Parata and Maragai varieties), Rurutu, Ra'ivavae, Rapa, Marquesan (in its 'eo 'enana mei Nuku Hiva, 'eo 'enana mei 'Ua Pou, 'eo 'enata varieties), Mangarevan;
- **non-territorial languages of France**: dialectal Arabic (Moroccan, Algerian, Tunisian, Syrian, Palestinian), Berber (Tashlhiyt and Kabyle), Judeo-Spanish (in its Ḥaketía and Djudyó varieties), Yiddish, Western Armenian, Rromani, LSF.

Particular efforts were made to map the eight customary areas of New Caledonia and the five archipelagos of French Polynesia, for which two varieties of Tahitian, three of North Marquesan ('eo 'enana) and South Marquesan ('eo 'enata) and four varieties of Pa'umotu were included. All field recordings, were collected in quiet rooms (in Wave

format, sampled at 44.1 kHz), and great care was given to the orthographic transcription of what was said. Technical problems had to be resolved (e.g. diacritics like macrons, which are important in Polynesian languages to indicate vowel length), in addition to difficulty with translation. It seems that the wealth of our linguistic heritage is of interest to the general public, since in about a year our site has attracted over half a million visits.

3. EXPLOITATION AND EXTENSION OF THE DATA

3.1. Back to the genesis of the speaking atlas

The principal investigator (PI) of this work conducted dozens of field surveys, all over Hexagonal France and the Overseas, to record the speakers who are often elderly people (average age = 60). This work, which revived a very classical approach in dialectology, led to the launch of a first version of the website, with one hundred transcribed recordings from Hexagonal France, in June 2017. The site has enjoyed a great success in print and broadcast media as well as in social networks, with the general public, since it received 300 000 visits over the summer of 2017. Over 200 different people wrote to us, about 60 of them spontaneously sending us their contributions (recordings, orthographic transcripts and signed consents). The recordings have most often been made with smartphones, which today give high quality results. This crowdsourcing dimension was made possible by a snowball effect created by the network of speakers we met (followed up by a few dozen e-mail messages in which the PI thanked them and invited them to disseminate the information) and by the media publicity that ensued. Extended in May 2018 to the many languages of the French Overseas and so-called non-territorial languages, the site continues to be regularly visited and alimented by spontaneous proposals for contributions, especially from Occitania, but also from distant islands.

We have not developed a complicated recording interface; on the other hand, we answered all those who wrote to us, enjoining them to advertise our work, possibly within clubs and associations of dialect speakers. As minority languages continue to unleash passion, the media multiplied communication.

We argue that this special human relationship, even via the Internet or the telephone, is crucial to learn about the local geography and history and leads to very enriching exchanges. Although we are in contact with a wide range of people, a particular profile emerges: we often deal with retired males who, without necessarily being "local scholars", mostly come from the rural world and have benefited from some social ascent (for example through teaching), who have thought about their dialects and are attached to them, knowing that their way of speaking will probably disappear with them or their generation. A bias remains: we mainly are in contact with people who are connected to the Internet, but we try to meet other categories of speakers.

3.2. First phonetic/phonological observations

Regardless of these methodological aspects related to the special session of the present conference, observations can be made at the phonetic/ phonological level, for the collected data. Here we limit the discussion to Hexagonal France.

The Latin /k/ has been maintained in Corsican, Catalan, South-Occitan as well as in Norman, to the north of what is known as the Joret line [11] and in Picard, resulting in forms such as *recauf(f)é* 'warmed'. This /k/ evolved to [ʃ] around central dialects (resulting in forms like *réchauffé*), into an interdental [θ] (noted <sh>) in Francoprovençal, resulting in forms like *reshodô,* and is affricated in [ts] in North-Occitan as in part of Burgundy, resulting in forms like *rétsindu*. Perceptively very salient, an [h] can be perceived not only in non-Romance languages but also in most varieties of Norman, where the corresponding phoneme is noted <h> (e.g. *hardes* 'cloak'), Poitevin-Saintongeais where the /h/ phoneme, noted <jh>, corresponds to French /ʒ/ (e.g. *voeyajheur* 'traveller') and Gascon, where the corresponding phoneme, noted <h>, comes from the laryngealisation of Latin /f/ (e.g. *hòrt* 'strong'). Similarly, a [ç] is perceivable in Burgundian Bresse transcribed as <sc'> (e.g. *sousc'iller* 'to blow'), in Franche-Comté near the Swiss border, transcribed <çh> (e.g. *çhioûçhaie* 'to blow'), in Poitou, also transcribed <çh> (e.g. the demonstrative *çhàu* ''that') and in Alsace bossue, noted <ch> as in German, even if the dialect spoken in our survey point of Wolfskirchen is not Alemannic but Rhenish Franconian. An [x] coming from different origins can also be heard in Breton and in Germanic languages and in dialects like Alsatian, also transcribed <ch> after a vowel like <o>, as well as in Lorrain, transcribed <hh>, in the Vosges.

An apical [ɾ] can be heard throughout the South, as well as in Burgundy and Maine [14]. Also, in Occitan, betacism (confusion between [b] and [β] or [v]) is found in Gascon, Languedocian and sporadically elsewhere, and is shared with Catalan (e.g. *arribaria* 'would arrive'). Rhotacism is found in Gascon, North-Occitan and sporadically elsewhere in the word *sorelh* 'sun', where it can also affect the final consonant /ʎ/, especially in Auvergne.

These remarks illustrate that some consonants (/h/, / θ/, /ç/, /x/, /ɾ/,/ʎ/, /β/) are missing in French, even though they may be matched with French units (e.g. /ɾ/ with /ʁ/, /ʎ/ with /lj/, /β/ with /b/ or /v/), as was done in [17]. Conversely, at the phonological level, the vowels /œ/ and /ø/, which are found in all the *Oïl* area, do not belong to the Occitan system, except perhaps in Eastern Languedocian and in the North of the domain with the Croissant. The above mentioned examples are among the issues which must be taken into account when considering the phonetic transcription of these varieties.

3.3. Towards automatic phonetic transcription and dialectometry

The phonetic specificities of the languages of France listed in the previous section, not to mention the Overseas and non-territorial languages, illustrated that the phonetic transcription task is not an easy one. However, as a first step in this direction, forced alignment experiments using French acoustic models with some adjustments were conducted for 140 survey points in Hexagonal France (of which 60 in the Occitan domain). The corresponding word list, composed of 4000 words, was transcribed using grapheme-to-phoneme converters developed for French and Occitan [5], the pronunciations of which were proposed as pronunciation variants to the alignment system. The acoustic models used were context-indepdendent phone models trained for French [2]: they represent a set of 35 phones, with 3 extra units for non-speech events (silence, breath and fillers). The breath model was used as a pronunciation variant for /h/.

The outputs of the alignment system will certainly have to be corrected, but as the boundaries between units are positioned quite accurately, the process will be sped up. Pending manual checking, dialectometry experiments were conducted, based on a simple inter-symbol edit distance, as in [10]. The results make sense insofar as the clusters that are obtained are consistent with expected divisions between non-Romance languages and Romance languages and, within the latter, divisions between *Oïl* and *Oc* languages, for which the Gascon dialect departs from the other Occitan dialects [7]. These promising outcomes encourage us to continue this work with the study of all our investigation points.

4. CONCLUSION

The speaking atlas presented here shows the richness of our linguistic heritage. It allows us to appreciate its diversity, directly and on a comparable basis (well known by phoneticians), which seems to incite great interest by the general public and not only specialists. The attractiveness of the website and the maps is probably no stranger to this success. We intend to continue this work and provide a phonetic transcription of all the recorded data. Thus, we will show that the same methodology can be applied to collect, document and analyse minority, low-resourced and endangered languages.

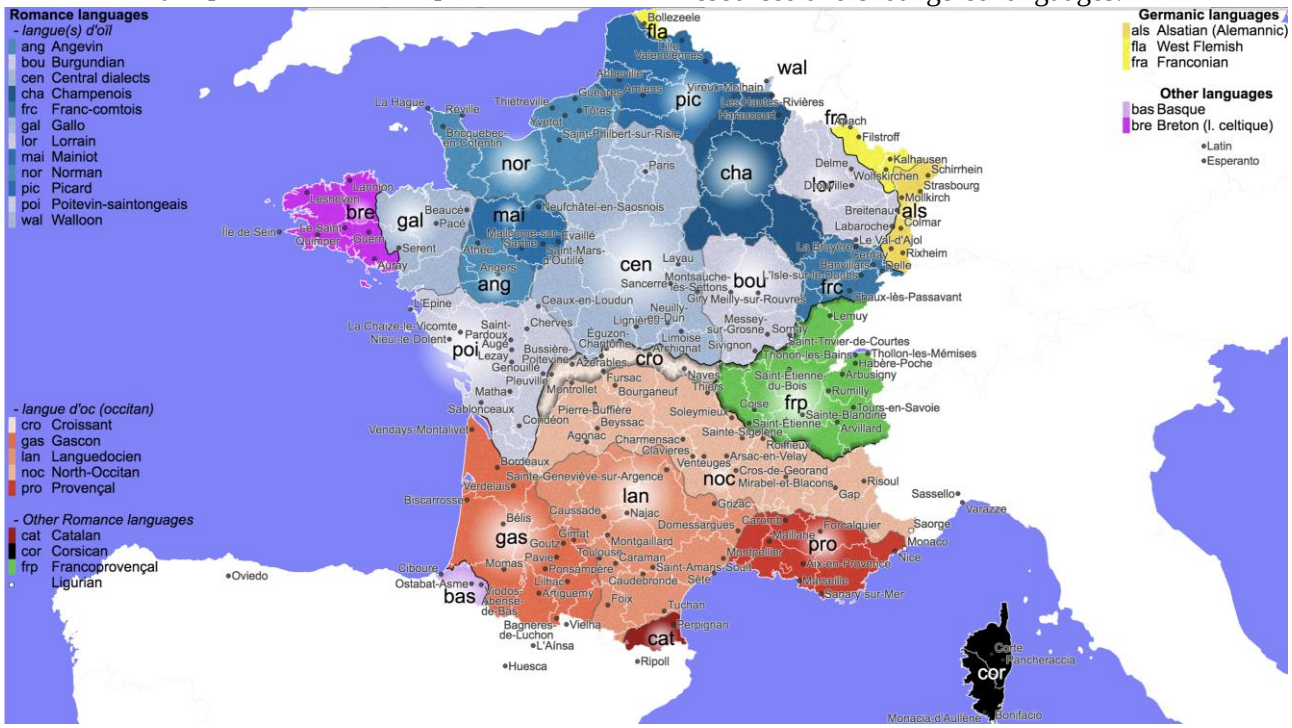

Figure 1: map of Hexagonal France. Warm colours (in the reds) were chosen for *Oc* varieties, cold colours (in the blues) for the *Oïl* varieties, while shades of yellow for Germanic languages.

5. REFERENCES

[1] Adda, G., Stüker, S., Adda-Decker, M., Ambouroue, O., Besacier, L., Blachon, D., Bonneau-Maynard, H., Godard, P., Hamlaoui, H., Idiatov, D., Kouarata, G.-N., Lamel, L., Makasso, E.-M., Rialland, A., Van de Velde, M., Yvon, F., Zerbian, S. 2016. Breaking the Unwritten Language Barrier: The BULB Project, *Procedia Computer Science*, 81, 8–14.

[2] Adda-Decker, M. & Lamel, L. 1999. Pronunciation variants across system configuration, language and speaking style, *Speech Communication*, 29, 83–98.

[3] Almberg, J. & Skarbø, K. 2002. Nordavinden og sola. Ein norsk dialektdatabase pånettet http://www.ling.hf.ntnu.no/nos. In: I. Moen, H.G. Simonsen, A. Torp & K. I. Vannebo (eds.), *Utvalgte artikler fra Det niende møtet om norsk språk.* Oslo: Novus Forlag.

[4] Bird, S. Hanke, F. R., Adams, O., Lee, H. 2014. Aikuma: A mobile app for collaborative language documentation, *Proc. Workshop on the Use of Computational Methods in the Study of Endangered Languages, pages*, Baltimore, 1–5.

[5] Boula de Mareüil, P. 1997. *Étude linguistique appliquée à la synthèse de la parole à partir du texte*. PhD thesis, Univ. Paris-Sud, Orsay.

[6] Boula de Mareüil, P., Vernier, F., Rilliard, A. 2018. A Speaking Atlas of the Regional Languages of France, *Proc. 11th International Conference on Language Resources and Evaluation*, Miyazaki, 4133–4138.

[7] Boula de Mareüil, P., Sichel-Bazin, R., Quint, N., Adda, G. 2017. Norme et variation à l'âge des corpus informatisés pour les langues régionales de France. In: C. Feuillard (ed.), *Usage, norme et codification : de la diversité des situations à l'utilisation du numérique*. Brussels: EME Éditions, 217–222.

[8] Cerquiglini, B. 1999. Rapport au Ministre de l'Éducation Nationale, de la Recherche et de la Technologie, et à la Ministre de la Culture et de la Communication. http://www.ladocumentationfrancaise.fr/var/storage/rapports-publics/994000719.pdf

[9] Gilliéron, J. & Edmont, E. 1902–1910. *Atlas linguistique de la France*. Paris: Champion.

[10] Heeringa, W. 2004. *Measuring dialect pronunciation differences using Levenstein distance.* PhD thesis, Rijksuniversiteit, Groningen.

[11] Joret, C. 1881. *Essai sur le patois normand du Bessin: suivi d'un dictionnaire étymologique*. Paris: F. Vieweg.

[12] Léonard, J. L. 2016. La valorisation des données dialectales d'oïl du liseré frontalier wallon recueillies par la mission Ferdinand Brunot en 1912 : enjeux pour la documentation des langues en danger. *Diachroniques*, 6, 87–120.

[13] Médélice, J. É., 2008. Présentation du projet de l'Atlas Linguistique Multimédia de la Région Rhône-Alpes et des zones limitrophes (ALMURA) et commentaires du poster. In: G. Raimondi & L. Revelli, *Dove va la dialettologia?* Alessandria: Edizioni dell'Orso, 199–205.

[14] Premat, T. & Boula de Mareüil, P. 2018. Le /R/ "roulé" en français et dans quelques langues régionales de France. *Proc. 32es Journées d'Études sur la Parole*, Aix-en-Provence, 55–63.

[15] Romano, A. 2016. La *BD AMPER, La tramontana e il sole* e altri dati su lingue, dialetti, socioletti, etnoletti e interletti del Laboratorio di Fonetica Sperimentale "Arturo Genre". *Quaderni del Museo delle Genti d'Abruzzo*, 41, 225–240.

[16] Srivastava, B.M.L., Sitaram, S., Kumar Mehta, R., Doss Mohan, K., Matani, P., Satpal, S., Bali, K., Srikanth, R., Nayak, N. 2018. Interspeech 2018 Low Resource Automatic Speech Recognition Challenge for Indian Languages. *Proc. 6th Intl. Workshop on Spoken Language Technologies for Under-Resourced Languages*, Gurugram, 11–14.

[17] Vieru, B., Boula de Mareüil, P., Adda-Decker, M.. 2011. Characterisation and identification of non-native French accents. *Speech Communication*, 53, 292–310.

[i] This research is part of the following projects: DGLFLF "Langues et Numérique 2017", ANR-17-CE27-0001-01 ("The linguistic Crescent: a multidisciplinary approach to a contact area between Oc and Oïl varieties") and ANR-10-LABX-0083 ("Investissements d'Avenir", Labex EFL, Axe 3, Opération LC4 – "Les parlers du Croissant : une aire de contact entre oc et oïl"). We are grateful to the Académie des langues kanak, to Jacques Vernaudon and to all those who have agreed to give us their time and lend their voices to this achievement.