Proceedings of the Language Technologies for All (LT4All) , pages 155–159
Paris, UNESCO Headquarters, 5-6 December, 2019.

A Speaking Atlas of Indigenous Languages of France and its Overseas

Philippe Boula de Mareüil[1], Frédéric Vernier[1], Gilles Adda[1], Albert Rilliard[1], lacques Vernaudon[2]

[1]LIMSI, CNRS & Université Paris-Saclay, Orsay, France
[2] EASTCO, Université de la Polynésie française, Faaa, PF
{philippe.boula.de.mareuil, frederic.vernier, gilles.adda, albert.rilliard}@limsi.fr
jacques.vernaudon@upf.pf

Abstract

The objective of this work is to show and valorise the linguistic diversity of France through recordings collected in the field, a computer interface (which allows viewing the dialect areas) and a work of orthographic transcription. We briefly describe a website (https://atlas.limsi.fr) presenting interactive maps of France and its overseas territories, from which the Aesop fable "The North Wind and the Sun" can be listened to and read in over 300 versions, in indigenous languages. There is thus a scientific and heritage dimension in this work, insofar as a number of regional or minority languages are in a critical situation.

Keywords: geolinguistics, dialectology, speaking atlas, indigenous languages

Résumé

L'objectif de ce travail est de montrer et de valoriser la diversité linguistique de la France à travers des enregistrements recueillis sur le terrain, une réalisation informatique (qui permet de visualiser les aires dialectales) et un travail de transcription orthographique. Nous décrivons ici un site web (https://atlas.limsi.fr) présentant des cartes interactives de France hexagonale et des Outre-mer, à partir desquelles la fable d'Ésope « La bise et le soleil » peut être écoutée et lue dans plus de 300 versions, en langues locales. Il y a ainsi une dimension à la fois scientifique et patrimoniale à ce travail, dans la mesure où un certain nombre de langues régionales ou minoritaires sont en situation critique.

1. Introduction

Language allows us to communicate but also to reflect our identity. A uniform language where, as in the mythical Tower of Babel, everyone would designate a brick or mortar in the same way, without metaphors, without ambiguity or affectivity, without polysemy, would undoubtedly have a technical utility. The argument that, if we all spoke French or English (or Globish), understanding would be easier, is difficult to counter. It was already in place in French revolutionary Abbé Grégoire (1794), in his report on the need and means of annihilating the "patois", from which we will quote an extract (without translating it).

> Proposerez-vous […] des traductions ? Alors vous multipliez les dépenses […]. Ajoutons que la majeure partie des dialectes vulgaires résistent à la traduction, […] les uns […] sont absolument dénués de termes relatifs à la politique ; les autres sont des jargons lourds et grossiers, sans syntaxe déterminée, parce que la langue est toujours la mesure du génie d'un peuple.

Today, fortunately, we no longer express ourselves in these terms; we are rather saddened by the death of languages, as well as by the disappearance of animal species. We will not push this parallel made by Hagège (2002) with living species too far, because language is above all a social construction. However, a world with no regional particularities would be boring and lack poetry; it would lack flavour like a meal without salt, without pepper. Nowadays, most people are attached to the diversity of languages, which all teach us something about Mankind. As Cyrulnik (2019) says:

> Il faut que le langage soit énigmatique afin de laisser place à l'interprétation. Un langage précis ne serait que désignation, signal de la chose, sans vie émotionnelle, sans vibration, juste une information pour déclencher la réponse.

The question of the linguistic norm quickly arises, as soon as one is interested in the diversity of languages and variation within languages. We can distinguish at least two types of norms: a statistical norm (objective, established by observable and quantifiable facts) and a prescriptive norm (subjective, which indicates a model promoted as "correct"), a cultural construct made up of social choices (Canguilhem, 1966; Gadet, 2003; Rastier, 2007). For many minority languages, in the absence of a prescriptive norm accepted by all, a great deal of variation comes into play. This is understandable, but it is a difficulty we quickly encounter when we go and do field work, for example asking speakers to translate the fable attributed to Aesop "The North Wind and the Sun", a text which has been used for over a century by the International Phonetic Association (IPA) — this story can be accessed in a hundred versions on the IPA website.

In the following, we elaborate on this fable, in order to highlight and promote linguistic diversity: we describe a speaking atlas which takes the form of a website presenting interactive maps of France and its Overseas Territories, where one can click on more than 300 survey points to listen to translations of this text, in regional or minority languages, and read a transcript of what is said. In total, between 2016 and 2018, around 60 languages were collected, half of them in Oceania. These languages do not have unanimously recognised and accepted standards (Caubet et al., 2002). Thus, 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 some extent). 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 interests of efficiency, noting what in pronunciation differs from French, while following the orthographic conventions of French. These systems, which have their advantages and disadvantages, are illustrated in the languages addressed here. Fixing the spelling of an indigenous language is an object of research per se, which may not be enough to reverse language shift (Fishman, 1991). However, in this rapidly changing world, in our society where computer- and smartphone-mediated communication holds a preponderant place, we believe that there is hardly any other way for the survival of many minority languages than learning how to write them.

2. Materials

2.1. Languages Collected

Aesop's fable (120 words in French, about 1 minute of speech) was recorded in Basque, Breton, Alemanic Alsatian, Franconian, West Flemish and Romance languages (Boula de Mareüil et al., 2018: see also references therein). It was then translated in the many languages of the French Overseas Territories (Caribbean, Pacific and Indian Oceans): the languages of Guiana (Goury and Migge, 2003; Léglise and Migge, 2007; Biswana, 2016), Creoles (Cayol et al., 1984–1995; Bernabé, 2005; Le Dû and Brun-Trigaud, 2011; Corn, 1990; Ehrhart, 1993), Kanak languages (Rivierre, 1979; Wacalie, 2013; Moyse-Faurie, 2014), Polynesian languages (Lazard and Peltzer, 2000; Vernaudon, 2017) and the languages of Mayotte (Laroussi, 2010). It was also translated in the so-called non-territorial languages of France such as R(r)omani (Courthiade, 2007, 2013) and the French sign language (LSF), with respect to which the French State acknowledges a patrimonial responsibility (Cerquiglini, 1999).

The website (https://atlas.limsi.fr) opens with Hexagonal (i.e. Metropolitan) France, divided in the 25 dialect areas listed below and in Figure 1. Another tab opens a map of the world, which allows navigation from creole to creole and gives access, by clicking inside various rectangles, to additional maps: 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). Also, a tab opens a map of the Euro-Mediterranean area, with the non-territorial languages of France. In summary, the languages collected are:

- **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 (with its Luxembourgish, Mosellan and Rhenish dialects);
- **Breton** (a Celtic language, with its Trégorois, Léonard, Cornouaillais and Vannetais dialects);
- **Basque** (a linguistic isolate, with its Lapurdian, Lower Navarrese and Souletin dialects);
- **French-based creoles**: Guadeloupean, Martinican, Guianese creoles, Reunion creole (from the lowlands and the highlands) and Tayo (the creole of New Caledonia);
- **Nengee languages** (English-based creoles, possibly influenced by Portuguese, of the descendants of Maroons who escaped from slavery in Suriname): Aluku, Ndyuka, Pamaka, Saamaka;
- Hmong (an Asian language brought to the French Guiana) and **Indigenous languages of America**: Kali'na, Wayana (Cariban languages), Arawak, Palikur (Arawakan languages), Teko, Wayãpi (Tupi-Guaraní languages);
- **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, Tamazight and Kabyle), Judeo-Spanish (in its Ḥaketía and Djudyó varieties), Yiddish, Western Armenian, Rromani, LSF (with an audio-visual recording which "doubled" in French by a researcher specialised in LSF, who also wrote an explanatory text about the body gestures which characterise this language).

Options enable the display (or not) of administrative boarders, the legend, the seas and new recordings (as well as their transcripts) outside of France: in Norman (in Jersey), Walloon (in Belgium), Catalan, Aranese Occitan, Aragonese, Basque, Asturian and Galician (in Spain), different Ligurian dialects of Italy, Haitian and Saint-Lucian (French-based) creoles, Bislama (English-based creole from Vanuatu), Fijian, Malagasy, Latin and even Esperanto. A specific checkbox allows the user to zoom on the Crescent, 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 — the limits of which appear fuzzy to highlight the transitional nature of this zone, between *Oïl* and *Oc* languages. Finally, a double orthography has been added for some varieties, in particular Berber (in Tifinagh and Latin alphabets), Western Arminian, Yiddish and Arabic dialects. A page "About" enables the visitor to know more about the ongoing project with some of our publications (Boula de Mareüil et al., 2018, 2019) and to download the data under a Creative Commons license.

2.2. Speakers and Protocol

The speakers we selected, most often elderly people (average age = 60), were from varied socio-professional backgrounds. They were recorded in quiet rooms and asked to sign consents for free distribution. A common protocol was applied, in which the speakers were asked to translate this story into their indigenous language, either directly

with the French text in front of them, or from a written text they preferred to read. The recordings were then intensity equalised, and great care was given to their orthographic transcriptions, which were checked by linguists.

Sometimes the speakers' productions moved away from literal translations to get closer to oral traditions. Some words gave rise to particularly enriching conversations:

- ***bise* 'north wind'**: tra(ns)montane in some Romance dialects, trade winds in some Polynesian dialects, Hebrew loans in some Jewish languages;
- ***voyageur* 'traveller'**: pilgrim in some Romance dialects, wanderer in some Germanic dialects;
- ***manteau* 'coat':** burnous in some Arabic dialects, rain protection in some indigenous languages of America; linen (*jii*) offered to "make the custom" or overcoat (*paleto*, actually a loan word of obscure origin) with which a traditional chief is clothed during his enthronement in some Kanak languages (see the illustration in Table 1).

In all cases, we paid a lot of care to the orthographic transcription of what was said, with in square brackets idioms of beginning (such as 'once upon a time') and idioms of end (of the type 'it is finished') which our Kanak speakers often wanted to add, The different translation strategies testify to the wealth of our linguistic heritage. It seems that this richness is of interest to the general public, in the sense that our site enjoyed a great success in print, broadcast and social media: it has attracted over 600,000 visits since its launch in 2017.

3. Conclusion and Future Work

3.1. Discussion

The aim, through this short outline of our speaking atlas, was to valorise the linguistic diversity of France, relying on a comparable basis which has didactic scope. This is almost urgent, inasmuch as a large number of indigenous languages of France are endangered. With this linguistic atlas, we hope to give prestige to dialects, to give them a positive image, for lack of being able to reverse the decline in their use — transmission among young people is not assured in many cases. It is probably inevitable that most dialects and minority languages of France will be supplanted by a more widely used language like French — which is also mortal. At a time when linguistic diversity and biological diversity are threatened, our profession of faith is that we will devote all our energy to delay the deadline. It is not (only) a matter of folklore tinged with exoticism and essentialism, reifying an idealised past (Bucholtz, 2003). Each language, each dialect provides formal means to express nuances of thought; each language, each dialect refers to a whole imaginary through what words evoke, through the play of sounds. Living with several languages opens up to the Other, it makes it possible to understand difference, it teaches people about the multiplicity of worldviews.

In the future, we plan to expand our linguistic mapping activity, combining field surveys and crowdsourcing: we still aim at enriching the speaking atlas of the languages and dialects of France, which on the other hand we recently extended to Italy, Belgium and Spain. The objective is to develop its educational aspect in two directions: starting from the existing materials, to which glosses as in Table 1 should be added, and from new surveys, two case studies will be considered (Romance languages and Polynesian languages) to immediately grasp variation through audio paths and the visualisation of isoglosses. Isolated words and/or linguistic notions will be presented, as in traditional dialectological atlases (Gilliéron and Edmont, 1902–1910; Charpentier and François, 2015). Visually and acoustically (the data are currently being transcribed phonetically by using forced alignment), different types and groupings of dialects will be illustrated.

3.2. Linguistic Mapping, Crowdsourcing and Dialectometry

Innovative visualisation methods will be developed, centered on the Gallo-Romance area on the one hand, on French Polynesia on the other hand — Romance and Polynesian languages presenting a certain homogeneity within each family. Linguistic variables will be selected: lexical items such as the words/concepts *wind*, *sun* and *cloak*, for example, taken from the recordings of the Aesop fable already mapped in more than 300 translations in the current version of our speaking atlas, to which words/concepts will be added from the *Atlas Linguistique de la France* (Gilliéron and Edmont, 1902–1910) and the Swadesh list collected during new surveys (Swadesh et al., 1971). The user will be able to follow phonetic and lexical changes by simply moving the mouse, possibly following courses which will be proposed. Isoglosses will also be generated more or less automatically, using clustering techniques, the results of which will be projected on maps, with different colour codes associated with different types and groupings of dialects. For these different aspects, we will also offer remote recording/transcription methods to enrich the database. In the longer term, solutions will be considered to go further (with one or more hundreds of words/concepts) and make the site sustainable.

The web will be used in order to display the fruits of research, as well as to collect new information, using a crowdsourcing methodology. This type of methodology will be particularly appreciable for French Polynesia, to avoid travelling in archipelagos very far from each other. Completing remote survey points, though, requires caution: we will take care of the linguistic content of the recordings and their transcripts. The breadth of the territory covered by this project and the diversity of the languages represented (Romance or even neo-Romance languages such as French-based creoles and Oceanian languages) opens up other avenues of research, which will require new collaborations.

In addition to the functionalities of the https://atlas.limsi.fr site, an attractive interface will also guide the journey through space, in Hexagonal France and French Polynesia — where extremely interesting lexical taboo phenomena (*pi'i*) exist. In order to go beyond the Aesop fable "The North Wind and the Sun" and to integrate a list of isolated words, for fauna and flora, for example — reviving with

the traditional practice of linguistic atlases —, we wish to design and/or improve an existing recording tool, LIG-Aikuma (Adda et al., 2016), to help field linguist in their investigations. Such an application will make it possible to capitalise on the network of dialect (or minority language) speakers we have, to ask them to record themselves. The program will need to be user-friendly, as speakers are often rather old. And this smartphone-based approach does not prevent the need for further field surveys, because human relationships are essential and because it may be important to be with the informants to start the process. The task of the field linguist equipped this way will be facilitated, and the processing of the data collected will be faster, to apply various dialectometric techniques.

4. Acknowledgements

This work was largely financed within the framework of the "Langues et Numérique 2016 & 2017" calls for proposals of the Délégation Générale à la Langue Française et aux Langues de France. We are grateful to the Académie des Langues Kanak and to all those who have agreed to give us their time and lend their voices to this achievement.

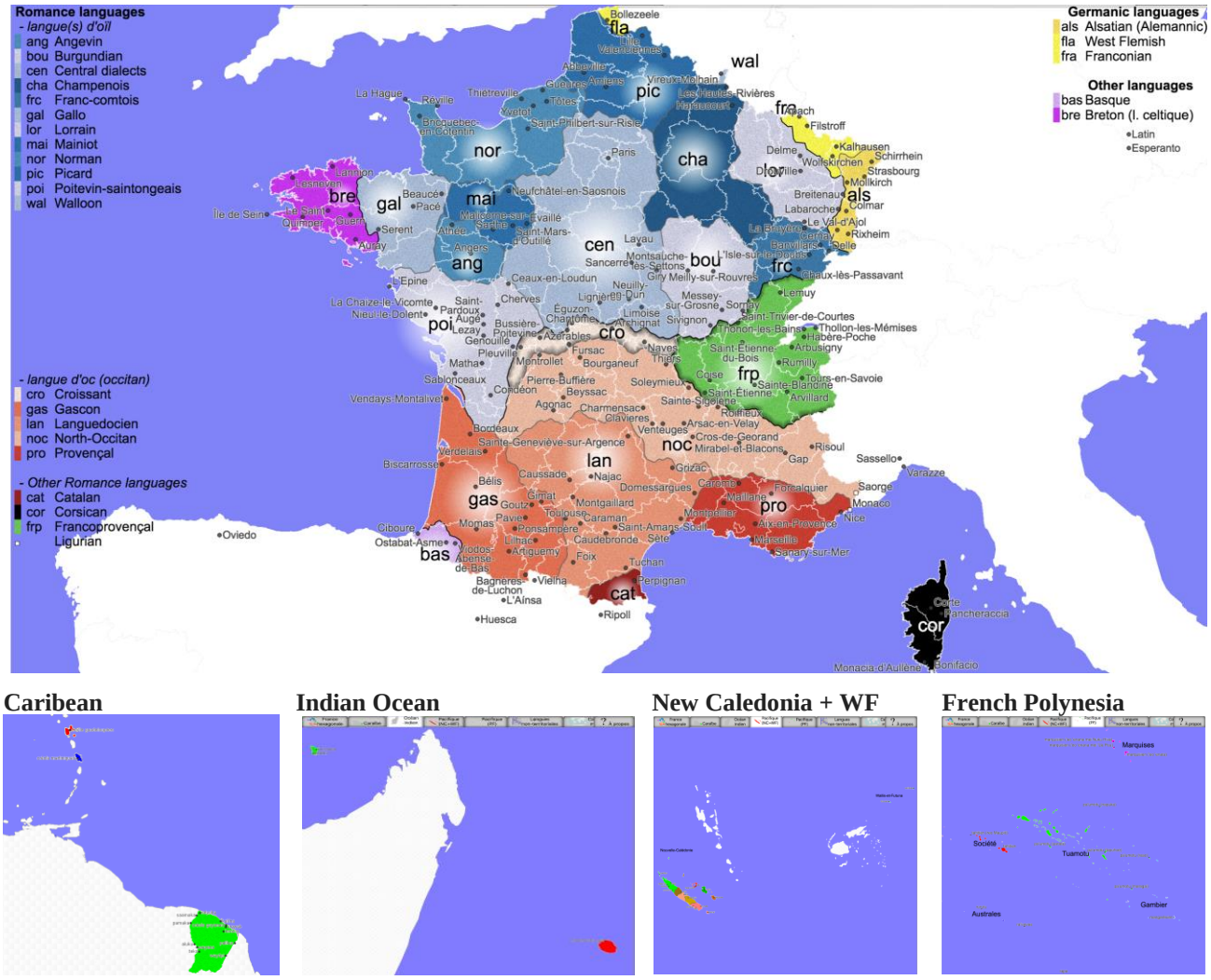

Figure 1: Maps of Hexagonal France and Overseas Territories (WF = Wallis-and-Futuna).

Naa	mwa	tûâ	taa	vinyié	nâ	nyî	vê mé	iitèè	naa	i pwa	ngê	taa	jii.
They two	TAM	see	a	man	and	he	come	meet	they two	covered	with	a	cloth

Nââ	ngê	tôâ	taa	xèxùù	na	yi	vê	bwè	nââ	iévititéé	mô	taa	paleto.
They two	TAM	see	a	man	who	he	come	towards	they two	wrapped	with	a	coat.

Table 1: Excerpt from the Aesop fable with literal translations in two Kanak languages (Numèè and Kwényï) spoken in the Djubéa-Kaponé customary area of New Caledonia. TAM = tense-aspect-mode particle.

5. Bibliographical References

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.

Bernabé, J. (2005). Guadeloupe et Martinique : Un survol sociolinguistique. *Langues et cité*, 5:6–7.

Biswana, H. (2016). Luta cultural e por direitos: reflexões dos Arowaka sobre a cultura, *Boletim informativo*, 2:22–26.

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

Boula de Mareüil, P., Adda, G., Lamel, L., Rilliard, A., Vernier, F. (2019). A speaking atlas of minority languages of France: collection and analyses of dialectal data. *Proceedings of the 19th International Congress of Phonetic Sciences*, Melbourne, 1709–1713.

Bucholtz, M. (2003). Sociolinguistic nostalgia and authentication of identity. *Journal of Sociolinguistics*, 7(3):398–416.

Courthiade, M. (2007). Jeu dialectes-langue. *Langues et cite*, 9:6–7.

Courthiade, M. (2013). *A succint history of the Rromani language*. INALCO, Paris.

Canguilhem, G. (1966). *Le normal et le pathologique*. Presses Universitaires de France Paris.

Caubet, D., Chaker, S., Sibille, J. (2001). *Codification des langues de France.* L'Harmattan, Paris.

Cayol, M., Chaudenson, R., Barat, C. (1984–1995). *Atlas linguistique et ethnographique de la Réunion*. Éditions du CNRS, Paris.

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>.

Charpentier, J.-M. and François, A. (2015). *Atlas linguistique de la Polynésie française*. Mouton de Gruyter, Berlin.

Corn, C. (1990). Tayo pronouns: a sketch of the pronominal system of a French-lexicon Creole language of the South Pacific. *Te Reo*, 33:3–24.

Cyrulnik, B. (2019). *La nuit, j'écrirai des soleils*, Éditions Odile Jacob, Paris.

Ehrhart. S. (1993). *Le créole français de St-Louis (le tayo) en Nouvelle-Calédonie*. H. Buske Verlag, Humburg.

Fishman J. A. (1991). *Reversing language shift: Theoretical and empirical foundations of assistance to threatened languages*. Multilingual Matters, Clevedon.

Gadet, F. (2003). *La variation sociale en français*. Ophrys, Paris.

Gilliéron, J. and Edmont, E. (1902–1910). *Atlas linguistique de la France*. Champion, Paris.

Goury, L. and Migge, B. (2003). *Grammaire du nengee. Introduction aux langues aluku, ndyuka et pamaka*. IRD Éditions, Paris.

Grégoire, Abbé H. (1794). *Rapport sur la nécessité et les moyens d'anéantir les patois et d'universaliser l'usage de la langue française*. Convention nationale, Paris.

Hagège, C. (2002). *Halte à la mort des langues*. Odile Jacob, Paris.

Laroussi, F. (2010). *Langues, identités et insularité : Regards sur Mayotte*. PURH, Rouen/Le Havre.

Lazard, G. and Peltzer, L. (2000). *Structure de la langue tahitienne*. Peeters, Paris.

Le Dû, J. and Brun-Trigaud, G. (2011). *Atlas linguistique des petites Antilles.* Éditions du CNRS, Paris.

Léglise, I. and Migge, B. (2007). *Pratiques et représentations linguistiques en Guyane.* IRD Éditions, Paris.

Moyse-Faurie, C. (2014). L Nouvelle-Calédonie et le statut des langues kanak : quelques repères historiques. *Langues et société*, 26:9–11.

Rastier, F. (2007). Conditions d'une linguistique des normes. In G. Steuckard, S. Sioufi (editors), *Les linguistes et la norme — Aspects normatifs du discours linguistique*, Peter Lang, Bern (pages 3–20).

Rivierre, J.-C. (1979). Langues de l'extrême-sud et plus particulièrement le nââ kwênyii, langue de l'Île des Pins. In A.-G. Haudricourt, J.-C. Rivierre, F. Rivierre, C. Moyse-Faurie, J. de la Fontinelle (editors), *Les langues mélanésiennes de Nouvelle-Calédonie*. Nouméa, DEC (pages 72–79).

Swadesh, M., Sherzer, J., Hymes, D. H. (1971). *The Origin and Diversification of Language*. Aldine, Chicago.

Vernaudon, J. (2017). L'origine des langues polynésiennes. *Langues et cité*, 28: 2–3.

Wacalie, F. S. (2013). *Description morpho-syntaxique du nââ numèè (langue de l'extrême-Sud, Nouvelle-Calédonie)*. Phd thesis, INALCO, Paris.