A Parallel Corpus of the Parable of the Prodigal Son: Building a Resource for Documenting Language Varieties in Mainland France

Lucence Ing*, Juliette Janès*, Sven Ködel◇, Benoît Sagot*
*Inria, Paris, France
{firstname.lastname}@inria.fr
◇Institut historique allemand, Paris, France
SKoedel@dhi-paris.fr

Abstract

This paper presents a historical parallel corpus of languages spoken in mainland France. It consists of a collection of versions of the Parable of the Prodigal Son, collected during the 19th century. The paper aims to present the interest of such a corpus, its constitution—through XML/TEI encoding, semi-automatic alignment and projection on linguistic maps—and its potential uses for the study of these low-resource languages.

Keywords: parallel corpus, history of language sciences, low-resource languages, language varieties

1. Introduction

France hosts multiple regional languages (Fig. 1),[1] many of which are now endangered, as they are spoken by only a limited number of speakers today (Cerquiglini, 1999). Among them, Romance languages include the *langues d'oïl* (e.g. Picard, Norman, Gallo) in the north, the *langues d'oc* (Occitan and its varieties such as Provençal and Limousin) in the south, and the *Franco-Provençal* varieties in the south-east. Other linguistic groups include Celtic (Breton) and Germanic (Alsatian, Platt). These languages form part of a historical linguistic continuum extending across mainland France and neighbouring regions.

Figure 1: Mainland France and neighbouring regions linguistic map.

Preserving and promoting these languages requires the development of language resources such as text corpora and annotated datasets, which are essential for both linguistic research and natural language processing (NLP). Several initiatives have contributed to this effort. The multilingual projects ANR RESTAURE (Bernhard and Vergez-Couret, 2015), focusing on Alsatian, Picard and Occitan, and DIVITAL,[2] covering Alsatian, Corse, Occitan and Poitevin-Saintongeais, aim to enhance the digital visibility of France's regional languages. Other efforts include language-specific corpora, such as PICARTEXT for Picard (Eloy et al., 2015), BaTelÒc for Occitan (Bras and Vergez-Couret, 2013), and MeThaL for Alsatian (Fabo, 2023).

This paper builds on this line of research by presenting a historical corpus of language varieties from mainland France other than standard French. It consists of multiple versions of the Parable of the Prodigal Son (PPS), a biblical passage from Luke 15:11–32, produced for language surveys in the 19th century. The corpus, released under a CC-BY licence, is encoded in XML/TEI with linked referentials, ensuring interoperability and facilitating reuse.

The paper discusses the historical importance of the corpus, describes its development, and evaluates its linguistic quality, using semi-automatic textual alignment and cartographic projection.

2. Data collection overview

The history of the Parable of the Prodigal Son in the development of the language sciences highlights the intersection of linguistics, culture, and politics. Breton lawyer Jacques Le Brigant (1720-1804) is credited with being the first to publish the parable in a local variety of Breton (Le Brigant, 1779). However, the interest in the diversity of France's regional languages really emerged only at the beginning of the 19th century, following the French Revolution. This interest arose in reaction to the Jacobin policy of eradicating local languages other than French, as expressed in Henri Grégoire's famous report on

[1] https://commons.wikimedia.org/wiki/File:Langues_de_la_France.svg?lang=fr

[2] https://divital.gitpages.huma-num.fr/fr/

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 3317–3324
11–16 May 2026.

the necessity of annihilating the "*patois*" (Grégoire, 1794). Initially, efforts were made to translate into several regional languages, but linguistic diversity soon came to be seen as an obstacle to national unity, leading to a more repressive linguistic policy.

By 1800, the central state emerging from the Revolution lacked precise knowledge of the languages spoken by its population. The perception of linguistic diversity also evolved, as popular culture and language came to be seen as evidence of the nation's origins. Therefore, within the broader effort to provide a statistical description of France, a linguistic survey was initiated. Its aim was to assess the number of speakers, to locate linguistic boundaries more precisely, and to identify and describe the varieties present on French territory.

As head of the statistics office at the Ministry of the Interior, Charles-Étienne Coquebert de Montbret chose the PPS as a linguistic sample, thereby creating a tool that would be used to document languages until the early 20th century. His choice was motivated by the parable's length, its popularity, and its vocabulary, which includes concrete and common words, facilitating its translation into vernaculars often spoken in rural settings.

The collection of translations of the parable, which began 1806 and lasted until 1812, benefited from the strong administrative hierarchy of Napoleonic France and the dense network of administrative agents across its territory. Coquebert de Montbret's linguistic survey thus relied on the personnel and expertise of the central state. Using a single text as a sample, produced by local informants following precise and identical instructions, ensured the uniformity and intelligibility of the collected data despite geographical distance.

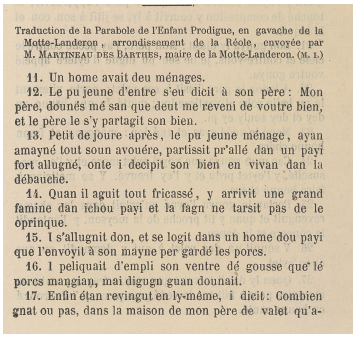
Over six years, the statistics office managed to gather approximately 500 versions of the parable, covering the entire territory and languages of contemporary France, as well as then annexed or controlled territories such as Belgium, Luxembourg, parts of the Netherlands, Germany, northern Italy, and Spain. A few years later, the Royal Society of Antiquaries of France continued the collection, adding a small number of translations through its network of corresponding members.

Napoleonic prefects and collaborators involved in Coquebert's survey also published isolated translations. The model was quickly transferred to other European countries, where collections of the PPS were published up to the end of the 19th century.

3. Building the corpus

3.1. Source Material

The corpus of translations of the PPS gathered by Coquebert de Montbret was never compiled into a single publication. While a few texts appeared in contemporary journals, the largest compilation remains the 90 versions published in 1824 (Coquebert de Montbret, 1824). This compilation was reprinted in 1831 (Labouderie and Coquebert de Montbret, 1831) and later reissued by Schnakenburg (1840) and Favre (1879). Most of the corpus, however, survives only in manuscript form (Coquebert de Montbret, 1812b,a). These documents, dispersed across French and European archives, have long been difficult to access due to fragmentation, lack of metadata and limited digitisation.


Traduction de la Parabole de l'Enfant Prodigue, en gavache de la Motte-Landeron, arrondissement de la Réole, envoyée par M. Martineau des Barthes, maire de la Motte-Landeron. (M. L.)

11. Un home avait deu ménages.
12. Le pu jeune d'entre s'eu dicit à son père : Mon père, dounés mé san que deut me reveni de voutre bien, et le père le s'y partagit son bien.
13. Petit de joure après, le pu jeune ménage, ayan amayné tout soun avouére, partissit pr'allé dan un payi fort allugné, onte i decipit son bien en vivan dan la débauche.
14. Quan il aguit tout fricassé, y arrivit une grand famine dan ichou payi et la fagn ne tarsit pas de le oprinque.
15. I s'allugnit don, et se logit dans un home dou payi que l'envoyit à son mayne per gardé les porcs.
16. I peliquait d'empli son ventre dé gousse que lé porcs mangian, mai digugn guan dounait.
17. Enfin étan revingut en ly-même, i dicit : Combien gnat ou pas, dans la maison de mon père de valet qu'a-

Figure 2: 1879 edition parable.

Recent mass digitisation initiatives have made many 19th-century printed sources accessible online, but only a few manuscript parables have been digitised to date (e.g. the National Library of France [BnF] and the municipal library of Rouen). The complex transmission history of those documents (multiple manuscript copies, administrative redactions, and editorial normalisation) requires contextualisation through transmitted co-texts (letters, notes, memoirs). Systematic metadata enrichment (localisation, dating, author identification) is therefore essential to enable reliable reuse and to support new investigations in historical sociolinguistics.

3.2. Methodology of Construction

The Favre (1879) edition (Fig. 2) digitised by the BnF and published on Gallica serves as the main source for the present corpus, the parallel corpus of the Parable of the Prodigal Son (PCPPS). The OCR[3] output provided by Gallica in XML/ALTO[4] is extracted by using its API. We then perform manual correction to ensure textual accuracy and to normalise OCR results. Thanks to the repetitive layout of each parable, a Python script converts the corrected ALTO files into a single XML/TEI (TEI Consortium, 2025) corpus. Following the Text Encoding Initiative (TEI) Guidelines, this encoding

[3]Optical Character Recognition.

[4]Standard format for OCR output, giving both the textual content and the position coordinates on the page. https://www.loc.gov/standards/alto/

```
<div xml:id="limo-1246-31" type="parabole" n="31"
     xml:lang="limo-1246" hand="#person_19" corresp=
     "#place_28">
    <head xml:lang="stan-1290"><lb/>Traduction de la
      Parabole de l'Enfant Prodigue, en gavache de
     la <lb/>Motte-Landeron, arrondissement de la Ré
     ole, envoyée par <lb/>M. Martineau des Barthes,
      maire de la Motte-Landeron. (M. I.)</head>
    <p n="11"><lb/><num>11.</num>Un home avait deu m
     énages.</p>
```

Figure 3: Example of parable version encoded with linked metadata.

```
<place xml:id="place_28" corresp="#limo-1246">
    <settlement>
        <name>Motte-Landeron</name>
        <idno type="geonames">3008042</idno>
    </settlement>
    <region><name>Gironde</name>
    <idno type="geonames">3015948</idno>
    </region>
</place>
```

Figure 4: Example of place information associated with Fig. 3.

preserves the structure of the printed edition (e.g. tagged and numbered paragraphs) while integrating metadata for each parable (Fig. 3).

The resulting TEI file is manually reviewed and sociolinguistic metadata are added for each translation. To support data reuse and document language variation, each translation includes the following metadata: the **language** (with Glottolog[5] identifiers, inferred when not explicitly mentioned); the **locality** (linked to precise geographical entities with Geonames[6] identifiers like settlement or region, Fig. 4); and the **collector** (name, profession, origin, residence and known languages).

This encoding ensures interoperability an long-term preservation of an enriched parallel corpus for the languages of France.

3.3. Content and Coverage

The PCPPS consists of 89 dialectal versions of the same 22 paragraphs, amounting to approximately 100,000 tokens. The encoded document covers 22 linguistic varieties, from the mainland France and neighbouring countries (mostly Belgium and Switzerland). Most of the versions published in this edition date from 1806–1824.

No parable originates from the Paris Basin and neighbouring regions. The recorded versions come from areas with stronger regional linguistic diversity, ranging from northern France to southern Occitania, Provence and Swiss Romandy. Table 1 provides an overview of the distribution of texts by variety, with labels inferred from collection locations.

[5] https://glottolog.org/
[6] https://www.geonames.org/

Lang. Family	Varieties	Quant.
Oil	Wallon	6
	Picard	4
	Bourguignon	1
	Normand	2
	Franc-Comtois	11
	Poitevin-Saintongeais	7
Occitan	Auvergnat	3
	Limousin	8
	Languedocien	16
	Gascon	1
	Vivaro-Alpin	11
	Provençal	9
Germanic	Platt	4
	Alemanique	1
Others	Franco-provençal	3
	Romanche	2

Table 1: Language varieties in the corpus.

4. Quality Analysis of the Corpus

Even though it illustrates France's linguistic diversity, the parable corpus has long been considered unreliable due to its inconsistent transcription practices (de Tourtoulon and Bringuier, 1876). Indeed, before the invention of the International Phonetic Alphabet at the end of the 19th century, there was no standardised way for transcribing the spelling of the words, and each transcriber could have his own way to transcribe the sounds. However, a recent study on the "Parable of Grégoire" in Lorrain has proven the data's linguistic quality (Duval, 2019), which shows that the choices of transcription, even if not standardised, were coherent. We therefore compare the lexicon and spelling of each word of some investigation points with other dialectal—later—sources to evaluate its linguistic reliability, based on the correspondence between the standardised spelling for a place and the spelling of the PCPPS data. It is still important to remind that, nowadays, even if IPA exists, it is not used for basic transcription of texts; some varieties are still not standardised and can be written in different ways.

4.1. Comparison with the data of ALF

The PCPPS is compared with data from the online edition (Gilliéron and Edmont, 2013) of the *Atlas linguistique de France* (ALF), published between 1902 and 1910, a century after Coquebert de Montbret's survey, and presenting a standardised way of sound transcription. Each of its 1,920 maps presents a word (produced in context) and its translation across the 639 localities in mainland France, chosen by the investigators to illustrate the diversity of spoken languages (e.g. map 722 for *jeune* ['young'], Fig. 6).

This approach follows a tradition of dialectometric studies. From the ALF to the *Nouveaux Atlas*

le (the)	père (father)	dit (said)
l’	pére	dit
l’	per	déjo
lo	pere	deheu
lo	père	djet
lo	père	dehé
lo	père	die
Son	père	dijy
lo	pere	diji
Lo	pere	dejeut
lou	pare	dizèt

Table 2: Alignment at the word level

.

est/fut (*is*/*was*)	tout (*very*)	proche (*close*) / approcha (*approached*)
fut	-	proche
-	-	aprustzét
est	tôt	près
foût	to	près
-	-	approcheuve

Table 3: Level of segmentation adopted in the alignment. Languages are French, Auvergnat and, for the last third lines, Walloon.

linguistiques de France (NALF), producing more precise regional data (Glessgen and Sauzet, 2020), linguistic variation is analysed through maps, enabling the visualisation of isoglosses across the entire linguistic diasystem. There are now online interfaces, such as Thesoc[7] for Occitan, based on the NALF series. We reuse the ALF data because they are closer in date to the 19th century sources than the NALF series. They are also already reused for dialectometric analyses (Goebl, 2003).

4.2. Parable Alignment

Mapping the PCPPS data implies identifying every word corresponding to an ALF map (e.g. all the forms that correspond to the word ‘father’). We therefore decide to align the parables at the word level (Table 2).

The question of the alignment granularity is non-trivial. In Table 3, an equivalence between *foût to près* and *approcheuve* can be observed. However, if we only encode this 3-to-1 word alignment, we miss the variation between *proche* and *près* (‘close’) and that between *tôt* and *to* (standard French *tout* [‘very’]) in those languages that do not use a single word to express ‘was close’. We therefore decide to perform token-level alignment, as illustrated in Table 3, accepting to lose the one-to-many mapping in cases such as *approcheuve* vs. *est tôt près*.

We first compute automatic alignments using collatex[8], a tool designed to analyse the variations between textual witnesses. It relies on the automatic identification of token-level similarity using the Levenshtein distance. However, the generated alignments are not accurate due to several challenges including:

- spelling variation, that can be extensive, as for the subjunctive of *aller* (‘to go’): *olle* in Franc-Comtois, *aie* in Poitevin;
- lexical choices, sometimes even syntactic structures, that may vary from one variety to the next, as a result of subjective translator choices (see Table 4 for an example).

Because of the numerous difficulties, the automatic alignment is not perfect. We thus manually corrected its output. Once all the paragraphs were aligned and corrected, we automatically generated a new TEI file with the collation (Fig. 5).

j’ (j’ / *I*)	a (aï / *have*)	pegchî (péchè / *sinned*)	conte (contre / *against*)	lu (le / *the*)		ci (ciel / *heaven*)
dj’ (j’ / *I*)	a (aï / *have*)	offaincé (offensé / *offended*)		l’ (le / *the*)	bon (bon / *good*)	diet (dieu / *god*)
j’ (j’ / *I*)	ai (aï / *have*)	maufait (meffait / *misbehaved*)	à l’enconte (à l’encontre / *against*)	du (du / *the*)		ciel (ciel / *heaven*)

Table 4: Three Walloon translations of part of verse 18 (“I have sinned against heaven”).

```
<app>
    <rdg wit="#stan-1290-1 #fran-1270-18 #fran
     -1270-19 #fran-1270-21 #poit-1241-28 #wall
     -1255-3 #poit-1241-30 #lang-1309-39 #wall
     -1255-4 #limo-1246-42 #wall-1255-5 #lang
     -1309-52 #auve-1239-57 #viva-1235-69 #viva
     -1235-70 #arto-1238-8 #norm-1245-84">homme</rdg
     >
    <rdg wit="#arto-1238-10 #lorr-1242-13 #auve
     -1239-2 #fran-1270-20 #limo-1246-31 #limo
     -1246-33 #auve-1239-38 #gasc-1240-43 #lang
     -1309-45 #lang-1309-47 #lang-1309-48 #cata
     -1291-49 #lang-1309-53 #viva-1235-58 #prov
     -1235-64 #viva-1235-67 #fran-1270-73 #stan
     -1289-74 #high-1290-87 #fran-1270-89 #fran
     -1270-90">home</rdg>
    <rdg wit="#wall-1255-11">oum</rdg>
    <rdg wit="#lorr-1242-12">oumme</rdg>
    <rdg wit="#lorr-1242-14">hame</rdg>
    <rdg wit="#lorr-1242-15">am</rdg>
```

Figure 5: Beginning of the collation in TEI of *homme* (‘man’) in all the Parable versions. Each `<rdg>` element presents a specific spelling. It is linked to the identifier of the witness(es) that record(s) it thanks to `@wit` attribute.

4.3. Projection on Maps

4.3.1. Methodology

As georeferenced maps are not publicly accessible at the moment, ALF map images are used as back-

[7] http://thesaurus.unice.fr

[8] https://interedition.github.io/collatex/pythonport.html

ground layers for our own. This allows us to compare the lexical entries from the parable with those recorded in the ALF. The maps, downloaded in CartoDialect, are georeferenced manually in QGIS[9]. The alignment is performed by matching the historical maps with an OpenStreetMap[10] base layer. We use reference points like political borders, river confluences and coastal landmarks were used to ensure accuracy. For every identified lexical item, the alignments described in Sec. 4.2 are compiled into a file containing all lemmas and their associated Geonames coordinates extracted from the TEI file. Finally, the compiled data are projected onto the maps associated with each lexical item.

4.3.2. Analysis of Maps

Projecting the parable data on maps can help studying phonetics and lexicon. It is also a step for evaluating the linguistic quality of the data, showing if a correspondence between PCPPS data and ALF data exists. Fig. 6, in Appendices, is an example of a produced map and represents the projection of the word *jeune* ('young'). The points corresponding to the forms of the PCPPS seem to correspond to the variation represented on the map: *tzoïné* is close to the spelling identified on point 616 (Dordogne), *joube* to point 741 (Tarn-et-Garonne), *djoueine* to point 814 (Haute-Loire), etc.

Table 5 presents the proportions of correspondences across 19 maps.[11] The analysis considers both lexicon and spelling levels. An investigation point which lies between two—or more—ALF points is identified as geographically uncertain, since it could correspond to one point or another. These points that differ from their surroundings may indicate an isogloss change. The geographic uncertainty can be applied to both lexicon and spelling.

Correspondence	%
Good	40.3
Maybe good (Fuzzy location)	12.1
Maybe good (Fuzzy transcription)	20.9
Maybe good (Fuzzy loc. and transcr.)	9.3
Bad transcription	13.9
Bad lexicon	3.5

Table 5: Proportions of correspondences on 19 maps (1,197 points).

Spelling uncertainty is another type of uncertainty that must be highlighted. For example, in Fig. 6, *june* is near point 65, whose form is *djun*. It is possible that the initial *j-* corresponds to the same spelling as *dj-*, [ʤ], but this remains uncertain. Due to the lack of phonetic transcription standards at the time, many forms correspond to this uncertainty. Moreover, both types of uncertainty can be combined: a survey point can be located between several ALF points that could—or could not—correspond to the current spelling.

The results in Table 5 show that the greatest part (40.3%) of forms from the PCPPS present a correspondence with ALF data. Nevertheless, 17.4% of the PCPPS data do not correspond to the ALF data, mostly because of differences in spelling. The remaining part of the data (42.3% in total) may be the most interesting, as it corresponds to uncertainty in location (leading to the study of isogloss changes) or in transcription (leading to study of the evolution of transcription norms).

5. Conclusion and Further Work

The constitution of the PCPPS is in progress. Currently, the corpus remains rather small and the data reflect only a limited portion of the linguistic variation in mainland France. It should soon be increased by additional parables from:

- the Coquebert de Montbret's surveys, including both edited sources and manuscript materials such as the Montbret collection (Coquebert de Montbret, 1812b) preserved at the municipal library of Rouen, in order to produce the first complete edition of this corpus;
- another unpublished survey, Bourciez (1886), which includes 4,444 parables collected in the South-West of France.

The corpus is also going to be deepen in complexity. Indeed, the TEI format enables to handle the diversity of the sources within a single file, including variations across the different editions of the Coquebert de Montbret's survey. Work is already underway to add the Labouderie and Coquebert de Montbret (1831) edition to the XML file.

The analysis of the linguistic quality of the corpus will be more systematic. First, the workflow for alignment and mapping will be further automated. Second, the study of correspondences between PCPPS data and ALF points will be extended to more data. This will provide an overview of how survey data aligns with ALF data, higlighting differences that could be explored.

This paper outlines the initial steps in building a historical parallel corpus of low-resource languages from unpublished sources. Its expansion will support research in the history of linguistics and aid the documentation and promotion of these varieties.

[9] QGIS is an open-source geographic information system (GIS) software used for viewing, editing and analysing geospatial data. https://qgis.org/

[10] https://www.openstreetmap.org

[11] The list of the maps used for the analysis can be found in Sec. 11, Appendices, with three other examples.

6. Acknowledgements

We would like to thanks the colleagues from the Bourciez project, in particular Alexandre Génadot, for sharing data and expertise.

This work was mainly funded by Inria under the "*Défi*"-type project COLaF (Corpus et Outils pour les Langues de France). It also benefited from the support of Benoît Sagot's chair in the PRAIRIE-PSAI institute, funded by the ANR as part of the "France 2030" strategy under the reference ANR23-IACL-0008.

7. Ethical Concerns

Our work does not involve any particular ethical considerations.

8. Data and code availability

All data of the PCPPS (original structured files, aligned data, maps) can be found on this Github repository: `https://github.com/DEFI-COLaF/Parabole.git`.

9. Bibliographical References

Delphine Bernhard and Marianne Vergez-Couret. 2015. Le projet RESTAURE. In Colloque sur les technologies pour les langues régionales de France (TLRF 2015), Les technologies pour les langues régionales de France, pages 96–100, Meudon, France. Ministère de la Culture et de la Communication - Délégation générale à la langue française et aux langues de France.

Myriam Bras and Marianne Vergez-Couret. 2013. BaTelÒc : a Text Base for the Occitan Language. In Proceedings of the First International Conference on Endangered Languages in Europe, Alcanena, Portugal.

Bernard Cerquiglini. 1999. Les Langues de France : 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. Technical report, Ministère de l'éducation nationale, de la recherche et de la technologie.

Charles de Tourtoulon and Octavien Bringuier. 1876. Étude sur la limite géographique de la langue d'oc et de la langue d'oïl. Paris.

Marc Duval. 2019. Grégoire patoisant ? Essai de localisation d'une parabole "en patois lorrain, communiquée par le comte Grégoire". Revue de linguistique romane, 83(331):335–378.

Jean-Michel Eloy, Fanny Martin, and Christophe Rey. 2015. PICARTEXT : Une ressource informatisée pour la langue picarde. In Proceedings of TALaRE 2015 - Traitement Automatique des Langues Régionales de France et d'Europe, Caen, France.

Pablo Ruiz Fabo. 2023. The MeThAL Alsatian Theater Corpus and Related Resources: Work Done and Perspectives. In 5èmes journées du Groupement de Recherche CNRS "Linguistique Informatique, Formelle et de Terrain ", pages 113–118, Nancy, France.

Martin Glessgen and Maguelone Sauzet. 2020. La trajectoire et l'exploitation lexicale des Nouveaux atlas linguistiques de la France. Bien Dire et Bien Aprandre, 35:9–46.

Hans Goebl. 2003. Regards dialectométriques sur les donées de l'Atlas linguistique de la France : Relations quantitatives et structures de profondeur. Estudis Romànics, 25:60–117.

Henri Grégoire. 1794. Rapport sur la nécessité et les moyens d'anéantir les patois et d'universaliser l'usage de la langue française. Paris.

Jacques Le Brigant. 1779. Élémens de la langue des Celtes Gomérites ou Bretons : Introduction à cette langue et par elle à celles de tous les peuples connus. Strasbourg.

TEI Consortium. 2025. TEI P5: Guidelines for Electronic Text Encoding and Interchange, version 4.9.0.

10. Language Resource References

Bourciez, Edouard. 1886. Recueil des idiomes de la Région Gasconne. Université Bordeaux-Montaigne.

Coquebert de Montbret, Charles-Etienne. 1812a. Fond Coquebert de Montbret. Bibliothèque municipale de Rouen.

Coquebert de Montbret, Charles-Etienne. 1812b. Notes et documents sur les patois de la France, recueillis par les préfets des différents départements de l'Empire, vers 1811 et 1812. Bibliothèque nationale de France, Département des Manuscrits.

Coquebert de Montbret, Charles-Etienne. 1824. Matériaux pour servir à l'histoire des dialectes de la langue française. Mémoires de la Société des Antiquaires de France.

Favre, Léopold. 1879. Parabole de l'enfant prodigue en divers dialectes, patois de la France, avec une introduction sur la formation des dialectes et patois de la France. Eds. L.Favre.

Gilliéron, Jules and Edmont, Edmond. 2013. CartoDialect, Edition numérique de l'Atlas Linguistique de la France. ANR ECLATS.

Labouderie, Jean and Coquebert de Montbret, Eugène. 1831. Mélanges sur les langues, dialectes et patois, renfermant, entre autres, une collection de versions de la parabole de l'enfant prodigue en cent idiomes ou patois différents, presque tous de France, précédés d'un essai d'un travail sur la géographie de la langue française. Bureau de l'Almanach du Commerce.

Schnakenburg, J.F. 1840. Tableau synoptique et comparatif des idiomes populaires ou patois de la France. Foerstner.

11. Appendices

To proceed to the analysis of the linguistic reliability of the data, we made a comparison the ALF data and the PCPPS data for the following words: *père* ('father'), *fils* ('son'), *donner* ('give'), *porcs* ('pig'), *pays* ('country'), *jeune* ('young'), *jour* ('day'), *garder* ('keep'), *deux* ('two'), *après* ('after'), *pain* ('bread'), *aller* ('go'), *avait* ('had'), *rentrer* ('return'), *faim* ('hunger'), *homme* ('man'), *ciel* ('heaven'), *veau* ('calf'), *ami* ('friend').

The followed images were selected among the 19 maps created for the analysis. Other maps are available on the Github project[12].

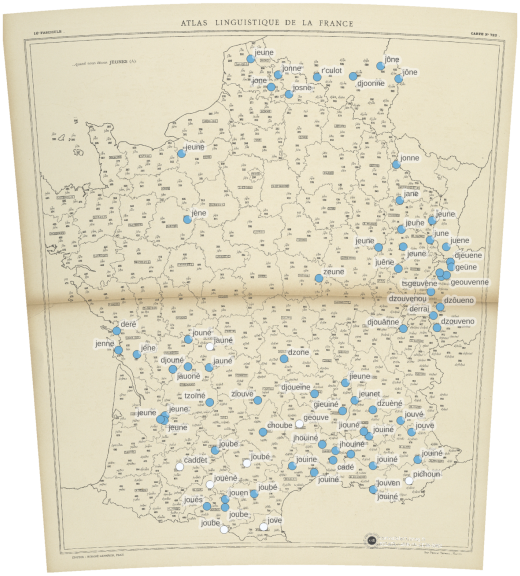


Figure 6: Map of the word *jeune* ('young') (blue is for specific localities and white for regions).

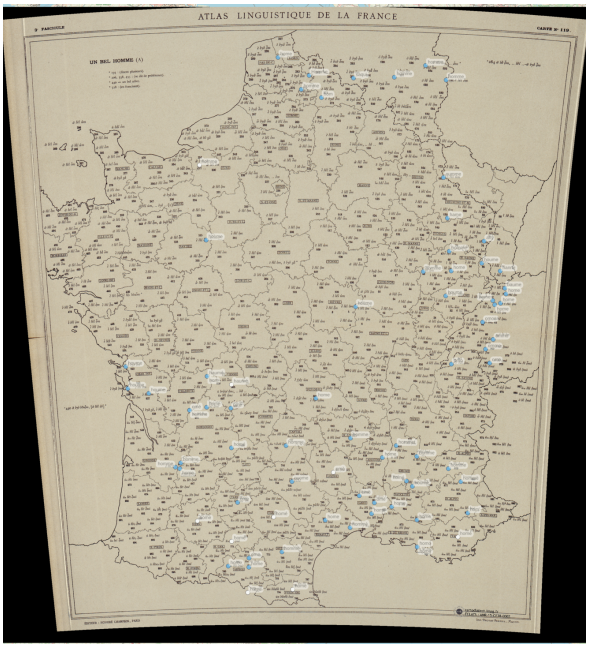


Figure 7: Map of the word *homme* ('man').

[12] https://github.com/DEFI-COLaF/Parabole/tree/main/4_visualisation_carte/2_cartes

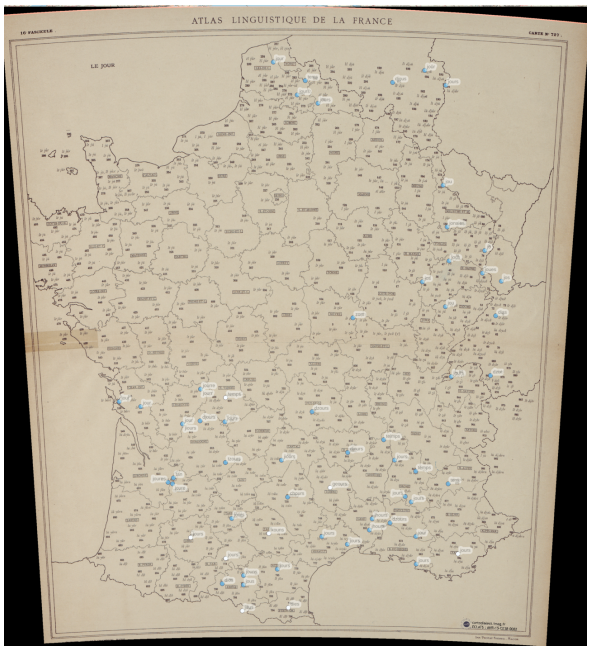


Figure 8: Map of the word *jour* (‘day’).

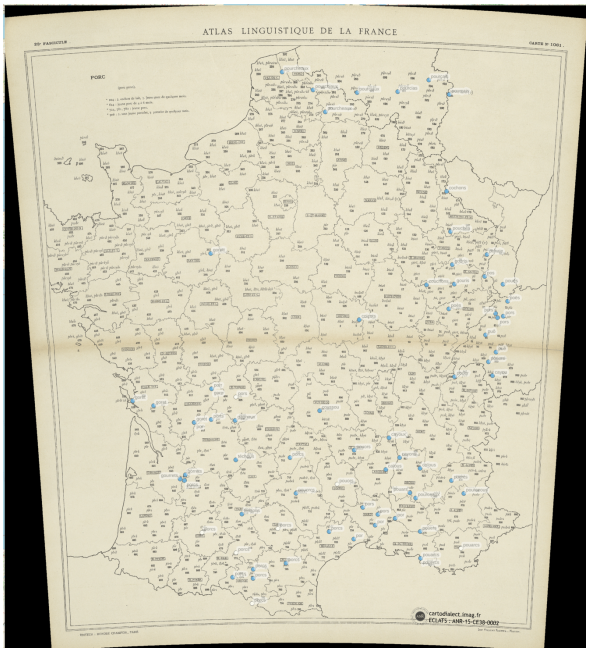


Figure 9: Map of the word *porc* (‘pig’).