A Speaking Atlas of the Regional Languages of France

Philippe Boula de Mareüil¹, Frédéric Vernier¹, Albert Rilliard^{1,2}

¹LIMSI, CNRS & Univ. Paris-Saclay, Orsay, France; ² Univ. Federal de Rio de Janeiro, Brazil {philippe.boula.de.mareuil, frederic.vernier, albert.rilliard}@limsi.fr

Abstract

The aim is to show and promote the linguistic diversity of France, through field recordings, a computer program (which allows us to visualise dialectal areas) and an orthographic transcription (which represents an object of research in itself). A website is presented (https://atlas.limsi.fr), displaying an interactive map of France from which Aesop's fable "The North Wind and the Sun" can be listened to and read in French and in 140 varieties of regional languages. There is thus both a scientific dimension and a heritage dimension in this work, insofar as a number of regional or minority languages are in a critical situation.

Keywords: geolinguistics, dialectology, speaking atlas

1. Introduction

Even if the modern Western world appears to be dominated by just a few widespread languages dialectologists in the field quickly observe a great deal of diversity. The idea of reporting this diversity on maps is not novel (Le Dû et al., 2005): from 1897 to 1901, E. Edmont traveled through France and its Gallo-Romance outskirts to carry out surveys in over 600 communes, mapped in the Atlas Linguistique de la France [ALF] (Gilliéron & Edmont, 1902-1910). Since the Second World War, the Atlas linguistiques de la France par régions (Séguy, 1973a; Tuaillon, 1976) have continued to document increasingly threatened languages and dialects - the limit between languages and dialects being ill-defined (Sibille, 2010). Recently, audio recordings have been digitised, at least for the Occitan (or Oc, Southern Gallo-Romance) domain, as part of the Thesaurus Occitan [Thesoc] (Sauzet & Brun-Trigaud, 2013) and the Francoprovençal domain, with the Atlas Multimédia de la région Rhône-Alpes [ALMuRA] (Médélice, 2008), but they are not entirely available and the Oil (i.e., Northern Gallo-Romance) domain is even more under-resourced (Léonard & Djordjević, 2009). Following the principle of paper dialectological atlases and benefiting from computer technology, the Corpus de la parole now gives online access to an audio corpus of great wealth (Jacobson & Baude, 2011). However, comparable data are sorely lacking, and despite national projects such as SyMiLa (http://symila.univtlse2.fr/), concerning syntactic variation, research on dialectal variation in France is somewhat lagging behind what is done in the Netherlands and Norway (Heeringa, 2004), Germany and Italy (Iannàccaro & Dell'Aquila, 2001), for example. The dominant position of the French language masks a linguistic diversity among the most exemplary of Europe, which the European Charter for Regional or Minority Languages (signed by France in 1999 but not ratified) proposed to promote. Similarly, Unesco insists on the need for multilingualism in cyberspace (Vannin et al., 2012); and general public-oriented sites are multiplying, offering people to record themselves all over the globe and asking social networks to vote for this or that recording, but without linguistic control. It is therefore important, even urgent, to collect recordings using a coordinated approach, applying a common protocol to give a better picture of the plurality of uses.

A few decades ago, it would have been easier to find speakers of dialects and regional languages in France.

However, the Internet, which now makes it possible to contact a number of associations for the protection and promotion of minority language, was less developed at that time, as were the means of collecting and storing large quantities of recordings. Since most of these dialects and languages are now endangered, we describe here a speaking linguistic atlas that aims to preserve them. This atlas takes the form of a website presenting an interactive map of France (for the moment limited to the metropolis, although recordings were also collected in the overseas), with over a hundred survey points on which one can click to hear as many speech samples and read a transcript of what is said. The orthographic transcription issue for the regional languages of France continues to be debated: particular attention was thus paid to this problem. In the absence of well-established written traditions and a

single authority, the regional languages of France do not have unanimously recognised and accepted standards (Caubet et al., 2002; Vannin et al., 2012). Consequently, the transcription solutions proposed vary from one language to another, even among the territorial languages of France, the spellings of which are based on the Latin alphabet and are meant to be close to the pronunciation (at least to a minimum). The orthographies adopted are more or less phonetic (reflecting a particular local pronunciation) or diasystemic (emphasising the unity of a set of dialects). Sometimes, the system is hybrid, for the sake of efficiency, noting what in pronunciation differs from French, while following the orthographic conventions of French. These systems have their advantages and disadvantages.

These different types of writing were tested through a case study: the transcription of the same text of a hundred words, which we will begin by presenting. We will then describe the surveys we conducted, the speakers whose recordings were collected and transcribed, the protocol we applied, and the mapping we established. We will present quantitative data at the conference.

2. Survey Points and Speakers

The material we have focused on consists of Aesop's fable "The North Wind and the Sun" (120 words in French, about 1 mn of speech), translated and recorded in more than 140 investigation points in as many varieties of regional languages of France. This text (used by the International Phonetic Association to describe a large number of languages and dialects of the world) was a way of reviving a 19th century tradition consisting of

translating the parable of the prodigal son. The corpus collected resulted in a digital database covering Romance languages (such as Occitan, Francoprovençal, Catalan and Corsican), Germanic languages (Alsatian, Franconian and West Flemish), Breton and Basque.

In order to meet these speakers, we contacted universities and local radio stations (if any), associations and public offices promoting regional languages and clubs of *patois* speakers, city halls, tourist offices and immersive schools (*diwan, calandretas*). We sent hundreds of e-mails and made hundreds of phone calls. A considerable amount of time was spent finding speakers of regional languages who are now a small minority, organising and optimising travel. This research required extensive documentation not only on dialects but also on the history and geography of the regions explored. In 2014–2016, we conducted about thirty field surveys, with a few days per mission, seeking to maximise the number of speakers per locality.

We then selected one speaker per locality, retaining only one commune within a radius of less than 20 km. We favoured native speakers, but have not disregarded neospeakers, in cities like Bordeaux or Nice. In some cases, there was little choice; in others, one of the speakers stood out for his/her oratory talent, the aesthetic nature of his/her voice or his/her local anchorage. The distinction between "neo-speakers" and "natural speakers" is not always obvious: there are native speakers of neo-Breton, for example; so-called "natural" speakers may have a highly Frenchified regional language; on the contrary, some speakers who have not inherited the dialect directly from their parents can commit to successfully replicating local particularities. When a speaker was born and has always lived in the same place, associating his/her recording with a point on the map is easy, but modern life encourages mobility. When a speaker explicitly said (s)he spoke the dialect of a place that was not his/her place of residence, we kept this place to associate it to him/her --usually the commune where the speaker had grown up.

The speakers, rather engaged in the cultural and linguistic field, were from varied socio-professional backgrounds. A typical profile was the following: a retired man from the farmer world (or workers' world, in the North and the East), having experienced social climbing, having worked for instance as a teacher. It should be noted that in the ALF, too, E. Edmont investigated mainly educated men, on average older in the North than in the South. All those who responded positively gave us an extremely warm welcome, expressing the moving feeling that their patois was going to disappear with them, in the Oïl domain and the Croissant (intermediate zone between Oil and Oc domains, so named because of its crescent shape). Without necessarily being "local scholars", they generally showed a very thorough knowledge of the culture, history and geography of their region, making the exchanges very enriching. Most of them were recorded at home or in a meeting room with a Zoom H4 recorder in Wave format (in stereo at a sampling frequency of 44.1 kHz and 16-bit quantisation, before intensity equalisation and conversion to MP3 for uploading). Each recording was associated with a signed consent for free distribution.

3. Protocol and Transcription

A common protocol was applied, in which the speakers were first asked to read the fable in French. They were then asked to translate this text into their regional language, either directly with the French text in front of them, or from notes they had preferred to take. Some speakers (especially in the Basque Country, Brittany, West Flanders and the Oil domain) wanted to write their translations entirely. This situation occurred either because the regional language and French were very different in their structures (with a different word order, as in Breton), or because, on the contrary, the regional language was close to French, as in the Oil domain, and the speakers needed "to get back into the swing of things". French, in France, dominates so much in everyday interactions and public space that the translation task is not easy. In Corsican, Catalan, Occitan, Francoprovençal, Alsatian and sometimes in the Oil domain, the speakers we recorded succeeded very well in translating the Aesop fable on the fly. A word like voyageur 'traveler' could pose difficulties, because at the time patois was spoken, some informants said, people did not use to travel). Still, translating the fables of La Fontaine or others was an exercise to which many of our informants were accustomed. Sometimes the speakers moved away from a literal translation to get closer to their oral traditions. These different translation strategies are also a testimony of the wealth and diversity of linguistic ingenuity.

For Occitan (*langue d'oc*), we transcribed the recordings into classical spelling (Alibert, 1935), sometimes drawing on what the speakers had provided us. We had the transcriptions checked and corrected by specialists who had access to the audio, as was the case for the Catalan and Corsican transcriptions we had produced.

The unified spellings of Basque (batua) and Breton (peurunvan) were well known by our Basque and Breton speakers, respectively. For Alsatian, a flexible system close to ORTHAL (Zeidler & Crevenat-Werner, 2008) was used, with written German (Schriftsprache) as a respecting dialectal peculiarities. For reference. Luxembourgish Franconian (spoken in our Apach survey point), the Luxembourg standard was followed, whereas for Moselle and Rhine dialects other systems were adopted, also inspired by modern German, but using the 'æ' digraph and grave accents to indicate/overcharacterise the open quality of some vowels, as well as consonants such as 'd' to note dull non-aspirated stop consonants, like [t] (Hudlett, 2004).

In the *Oïl* domain, we sometimes face individual spellings, such as in Berrichon (among central dialects), Lorrain or Walloon, for which in France there is no real standard. This also applies to three survey points in the linguistic Croissant (Pleuville, Bussière-Poitevine and Éguzon-Chantôme), for which a French-like spelling was adopted. The writer thus encodes words as he wants them to be heard, respecting his own phonetics. More often than not, as proposed by Contejean (1876), it is a Frenchified spelling that is applied by our *Oïl* informants, with the following rules:

- preservation of the plural mark for nouns, pronouns and adjectives (in general s);
- preservation of the person mark for conjugation;

- a concern for clarity, avoiding accumulating elisions through apostrophes (markers of an orality considered inherent in *patois*).

The Feller-Carton system, for Picard (Carton, 2009), also proposes two principles, with subregion-dependent rules to solve cases where these two principles are in contradiction: (1) priority to the French spelling provided that it does not create ambiguity; (2) priority to the phonetics of Picard.

We will see how or if speakers resolved possible contradictions. In all cases, we made sure of the adequacy between what was said and what was transcribed, because it is frequent not to write exactly the words that were spoken and not to read exactly the words that were handwritten or printed. When necessary, we deleted disfluencies in the audio signal, as well as comments like *es pas aisit* ('it is not easy' in Oc-citan) which were not part of the fable.

4. Cartography

The survey points, defined by their latitude and longitude, were arranged on a map of France where we included boundaries between linguistic domains, in addition to the borders between departments which delimit administrative regions. The former are eminently more debatable and generally less abrupt (Paris, 1888; Gilliéron & Mongin, 1905). If one looks at isoglosses drawn according to different criteria, they may not coincide, and it is then difficult to determine which ones are preferable to use. It is also known that the intercomprehension criterion, sometimes put forward in folk linguistics, quickly finds its limits: it can be asymmetrical and may depend on the communicative situation. Since any classification is questionable (Goebl, 2002; Gaillard-Corvaglia et al., 2007, Sumien, 2012), the one we propose has no pretension to be definitive. It retains 25 regional languages or primary dialects, grouped as shown in Fig. 1. We kept a classical partition in Romance languages (Oil, Oc, Catalan, Francoprovençal and Corsican), Germanic languages (Alsatian, West Flemish, Franconian) and "Other languages" (Basque and Breton). The latter languages, on the French territory, are traditionally subdivided into dialects: Luxembourgish, Moselle and Rhine for Franconian, Labourdin, Lower-Navarrese and Souletin for Basque; Trégorrois, Léonard, Cornouaillais and Vannetais for Breton. Although each of these ten dialects is represented by a survey point, we did not include these labels on the map for scale reasons.

It turns out that the majority of associations for the defence and promotion of Oil and Oc languages insist on the plural in the first case and on the singular in the second case, even if Gascon is unquestionably quite distinct from the other Oc varieties (Chambon, & Greub, 2002). These two points of view being difficult to reconcile (Abalain, 2007), out of respect for our informants, we indicated "langue(s) d'oil" and "langue d'oc (occitan)" in the legend. Warm colours (in the reds) were chosen for Oc varieties, cool colours (in the blues) for Oil varieties, while green was retained for

Francoprovençal and shades of yellow were kept for Germanic varieties. Thicker lines emphasise the frontiers between these different domains, while the limits of the Croissant (attached to Occitan), appear blurred to suggest the transitory nature of this linguistic area.

In order to draw these linguistic limits, we classified the 600+ ALF suvey points in our 25 categories, based on various regional atlases and publications: for the Oïl domain (Bourcelot, 1966-1978; Brasseur, 1980-2011; Carton & Lebègue, 1989-1997; Dondaine, 1972-1978; Dubuisson, 1971-1982; Guillaume & Chauveau, 1975-1983; Lanher et al., 1979-1988; Massignon & Horiot, Simoni-Aurembou, 1973–1978; Taverdet, 1971-1983; 1975–1980; Martin, 2015); for the Francoprovençal domain (Gardette, 1952; Tuaillon & Contini, 1996); for the Oc domain (Séguy, 1973a, 1973b; Desrozier & Ros, 1974; Sumien, 2008). The Occitan domain was subdivided, following Tuaillon and Contini (1996) into: Gascon, Languedocien, Provençal, North-Occitan and Croissant. We could have subdivided the North-Occitan subdomain (and the Croissant) into Limousin, Auvergnat and Vivaro-Alpine. However, the more we multiply borders, the more we create problems: speakers from Velay, for example, are difficult to classify. Featuring the Croissant, on the other hand, shows the shape of this set of transitional dialects between Oil and Oc. As for the domain of central dialects of Oil, we have not sought to subdivide it into Tourraine, Orléanais, etc., for lack of speakers of these dialects.

An algorithm was then designed to draw lines passing in the middle of two points of different categories, to smooth out the contours and to colour the zones so defined. The smoothing of the boundaries between dialectal areas, more or less sawtooth-like, can be set by the user, using a slider. To refine the layouts, a few points were added near the linguistic borders of the Basque, Breton, Catalan, Alsatian, Flemish and Franconian domains, as well as at the confluences of the Gascon, Languedocian and North-Occitan areas. In addition, a "Search" field is offered to the user, and options are provided for displaying or not boundaries between departments, ALF points and our survey points, Jersey, French-speaking Switzerland and Belgium on the one hand, Overseas France on the other.

The resulting interactive map, with over 140 points that can be clicked on to hear as many versions of the same text, Aesop's fable "The North Wind and the Sun", is available at https://atlas.limsi.fr (in French and English). Its French translation can be listened to (and read) by clicking on Paris, at that address.

The site makes it possible to immediately appreciate how certain geographically close dialects differ linguistically. This is the case for survey points in Bresse, between Burgundy and Savoy (Francoprovençal), points in Romance (Welche) and German-speaking Alsace, and a fortiori points in the Basque Country and Béarn (Gascon), gathered in the department of Pyrénées-Atlantiques. It is striking (although not surprising) that no dialect speaker was found around Paris: for central dialects, we only found speakers of Berrichon-Bourbonnais. In the East, especially, it is as if dialects had been relegated to Belgian, German and Swiss borders. We found speakers of Champenois and Lorrain almost only in the Ardennes and the Vosges, respectively. The annexation of Moselle by Germany during World War II may have given some respite to dialects (including Romance dialects, which were not prohibited as French was), but the speaker we recorded in that department is today an exception.

5. Conclusion and Future Work

This multimedia atlas provides a kind of colour photograph of the dialectal rainbow of (hexagonal) France: an instantaneous photograph of current uses, a tool for teaching variation, showing that the territorial languages of France, though threatened, are still today a precious reality (D'Hervé, 2005). It does not tell us much about the vitality of these languages, the number of their speakers, the social dynamics that are at work, despite the amount of data collected. Nor does it tell us how rewarding the personal meetings with the informants were. The practice of regional languages has become scarce and is now a minority: the chain of intergenerational transmission within family cells is often interrupted. It was therefore a challenge to give greater visibility to the diversity of the French linguistic landscape.

The problem of speaker representativeness remains a real issue, with respect to a traditional dialectological framework that assumes an irremovable link between the land and the speaker. This framework is more and more outdated in a country dominated by mobility and urbanisation (Mufwene & Vigouroux, 2012). Large cities such as Marseilles, Toulouse or Montpellier necessarily melt many differences, and the influence of French on regional languages is now inevitable. The applied protocol (a translation) can also have an influence, prompting us to take a critical look at the productions so elicited. However, depending on the speakers, the translation can favour both calques and a search for maximum deviation from French.

We wanted to reach linguists and teachers, but also the general public. This objective was achieved in view of the success of our site (over 300,000 visits by the summer of 2017), printed and audiovisual media as well as social networks. The aim was to catch up on the delay taken by France in the valorisation of its minority languages. Associating them to modernity (with innovative cartographic tools to present boundaries between dialectal areas), will hopefully help revitalise our linguistic heritage, or at least give it recognition as a vehicle for creativity, for lack of being able to counteract the decline in minority language practices.

We will continue to give media coverage to this speaking atlas, which we are going to extend to the French Overseas Territories. To go further, it is possible not only to use the web to display research results, but also to collect new information, using a crowdsourcing methodology now very much in fashion. Yet, proposing to complete surveys through the Internet will require some precautions for minority languages.

6. Acknowledgements

This work was largely financed within the framework of the DGLFLF "Langues et Numérique 2016" call for proposals. We thank F. Manni and M. Russo for their cooperation in this project, as well as G. Brun-Trigaud, who provided the latitudes and longitudes of more than 600 ALF points, and K. Biers for comments on an earlier version of this paper. We would like to express our sincere gratitude to all those who have agreed to give us their time and lend their voice to this achievement.



Figure 1: screenshot of the site < https://atlas.limsi.fr/index-en.html>.

7. Bibliographical References

- Abalain, H. (2007). Le français et les langues historiques de la France. Éditions Jean-Paul Gisserot, Paris. Alibert, L. (1935)., Gramatica occitana segon los parlars lengadocians. Societat d'Estudis Occitans, Toulouse.
- Bourcelot, H. (1966–1978). Atlas linguistique et ethnographique de la Champagne et de la Brie. Éditions du CNRS, Paris.
- Brasseur, P. (1980–2011). Atlas linguistique et ethnographique normand, Éditions du CNRS, Paris.
- Carton, F. (2009). Pourquoi et pour qui on transcrit ? Les graphies du picard moderne. *La Linguistique*, 45(1):113–123.
- Carton, F. & Lebègue, M. (1989–1997). Atlas linguistique et ethnographique de la Picardie. Éditions du CNRS, Paris.
- Caubet, D., Chaker, S. & Sibille, J. (2001). *Codification des langues de France*, L'Harmattan, Paris.
- Chambon, J.-P. & Greub, Y. (2002). Note sur l'âge du (proto)gascon. *Revue de linguistique romane*, 66:263–264.
- Contejean, C. (1876). *Glossaire du patois de Montbéliard*. Société d'émulation de Montbéliard, Montbéliard.
- D'Hervé, G. (2005). Le gallo dans l'enseignement, l'enseignement du gallo. *Marges Linguistiques*, 10:262–281.
- Desrozier, P. & Ros, J. (1974)). L'ortografia occitana. Lo lemosin, Centre d'Estudis Occitans, Montpellier.
- Dondaine, C. (1972–1978). Atlas linguistique et ethnographique de la Franche-Comté. Éditions du CNRS, Paris.
- Dubuisson, P. (1971–1982). Atlas linguistique et ethnographique du Centre. Éditions du CNRS, Paris.
- Gaillard-Corvaglia, A., Léonard, J.-L., Darlu, P. (2007). Testing Cladistics on Dialect Networks and Phyla (Gallo-Romance Vowels, Southern Italo-Romance Diasystems and Mayan Languages. Proc. 9th Meeting of the ACL SIG in Computational Morphology and Phonology, Prague (pp. 23–30).
- Gardette, P. (1952). Atlas linguistique et ethnographique du Lyonnais. *Revue de géographie de Lyon*, 27(1): 90–92.
- Gilliéron, J. & Edmont, E. (1902–1910). Atlas linguistique de la France. Champion, Paris.
- Gilliéron, J. & Mongin, J. (1905). Étude de géographie linguistique : « scier » dans la Gaule romane du Sud et de l'Est. H. Champion, Paris.
- Goebl, H. (2002). Analyse dialectométrique des structures de profondeur de l'ALF. *Revue de linguistique romane* 66(261–262): 5–63.
- Guillaume, G. & Chauveau, J.-P. (1975–1983). Atlas linguistique et ethnographique de la Bretagne Romane, de l'Anjou et du Maine. Éditions du CNRS, Paris.
- Heeringa, W. (2004). *Measuring dialect pronunciation differences using Levenstein distance*, PhD thesis, Rijksuniversiteit, Groningen.
- Hudlett, A. (2004). Charte de la graphie harmonisée des parlers franciques platt de la Moselle germanophone. Publications du Centre de Recherche sur l'Europe Littéraire, Mulhouse.

- Iannàccaro, G. & Dell'Aquila, V. (2001). Mapping languages from inside: notes on perceptual dialectlogy. *Social & Cultural Geography*, 2(3): 265–280.
- Jacobson, M. & Baude, O. (2011). *Corpus de la parole* : collecte, catalogage, conservation et diffusion des ressources orales sur le français et les langues de France. *Traitement automatique des langues*, 52(3):47–69.
- Lanher, J., Litaize, A., Richard, J. (1979–1988,). *Atlas linguistique et ethnographique de la Lorraine romane*. Editions du CNRS, Paris.
- Le Dû, J., Le Berre, Y., Brun-Trigaud, G. (2005). Lectures de l'atlas linguistique de la France de Gilliéron et Edmont. Du temps dans l'espace. CTHS, Paris.
- Martin, F. (2015). Espaces & lieux de la langue au XXI^e siècle en Picardie : approche complexe de la structuration des pratiques linguistiques en situations ordinaires – enquêtes en Picardie. PhD thesis, Université de Picardie Jules Verne, Amiens.
- Massignon, G. & Horiot, B. (1971–1983). *Atlas linguistique et ethnographique de l'Ouest*. Editions du CNRS, Paris.
- 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 dialetto-logia?*, Edizioni dell'Orso, Alessandria (pp. 199–205).
- Mufwene, S.S. & Vigouroux, C.B. (2012). Individuals, Populations, and Timespace: Perspectives on the Ecology of Language. *Cahiers de sociolinguistique*, 38(2): 111–137.
- Paris, G. (1888). Les parlers de France. Revue des patois gallo-romans, 17:161–175.
- Séguy, J. (1973a). Les Atlas linguistiques de la France par régions. *Langue française*, 18: 65–90.
- Séguy, J. (1973b). La dialectométrie dans l'Atlas linguistique de la Gascogne. *Revue de linguistique romane*, 37:1–24.
- Sibille, J. (2010). Les langues autochtones de France métropolitaine. Pratiques et savoirs. In C. Gruaz, C. Jacquet-Pfau (Eds.), Autour du mot : pratiques et compétences, Lambert-Lucas (pp. 69–85).
- Simoni-Aurembou, M.-R. (1973–1978). Atlas linguistique et ethnographique de l'Île-de-France et de l'Orléanais, Éditions du CNRS, Paris.
- Sumien, D. (2008). Classificacion dei dialèctes occitans. *Linguistica Occitana*, 7 : 1–55.
- Sumien, D. (2012). Le catalogue des langues romanes : clarifier les critères et les idéologies. *Revue des langues romanes*, 116(1):5–33.
- Taverdet, G. (1975–1980). Atlas linguistique et ethnographique de la Bourgogne. Éditions du CNRS, Paris.
- Tuaillon, G. (1976). Comportement de recherche en dialectologie française. Éditions du CNRS, Paris.
- Tuaillon, G. & Contini, M. (1996). *Atlas Linguistique Roman (ALiR)*. Istituto poligrafico e zecca dello stato, Rome.
- Vannin, L. & Le Crosnier, H. (2012). Net.lang. Réussir le cyberespace numérique. C1F Éditions, Caen.

Zeidler, E. & Crévenat-Werner, D. (2008). Orthographe alsacienne : bien écrire l'alsacien de Wissembourg à ferrette. Jjérôme Do Bentzinger Colmar.