frontiers | Frontiers in Complex Systems

TYPE Original Research
PUBLISHED 20 December 2024
DOI 10.3389/fcpxs.2024.1429114



OPEN ACCESS

EDITED BY
Marco Patriarca,
National Institute of Chemical Physics and Biophysics, Estonia

REVIEWED BY
Gianmario Raimondi,
Università della Valle d'Aosta, Italy
Mikhail Tamm,
Tallinn University, Estonia
Jacques François,
Université de Caen Normandie, France

*CORRESPONDENCE
Jean Léo Léonard,
leonardjeanleo@gmail.com

RECEIVED 07 May 2024
ACCEPTED 15 August 2024
PUBLISHED 20 December 2024

CITATION
Léonard JL (2024) Revisiting Southern Gallo-Romance from a complexity theory standpoint: Occitan.
*Front. Complex Syst.* 2:1429114.
doi: 10.3389/fcpxs.2024.1429114



# Revisiting Southern Gallo-Romance from a complexity theory standpoint: Occitan

Jean Léo Léonard*

Laboratoire Dipralang (EA 739), Université Paul-Valéry Montpellier 3, Montpellier, France

In this paper, the inner structure of the Occitan dialect network is revisited in the light of a range of cumulative (Ward's method) vs. reductive (Complete linkage, Groupe Average, Weighted Average) hierarchical algorithms provided by Gabmap, an Online dialectometric application for calculating distance/similarities by edit distance (Levenshtein algorithm). Reticularity of the Occitan geolinguistic space is addressed through connectograms using Gephi, and Multidimensional Scaling is also used to some extent. After sketching the canonical classifications of the Occitan geolinguistic space (Bec, Ronjat), providing the "eponymous dialects", we explore the deep patterns of this diasystem, bringing to light a hierarchy of systemic entities constituting an array of "invisible dialects", corresponding to entities of various size and functions (*macrodialects, dialects, subdialects, varieties, hubs, small worlds, buffer zones, default dialects*). The approach is based on concrete linguistic data from the THESOC database (Université de Nice/CNRS), contrasting the major isoglosses (macrodialectal features) with the "intricate variables", i.e., segmental *nexi*, extracting data relating to strategic points in the complex dialectal network from reductionist algorithms.

KEYWORDS

**Occitan, dialect, dialectometry, taxonomy, edit distance, diasystem, external factors, language dynamics**

## 1 Introduction

The classification of dialects within a dialectal network or linguistic domain cannot be an end in itself, nor can it be limited to objectives such as mere "territorial mapping." This article proposes an entirely different perspective, geared towards the benefit that the theory of Complex Adaptive Dynamical Systems (CADS) can derive from the classificatory practice in dialectology. Our aim here is thus not so much to propose a new classification of a linguistic domain (in this case, Occitan within the Gallo-Romance dialect network), or to merely confirm or refute current knowledge in this field, but rather to harness the algorithmic resources used in implementing Complexity Theory to enrich the empirical and epistemological horizon thereof. Such an approach will also reciprocally enrich the emergent paradigm of Language Dynamics. The primary aim of this paper is therefore to harness the robust analytical power of Complexity Theory (CT) to fathom the intricate linguistic landscapes of the Occitan dialect continuum. By employing advanced algorithmic tools and dialectometric methods, namely, here, those made available through Gabmap, not only do we seek to highlight the complex interplay of historical,

social, and linguistic factors shaping this geolinguistic domain, but we also dare say that "old classifications", or "good old taxonomies" may turn out to be heuristic grids to explore the inner structures of CADS such as dialect domains. We advocate for a quantitative analysis that can grasp the typological trends and inner diversity within well-defined linguistic domains, such as *diasystems* (languages seen as holistic complexes integrating all dialect varieties: see Weinreich, 1954 and below). Specifically, our study leverages CT to provide a nuanced and comprehensive view of dialects, subdialects, and varieties, moving beyond mere classification to uncover the underlying structures and patterns that govern linguistic variation in space and time. Our research aims to apply the principles of CT to dialectology, showing how dialects can be systematically categorized and understood from a wide array of both quantitative and qualitative standpoints. Using a phonological cognate set from the THESOC database, employing various hierarchical clustering methods, and exploring network-based models, we intend to demonstrate how CT can enrich our empirical and theoretical understanding of linguistic diversity. This endeavor should not only foster our knowledge of the Occitan domain but also contribute to the broader fields of General Dialectology (GD) and General Systems Theory (GST), providing innovative tools and methods for future research. In summary, our paper sets out to: (i) apply Complexity Theory to dialect variation within the Occitan linguistic domain, (ii) implement Gabmap and other computational tools to analyze and map dialectical diversity, (iii) challenge traditional classifications with data-driven and various algorithmic methods, (iv) sketch a model of *dialect ontology* (clusters of varieties salient in space and time, i.e., having a geohistorical and linguistic identity, that incorporates findings from hierarchical and network analyses, unravelling *visible* as much as *invisible* dialects/entities, (v) provide new insights on methodologies for the study of dialects and their typological features, as we systematically seek to enhance the heuristic properties of both qualitative and quantitative methods, within the framework of CT[1].

Dialectometry grasps typological trends, as suggested in (v) above, rather than phylogenetical evidence. In this respect, we consider Gabmap to be an outstanding heuristic tool, as the team who has been engineering it for over 20 years already (Nerbonne, Heeringa, Bolognesi, Prokić, Weiling, etc.) has shown remarkable epistemological awareness, developing and integrating into its general design many strategic mathematical and computational devices to enhance its heuristic accuracy in processing dialectical complexity. This concern for developing algorithmic solutions to comprehend both surface and deep structures of any geolinguistic space matches the demands of Complexity Theory (CT), in terms of providing a wide and robust array of holistic results on topics from social sciences–as linguistics or dialectology.

In the case of CT applied to a geolinguistic domain, especially a huge one such as Occitan (see Figure 1), we face a real challenge: the magnitude of the superdivisions (macrodialects, such as Arverno-Mediterranean on a vertex NW-SE vs. Aquitano-Pyrenean and Central Occitan, from Bordeaux to Narbonne or Montpellier) and subdivisions (Limousin, Auvergnat, Vivaro-Alpine and Provençal on the one hand, Gascon and the wide and intricate complex of Languedocian on the other hand).

These dialects in turn split into many subcomponents (subdialects, clusters of varieties, varieties). Jules Ronjat (1864–1925), author of the first landmark classification of Occitan dialects (Ronjat, 1941, 1–55)[2] divided the domain of "langue d'oc", i.e., Occitan, into five major dialects, which he simply called "groups" (A: Provençal, B: Languedocien-Guyennais; C: Aquitain; D: Auvergnat-Limousin, E: Alpin-Dauphinois). He then divided these clusters into "branches" and "subgroups" – the former were ranked by Roman letters, the latter by Greek characters, e.g., A = $\alpha$ to $\delta$ (=4 subgroups: Rhodanian, from which Frédéric Mistral [1830–1914] designed his literary koine; Maritime Provençal, Nice and Forcalquier at the threshold of the Alps)[3]. Following this taxonomy, Ronjat further divided each group or dialect into complex sets, as in the case of Languedocian, endowed with no less than 16 subgroups (or subdialects), of which at least two encapsulated three "branches": Western vs. Eastern Languedocian vs. Guyennais at a higher level under his $\alpha$ complex, as opposed to specific varieties, such as Comté de Foix (excluding Couserans), Capcir and Villefranche under a $\theta$ complex. In all, Ronjat's classification of what we today call "Occitan dialects" displayed no less than 5 major dialects (from A to E), approximately 19 subdialects and dozens of varieties we would today call *locolects* or at best *town dialects* (see spots on maps, Figure 1 above) – hierarchy is not always clear-cut between the two levels of subdivisions, rural and urban, since Ronjat contrived his nomenclature for the sake of indexing sets of phonological and structural variables (isoglosses), rather than for classification proper. He also took into consideration the buffer zone he dubbed (in 1913 already) "Croissant" (Crescent) at the top of the map, above the D dialect, where Oïl and Oc diasystems merge into a mixed dialect. In many aspects, his work is a landmark in Gallo-Romance dialectology and dialect classification, and can be compared to major breakthroughs such as that of Graziadio Isaia Ascoli

1 Occitan is one of the three major Gallo-Romance languages spoken in central-western Europe.

2 The "appendice" inserted in vol. IV (1941) of his comparative grammar of Occitan dialects, in Ronjat (1930–41), dedicated to dialect variables in the Occitan diasystem–although Ronjat couldn't have used the term, since his comparative grammar was published over a decade before Uriel Weinreich's pioneering article (1954). Pierre Bec's approach to the dialectical network as a complex and hierarchized *diasystem* is most explicitly formulated in his 1972 programmatic article published in Occitan.

3 See Sumien (2009), who provides an explicit account of this intricate taxonomy, which was initially contrived by Ronjat as an editorial tool to index each variety, subdialect and dialect throughout his comparative argumentation on phonology and grammar of Occitan, in his monumental essay, published between 1930 and 1941 by the *Revue des Langues Romanes*, Montpellier. Sauzet (2016) provides a nice contextualization of Ronjat's conception of Occitan as a genuine complex language, highly hierarchized into clear cut dialect components, in straightforward opposition to the prejudices of his time.

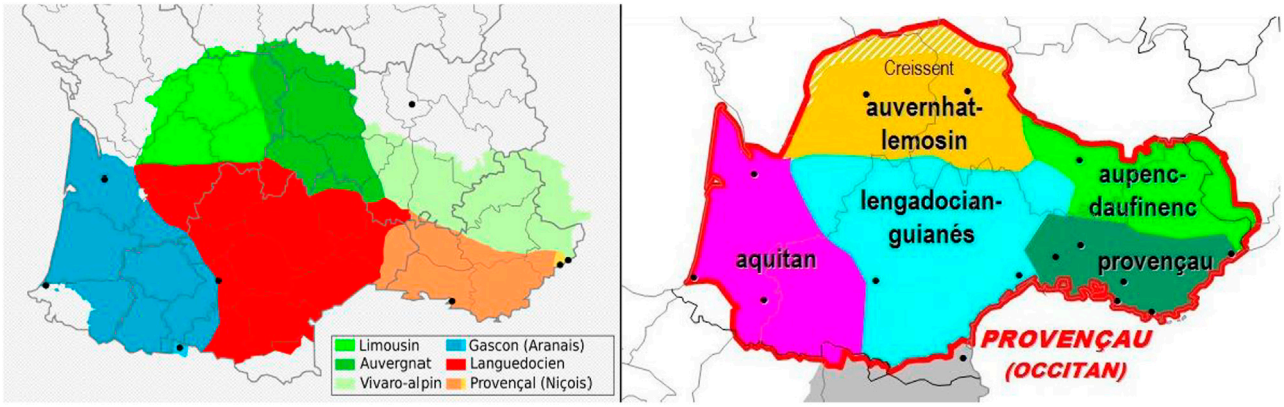



FIGURE 1
Occitan dialects (on the right: Ronjat's taxonomy, translated into Occitan). Left panel: reproduced from Jfblanc via Wikimedia Commons, licensed under CC BY-SA 4.0. Right panel: reproduced from Domergue Sumien - Revue Linguistica Occitana via Wikimedia Commons, licensed under CC BY 3.0.

**TABLE 1** Granular (Ronjat) vs. Reductionist (Bec) methods and models applied to the Occitan Dialect Continuum.

| Criterion | Ronjat's classification | Bec's classification |
|---|---|---|
| Primary Division | Five major dialects (Groups A to E) | Two macrodialects: Arverno-Mediterranean (ArM) vs. Aquitano-Pyrenean/Central Occitan (AqP/CO) |
| Subclassification | Dialects into branches and subgroups | Typological traits leading to different hierarchical models of dialects |
| Basis of Division | Strong "neogrammarian laws" combined with detailed phonological variables and some morphological traits (inflectional endings) | Phonological, morphosyntactic variables, ethnolinguistic criteria (as in Figure 2B). Strong "neogrammarian" phonetic laws, such as CA/GA palatalization, betacism, etc. |
| Mapping of Dialects | Fixed, clear-cut geographic boundaries, as in Figure 1 above | Vicarious hierarchies with abstract boundaries, as in Figures 2A,B above |
| Notable Features | Focus on philological isoglosses, mainly provided by the Felibrean literary tradition, in addition to Gilliéron's ALF (Atlas Linguistique de la France, 1902–10) and his own sporadic fieldwork observations | Bold reductionist approach, enhancing "near-decomposable" systems. Approach rooted on Weinreich's notion of diasystem. Nevertheless, as Ronjat, data mostly from ALF and current philological sources. Ronjat's major essay (1930–41) as a source of inspiration too |

(1829–1907) with his discovery of Francoprovençal as a third sub component of the domain (Ascoli, 1875; Wüest, 2003).

Interestingly enough, at the end of the 19th Century and at the very beginning of dialectology in France, Paul Meyer and Gaston Paris' doctrine, known as the *Dialectal Continuum* postulate greatly contributed to spreading the idea that dialects were merely abstract constructs, deprived of any substantial reality, especially in the Gallo-Romance domain, which Gaston Paris compared to a "tapestry" where linguistic traits melted into one another (Paris, 1888). This denial of the existence of dialects amounted to a sort of null hypothesis, induced by the French monolingual ideology inherited from the French Revolution, and it has been contradicted many times since–notably by Camproux (1962: 759–762) and especially by the bulk of dialectometric studies. Yet it had made a strong impression on the young Ferdinand de Saussure, who supported it enthusiastically in his lectures on general linguistics and in public presentations (Saussure, 1891; Saussure, 1913), in spite of Ascoli's seminal essay on the ontology of Francoprovençal, to which Gaston Paris' solemn talk at the famous "réunion des sociétés savantes" in 1888 was an *ad hoc* denial. Beyond ideological factors in this particular case, Null Hypotheses may be heuristic claims, when used to start examining intricate empirical questions. CT may bring relevant insights, and may be challenged at the same time, given the variegated array of results data processing can bring to the debate. As we shall see, not only do *dialects*, *subdialects* and *varieties* exist, –they can even be hierarchized in many different ways, but their ontology paves the way for many heuristic concepts relevant to CT, among which *hubs, small-worlds, singletons*, etc., as in Table 3, Section 3.8 below.

The territorial structure of the Occitan domain, which is both huge[4] and dense, also implies a high degree of complexity, in terms of external factors: three variegated highland complexes (the Massif Central in the Central-eastern zone, the Pyrenees in the South and the Alps in the East), major rivers, such as the Garonne, Dordogne, Loire and Rhône, two seas (the Atlantic Ocean and the Mediterranean), innumerable plains and piedmonts linked to highland vertices, conspicuous realms of the past and towns with

4 Roughly 190 000 km2 (Metropolitan France = 551 695 km2). Compare to the surface area of countries such as Senegal 196 712 km2 and Uruguay 173 626 km2.

a dense history (Bordeaux, Pau, Bayonne, Toulouse, Marseille, Valence, Grenoble, Clermont-Ferrand, Limoges, etc.), international borders with Italy and Spain, etc. These are specific conditions generating a complex interplay of interactions between rural and urban communities, regions and subregions, providing a unique opportunity for CT to apply tools and methods, as we intend to do here.

Section 2 below introduces the reader to the main canonical subdivisions of Occitan. Section 2.1.1 breaks up conventional entities and challenges the Occitan Dialect Classification (hence, OCD) with Pierre Bec's macrodialects–or superdialects–, introducing a more reductionist hierarchization of the main and subcomponents of Occitan as a major *diasystem*–although Bec more specifically used Weinreich's term for the Gascon dialect as a subcomponent of the Occitan domain (Bec, 1973: 26 § 20). In 2.2.2, Bec's major variables are presented as typological traits founding two alternative great divides between (i) the so-called Arveno-Mediterranean (ArM) vs. the Aquitano-Pyrenean + Central Occitan (i.e., Languedocian, hence AqP/CO), (ii) Northern vs. Southern Occitan. These traits are listed as a typological block, according to the former division. In section 3, we address one of these primary components of the diasystemic supercoordinates on the basis of two archetypical lemma (*castanhièr* and *abella*), respectively for the reflexes of the Latin velar stops before low vowels and betacism. In Section 4, we delve into our case study, consisting in applying CT to dialect variation on the basis of the THESOC database. In Sections 3.1, 3.2, we provide the basic tenets of our THESOC phonological database of 71 items (regular cognates), 662 localities; in 3.3 the results of Hierarchical Cluster Analysis (Joe Ward's Method); in 3.4 Complete linkage; in 3.5, Group Average; in 3.6 Weighted Average. In 3.7 the main external factors accounting for part of the variegated (or *vicarious*) patterns observed with these algorithms will be enumerated, bringing us to Section 3.8, where we will sketch a model of properties for the ontology of dialects, subdialects, varieties and other entities. In Section 3.9 we will compare the previous results obtained through hierarchical methods to Multidimensional Scaling (Section 3.9.1) and threshold graphs, i.e., connectograms (Section 3.9.2). In 3.10, we will take advantage of some reductionist taxonomies to extract data from particularly interesting varieties, out of the THESOC database, in an attempt to use dialectometry for data mining, for the sake of dialectical typology and fine-grained typological comparison of dialects and varieties within the diasystem. In Section 4, we will draw conclusions from our endeavor both for Complexity Theory and for General Dialectology.

# 2 Canonical taxonomic methods

## 2.1 Qualitative hierarchy and geographic stratification of the Occitan diasystem

In order to better understand the data provided by the linguistic atlases of the Occitan domain, through the THESOC database and results from Gabmap as a tool for CT applied to dialectology, we will constantly keep in mind what Herbert A. Simon used to call "the architecture of complexity" (Simon, 1962), through preliminary maps and variables, as in Figure 1 above, to compare with reductionist maps in Figures 2A, B below: i) hierarchy, ii) evolutionary processes, iii) dynamic properties of components and subcomponents (or "Nearly Decomposable Systems", as Simon would call them) and iv) description or modeling of complex systems–here, in this particular case, of a *diasystem*, intended as a kind of complex system, in the realm of linguistic diversity. One heuristic way of handling these four tenets of CT applied to GD is to rely on handbooks, such as Pierre Bec's *Manuel pratique d'occitan moderne* (Bec, 1973) and an authoritative overview of Occitan (Bec, 1963), considered as the *Vade Mecum* of scholars in this field of research.

## 2.2 The isoglottic-typological approach

### 2.2.1 From nomenclature (Ronjat) to higher-ranking hierarchies (Bec)

Indeed, Bec follows the main lines of Ronjat's pioneering classification of Occitan dialects, but he ventures farther in ordering the hierarchy (*macrodialects*) and affiliation of parts of the whole (*dialects* and *subdialects*), resulting in alternative (or vicarious) descriptions of the (dia)system. In Figure 2, Bec postulates two models of superordinate areas (macrodialects): on the one hand (to the left of the figure, i.e., Figure 2A), a clear-cut North vs. South dichotomy (or great divide), with a macrosubdivision in the latter (Gascon vs. Southern Occitan as a whole, encompassing both Languedocian and Provençal, matching Ronjat's former "groups A & B", in contrast with "group" C or Aquitain). Unlike Gascon in the SW, the remainder of the three dialects in the North (corresponding to Ronjat's "groups D & E") are sparse in the upper part of the map, without frontiers, reminding thus *a posteriori* Simon's "near-decomposable" (sub)systems within the Occitan diasystem as a whole. To the right of the figure (Figure 2B), Bec suggests an alternative description or modeling of the same diasystem (Bec, 1963; 1972; 1973), opposing two macro-areas or superordinates, which he defines in geographic and ethnolinguistic terms: from NW to SE, i.e., the "Arverno-Mediterranean" macro-area, as opposed to the SW and the center of the domain, with two major components in the South, which Bec calls "Aquitano-Pyrenean" and "Central Occitan" (AqP/CO)[5]. In doing so in 1963, Bec's work marked a breakthrough in the classification of Occitan dialects. He was introducing two innovative models that provided new insights on the inner structure of the Occitan dialect continuum. Figure 2A presents a major North vs. South division, with the South further splitting into Gascon and Southern Occitan (including Languedocian and Provençal), while the North displays a continuum of more dispersed dialects. Figure 2B, in turn, offers an alternative view, identifying two major zones, with more ethnolinguistic content: the so-called

5 See Zufferey (2008) for a presentation of the Occitan diasystem within a broader framework (Gallo-Romance as a whole). Moreover, besides Bec's and Ronjat's taxonomies, a pioneering study by Léon Lamouche (1901) is still worth citing. Although Ronjat criticized Lamouche's intuitions, his own classification owes a lot to the former. NB: both authors were wrong in including Catalan in their taxonomy (see Guiter, 1973: 80).

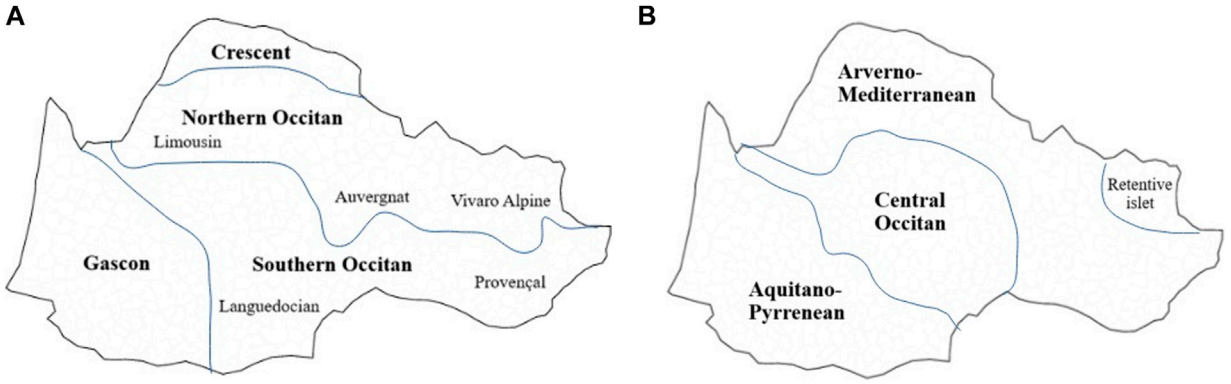



FIGURE 2
Pierre Bec's maps of vicarious hierarchies of Occitan dialects (Bec, 1963). **(A)** Bec's North vs South Great divide. **(B)** Bec's NW-SE to central and SW oblique divide. The maps are based on data originally presented in Bec (1963), substantially modified for the sake of this article.

"Arverno-Mediterranean" area (obliquely from Northwest to Southeast) and the "Aquitano-Pyrenean" and "Central Occitan" areas in the Southwest and central regions. In Figure 2B, Bec points at a kind of "default area", on the basis of evolutionary processes defined in his own terms as an "ilot de conservatisme" ("retentive islet"), reminding Simon's second tenet above. This variable hints, among other traits, at the robustness of etymological final stops, and confirms the isoglottic and strongly philological nature of Bec's taxonomy.

Both Figures 2A, B are empirically accurate and intercomplementary, together offering a richer, more nuanced understanding of Occitan dialects than previous classifications or nomenclatures, which relied mostly on "eponymous dialects".

Bec's approach can be seen as bold and daring, because he challenged the traditional academic view that each Occitan dialect should be well-defined and distinct. Instead, he introduced a heuristic reductionism by proposing broader classifications and higher-level hierarchies to encompass the dialect continuum. What is remarkable in these maps is the *epoché*, in phenomenological terms (i.e., *Tabula Rasa* of canonical views or prejudices), Bec decides to apply to the issue of defining and placing dialects and subdialects on a map, as all dialectologists had done so far. From the standpoint of CT, he consistently follows what Simon calls a "near decomposable" approach to ethnolinguistic categories such as "dialects" and "subdialects", as sketched out in the first map to the left of Figure 2 (i.e., Figure 2A), where these (near) subcomponents are left without definite borders. His approach is purely linguistic and, as such, highly abstract, although anchored in phonological data–as suggested by the label "îlot de conservatisme", i.e., "retentive islet" in the upper right side of the map, where "Vivaro-alpin" or "Provençal Alpin" would have been expected. In both maps, phonological criteria are used for isoglosses, such as Latin CA-, GA-palatal vs. non palatalized onsets, betacism (*b, β, v* > *b*), palatalization of Latin inner cluster CT, degrees of palatalization or affrication/disaffrication of onsets, etc.

Table 1 outlines a contrastive glimpse at the main differences between the classification approaches of Jules Ronjat and Pierre Bec, highlighting their similarities, differences, and the implications of their models on the understanding of Occitan dialects.

Ronjat's classification, on the one hand, offers a detailed, granular approach, focusing on the fine distinctions between individual dialects based on phonological and structural features. This provides a solid foundation for understanding the specific characteristics and boundaries of each dialect. Bec's classification, on the other hand, introduces a broader perspective by grouping dialects into larger macrodialects, using a combination of phonological, morphosyntactic, and ethnolinguistic criteria. This approach highlights overarching patterns and relationships, offering a more holistic view of the dialect continuum. Together, the two approaches provide a comprehensive framework: Ronjat's detailed mapping of individual dialects (the *Granular view*) complements Bec's broader, more integrative approach (the *Reductionist view*). This combination of standpoints allows for a deeper and more nuanced understanding of Occitan dialectology, balancing fine-grained distinctions with larger structural patterns. As we'll soon see, dialectometric tools allow to go ever far beyond canonical views on nomenclatures of a dialect continuum such as Occitan. Yet, both approaches (qualitative and quantitative) are undoubtedly powerful–especially diasystemic modeling–and the application of CADS principles provides a heuristic metatheory, highly relevant for CT.

### 2.2.2 Modeling Bec's typological traits

We will now have a closer look at these qualitative criteria, which we will adapt to our CT approach, in terms of structural variables within the diasystem, and how we can define and hierarchize these components as the building blocks of Occitan dialect classification. Nevertheless, our purpose will be to go beyond these constructs, which Nerbonne and Kretzschmar (2003) call "Eponym dialects", which surface in handbook maps and canonical descriptions such as the one initially proposed by Ronjat, to look for another type of component–or, more properly here, *near decomposable components* or NDC, which Nerbonne & Kretzschmar (*ibidem*) call "the invisible dialects". In terms of CT it means we will be in search of *emerging* structures. We will focus on the macro areas identified by Bec as they appear on the second map (Figure 2B): the "Arverno-Mediterranean" complex (ArM) to the right of the map vs. the Aquitano-Pyrenean/Central

TABLE 2 Advantages and disadvantages of hierarchical clustering methods.

| Hierarchical clustering method | Advantages | Disadvantages |
|---|---|---|
| **Hierarchical Cluster Analysis** (WM) Hierarchy | Unravels roughly symmetric hierarchical relationships between dialects | Can produce clusters that are too detailed and specific. It might therefore create more categories or groups than necessary, making it harder to grasp broader patterns of intricacy in the data |
| Visualization | Provides a clear-cut visual representation of dialect clusters through dendrograms, at least in the surface | May produce dendrograms that are difficult to interpret, particularly for readers unfamiliar with cluster analysis techniques |
| Categories | Allows for the identification of dialects and subdialects | Selection of the appropriate linkage criterion (e.g., here, Ward's Method) can impact the resulting dendrogram |
| **Complete Linkage Hierarchy** | Tends to create compact, uniform clusters | Enhances the "chaining phenomenon", where clusters are stretched out along a chain-like structure, blurring granularity. Prone to unravel deep *flat hierarchies* in the quantitative dataset. |
| Visualization | Maintains a clear separation between clusters, making it suitable for identifying distinct dialect groups | Sensitive to outliers, which may affect cluster formation and generate asymmetries |
| Categories | Highlights sharp boundaries between dialects, helping in the identification of dialectal boundaries | May not be suitable for datasets with non-linear or irregular cluster shapes, enhancing *a posteriori* erratic (smaller) cluster or "default dialects" |
| **Group Average (GA) Hierarchy** | Creates balanced clusters by considering the average distance between all pairs of data points in two clusters | Sensitive to the presence of outliers, which can skew cluster formation, but provides clues on singletons and exclaves, default dialects, intensively contact lects or alloglottic lects |
| Visualization | Preserves discrete clusters and outliers, enhancing the representation of dialect diversity | Can produce clusters with varying sizes, leading to imbalanced groupings |
| Categories | Provides a smoother representation of dialect clusters compared to Complete Linkage | May not be suitable for datasets with highly unevenly distributed data points –this is not the case here with our THESOC dataset, as suggested by tools 1–2 (section 3.1) and statistical distribution of values in Figure 7 and section 3.2) |
| **Weighted Average (WA) Hierarchy** | Combines aspects of Complete Linkage and Group Average, offering a balanced approach to cluster formation | Requires careful consideration of weighting factors, which can introduce subjectivity into the analysis |
| Visualization | Maintains the overall structure of the taxonomy while enhancing broader subgroups | May not capture nuanced relationships between dialects as effectively as other methods |
| Categories | Offers a compromise between preserving discrete clusters and capturing broader trends in dialect distribution | The effectiveness of WA may vary depending on the weighting scheme used |

Main clustering methods are highlighted in bold in the table, in terms of visualisation and enhancement of the categorical properties revealed by the algorithm.

Occitan complex (AqP/CO), centered on Gascon and neighboring varieties, to the left of the same map, following a line from Bordeaux in the NW to Narbonne in the SE. The former set of *diasystemic variables* (here, isoglosses) will be called Block I, whereas the latter will be called block II. T stands for *trait*; at the end of each line describing the variable, we characterize the diasystemic natural class of dynamic process by an abbreviation[6]: e.g., T1 = {-CT-}$^{\text{VOC|PAL\_AFFR}}$. Symbols ≡ and ≈ stand for inherited stability of patterns vs. intense innovation or restructuration, such as for morphosyntax (MS) ArM T5: {MS≈} vs. AqP/CO T8 {MS≡}.

(1) Bec's Typological Traits defining the Great Divide in the Occitan Space (Bec, 1970: 18-20)

(1.1) Block I (ArM):

T1: Palatalization and affrication (encoded here with exponent $^{\text{PAL\_AFFR}}$ in DS label) of implosive Latin cluster -CT- (or *nexus*), ex: FACTU 'done, deed' > Oc. <*fatʃ*>. Diasystemic label: {-CT-} $^{\text{PAL\_AFFR}}$.

T2: Palatalization of Latin velars before low vowel CA-, GA-: **c*, **ɟ*, > *ʤ*. DS label: {K/G} $^{\text{PAL\_(AFFR)}}$.

T3: Distinction/b/ ~/v/. DS label: {B:V}.

T4: Final and medial consonant dropping. DS label: {C→ 0/_ #, V_V}.

T5: Restructuring of inherited morphosyntactic patterns (articles, number, verbal inflection, etc.). DS label: {MS≈}.

T6: Widely predominant -*e* ending in 1SgPR.IND, except in Maritime Provençal (-*i*) and Alpine Provençal (-*o*). DS label: {-E/-O/-I}$^{\text{1SG.PR.IND}}$.

This macro area ArM is therefore grounded on a natural class which can be described through the following declarative formula:[7]

6 We use the same kind of indexation of diasystemic variables in Léonard, 2016, for other Gallo-Romance varieties (Oïl dialects in NE France at the border with Belgium).

7 Abbreviations: DS stands for Diasystem, PHON for Phonology and MS for Morphosyntax.

DS I: {PHON: {T1: {-CT-}$^{AFFR}$, T2: {K/G} $^{PAL.AFFR}$, T3: {B:V}, T4: {C→ 0/_#, V_V}}; {MS {T5: {MS≈}, T6: {-E/-O/-I}$^{1SG.PR.IND}$}}

(1.2) Block II (AqP):

T1: Non palatalization of Latin velar onsets before low vowel (CA, GA). DS label: {K/G≡}⌐$^{PAL}$.

T2: Voicing of implosive stop in Latin -CT-cluster, ex: FACTU 'done, deed' > Oc. <*fajt*>. DS label: {CT}$^{VOC}$.

T3: Low Vowel Raising as a subsequent ordered rule from T2: <*fajt*> → <*fèjt, hèjt*>. DS label: {EYT}$^{Raising}$.

T4: Disaffrication of Latin palatalized voiced velar *ɟ > ʤ > ʒ: <ʤet'a> 'throw' → <ʒet'a>. DS label {J}$^{DISAFFR}$.

T5: Betacism, or labial obstruent merger: <*b, v* > → <*β*>. DS label: {B}$^{MERG}$.

T6: Final consonant retention or robustness. DS label {C≡/_#, V_V}.

T7: Palatalization in ʃ of Latin cluster -IS-: *pareiser* → *pareiʃer*, *peis* → *peiʃ* 'seem'. DS label: {(I)SH}.

T8: Robustness of inherited complex morphosyntactic patterns (articles, number, verbal inflection, etc.). {MS≡}.

T9: *-i* ending in 1SgPR.IND. *kant-i, bat-i, ves-i*: 'I sing', 'I hit', I see'. DS label: {-I}$^{1SG.PR.IND}$.

T10: Prepositional marking of personal agreement object (AGRO $^{[+hum]}$): *l'aimi a mon paire* 'I love my father'. DS label: {AGR.O}$^{PREP.\ hum}$.

In turn, the AqP macrodialect comprehends the following diasystemic traits, summed up in a declarative formula:

DS II: {PHON: {T1: {K/G≡}⌐$^{PAL}$, T2: {-CT-}$^{VOC}$, T3: {EYT}$^{Raising}$, T4: {J}$^{DISAFFR}$, T5: {B}$^{MERG}$ T6: {C≡/_#, V_V}, T7: {(I)SH}}; {MS {T8: {MS≡}, T9: {-I}$^{1SG.PR.IND}$, T10: {AGR.O}$^{PREP.\ hum}$}.

This list should be read as a fragment of the *diasystemic grammar* of Occitan in the broad sense: for instance, in block I (ArM, i.e., Arverno-Mediterranean), traits 1 to 4 (T1-4) are building blocks of syllables (onsets and nuclei) for making up words, while T5-6 describe central components of the grammar, through morphosyntax (T5) and verb inflection (T6). From a diasystemic standpoint, we consider that all these traits–phonological and grammatical–defined in the philological tradition, as in Ronjat and Bec's essays, where they have been given the status of isoglosses, are actually part of *grammar*. A grammar of a specific nature: a *polylectal grammar*, hence PG (Berrendonner et al., 1983; Puech, 1979; Diller, 2006; Evans, 2003; Mühlhäusler, 1992), in other words, a *diasystem*–see Weinreich (1954); Grassi and Telmon (1979); Léonard (2020a). Both terms (PG & DS) were initially coined to define *grammar* considered not only as corresponding to the description of a single norm (language, dialect), but above all, as a holistic complex integrating all the norms and sub-norms (*lects*) and structural patterns (phonological, grammatical, lexical, morphosyntactic, semantic) within a comprehensive framework enabling the linguist to fully describe the inner structural diversity of a dialect network.

## 2.3 From qualitative to quantitative description in dialect classification

These concepts (*polylectal grammar, diasystem*) enable exploration of the complexity of variation within any language or linguistic domain. This radically differs from the classical descriptive tradition consisting in presenting a canonical variety placed above all other constitutive geolinguistic, ethnolinguistic and sociolinguistic subsystems. While any language is already a complex system in itself, with its two major strands–phonology, with an inventory of, say, 6 to 122 consonants (see WALS' chapter 1 Online: https://wals.info/chapter/1), 2 to 14 vowels (see WALS' chapter 2 Online: https://wals.info/chapter/2), dozens of affixes and clitics and about 100 morphosyntactic patterns (see https://wals.info/feature), and over 150 000 words as basic lexicon for any language–, these building blocks can be multiplied by the number of (dia) lects available in a dialect network.[8] As "dialect variation is distributed across the grammar but in uneven portions" (Dunn, 2023: 19), the intensity of this complex combinatorics is somehow unpredictable. Nevertheless, as soon as one grasps the core of the stratification for one language, predictability comes back into play, as for any complex system, while preserving a margin for unpredictability and fluctuation (Polian et al., 2014).

Bec's variables are also to be found in Ronjat's seminal work on Occitan Dialect Classification (ODC).[9] The typological traits above (DS blocks I and II above in Section 2.2.2.) are indeed heuristic dialectal features based on which a dialect classification endeavor may be undertaken. Although "no individual node within the grammar captures dialectal variation as accurately as the grammar as a whole" – another statement by Jonathan Dunn (*op. cit.*: 18), with which we broadly agree, some variables and corresponding (phono)lexical items may turn out to be highly indexical, as in the spatial distribution of Southern Occitan (T1) & AqP/CO (T5) in Figure 3. By integrating these typological traits within a CT framework, we consider both phonological and grammatical traits as part of a *polylectal grammar*, i.e., a comprehensive system that accounts for the structural diversity of the dialect network. This approach allows us to see the dialects not just as isolated entities but as parts of a dynamic, interconnected system, providing a deeper understanding of Occitan dialect variation. Using Gabmap, we generated distribution maps for these features, showing how classical "strong isoglosses" (Bec's typological traits, in [1] data set above in 2.2.2.2) can be very insightful indeed. The areas highlighted by these features match the sophisticated hierarchies obtained through advanced clustering methods like Complete Linkage or Group Average, as we'll soon see, in Section 3 below (Sections 3.4 and 3.5, specifically). The mapping approach in Figure 3 nevertheless demonstrates how traditional phonological features can effectively reveal the structure of dialect macroareas, providing a clear, empirical basis for understanding dialect variation dynamics.

The dark areas at the bottom of the maps roughly match the extension of two major features structuring, on the one hand Bec's broader Southern Occitan (as in Figure 2A in Section 2.2.1. above), and on the other hand, the Aquitano-Pyrenean area (as in Figure 2B): T1 (retention of velar stops before low vowel) and T5 (betacism) –the latter being considered by P. Bec as characteristic of the "Iberian type" of this macrodialect. The term "Iberian" has been sporadically suggested in the

8 See also Miestamo (2017).

9 See also Allières (2003: 223–237) for a more up to date survey of diasystemic variables in the Occitan dialect network, and Zufferey (2008).

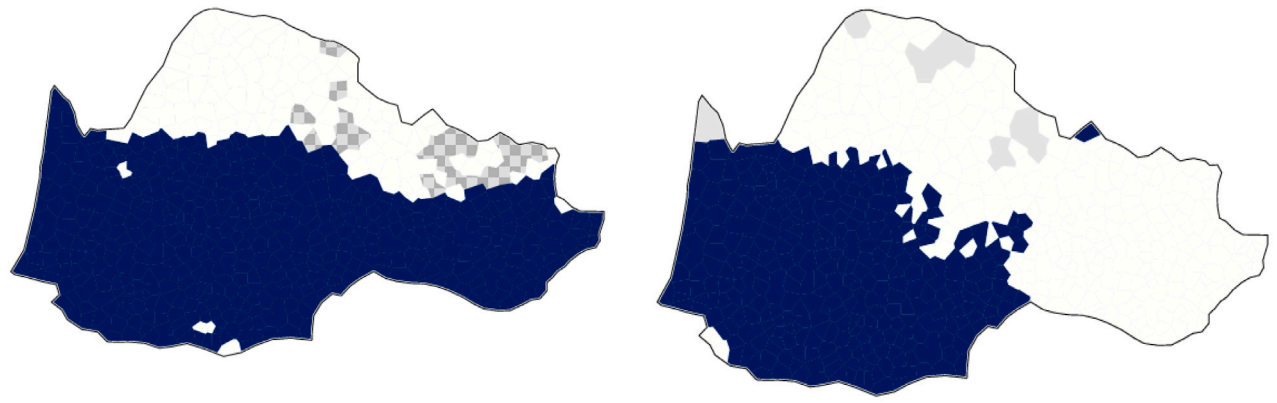



FIGURE 3
Block II. A sample of AqP/CO. T1 & T5, THESOC, Gabmap (distribution map).

literature on Occitan phonology, half as a typological trend and a substratum in continuity with Ibero-Romance languages, and might therefore be misleading if intended as such. Here, we consider this trait as a mere phonological trend of spirantization of lenis voiced stops *b, d, g*, surfacing as approximants *β, δ, γ*, as in spoken European Spanish and Portuguese, indeed, but without necessarily any substratic implication. The "Iberian substrate hypothesis" has also been advocated by D. Sumien (2009), embracing Gascon and Languedocian together with Catalan, in spite of Guiter's dialectometric results invalidating the hypothesis of Catalan as part of the Occitan diasystem[10].

# 3 Results

## 3.1 A complexity theory approach: vicarious models

Distribution maps of variables attached to a few lexical items as in Figure 3 above highlight at the same time the aprioristic approach of dialectologists, who rely on few qualitative variables to build up their models of dialect classification (as the two lists above of DS I and II, in [1]) and the highly heuristic power of these isoglosses.

However, not all areas are so clear-cut, and we know that isoglosses tend to fluctuate, depending on the principle of lexical diffusion or contact (adstrats and superstrats): the T1 variable, for example, of non-palatalization (in [1.2] above, within Block II, i.e., in the AqP macrodialect), would show a much less unitary southern macro-area, with a considerable gap to the SW, where Gascon has adopted lexical superstrat forms of the *chapew* [ʃapɛw] type ('hat'), with a palatal initial onset /ʃ/ (88 items in the THESOC database, all located in Gascony, where we would have expected the Gascon forms <*capeu, capeth*>). We will call these word-scale restrictions "phonolexical" forms.

In order to reduce the bias caused by fluctuations in the spatial distribution of isoglosses and to grasp all the units that make up the diasystem, a quantitative dialectometric approach is needed. Our main tool for this purpose here will be edit distance, using the free online software Gabmap (https://gabmap.let.rug.nl/).[11]

10 We also toe the line of Henri Guiter on this issue: our own results, from Louis Michel's Roussillon Catalan data (Léonard and Albinet, 2023), applying Levenshtein distance (categorical instead of string tokens, i.e., like Guiter's so-called "Méthode Globale"), read as follows: compare inner diversity of Northern Catalan, e.g., Banyuls'indices of edit distance with Port-Vendres 0.0294,118, Collioure 0.0196,078, Argelés 0.0392,157, St-Cyprien 0.0784,314, Bacarès 0.0980,392, as opposed to Catalan vs. Occitan: Banyuls vs. Occitan Lang. Narbonnais Leucate, 0.892,157 Narb. La Nouvelle 0.901,961 Narb. Grussian 0.901,961 Narb. Fleury 0.901,961; Biterrois Valras 0.941,176, Agatois Agde 0.960,784 Montpellierain Sétois 0.970,297 Mtp. Palavas 0.970,297, Provençal Grau-du-Roi, 0.99, Les Saintes 0.990,196, Les Martigues MTG 0.970,297 (original data from Michel, 1964).

11 See Leinonen et al. (2016) on processing dialect data with Gabmap, and Levenshtein's paper (1966) seminal paper on the premisses of edit distance. See Dubert and Sousa (2016) about the three main tools nowadays available for dialectometry Online: *VDM* (Visual Dialectometry, from the Salzburg School of dialectometry), *Gabmap* (from the Groningen school) and *Diatech* (from the Bilbao school), and for an appraisal of the advantages and disadvantages of qualitative vs. quantitative approach in geolinguistics. Moreover, let's mention that Bayesian methods, as applied by Hartmann to Germanic languages, may provide an even broader horizon for CT than classical dialectometry; yet they are more currently used for comparative purposes, rather than for dialectology proper (Hartmann, 2023).

### 3.1.1 Structure of the data

As Guylaine Brun-Trigaud puts it, in her contribution to a forthcoming paper (Léonard, 2024), the THESOC (Thesaurus Occitan) database was established in the 1990s by Jean-Philippe Dalbera (Université de Nice, Fr) under the auspices of the CNRS (Centre National de la Recherche Scientifique, France's largest governmental research organization, dedicated to advancing scientific knowledge across various disciplines). This endeavour emerged after a failed early attempt to digitize complex phonetic data, which led to the disbandment of the Atlas Régionaux group within CNRS and left a wealth of data unpublished. THESOC aims to centralize and make accessible all regional atlas data from the Occitan domain. Currently, THESOC houses nearly 1.5 million data tokens and continues to integrate unpublished data, which is accessible via its website (http://thesaurus.unice.fr). The database encompasses approximately over 8,000 notions from the regional atlases, though these figures don't fully capture the extent of comparable data. The Occitan surveys were conducted using three different questionnaires, each with its own proponents. Gardette's questionnaire was initially used, followed by Dauzat's, which influenced the ALG surveys. Finally, Nauton developed his own questionnaire, subsequently adopted by several atlases (ALMC, ALAL, ALP, ALLOc, ALCe) and partially by ALLOr, which eventually conformed to Nauton's format. Despite the diversity in methodologies, only 400 maps are common across the atlases strictly within the Occitan domain. The 71 cognates used here to implement phonological analysis with Gabmap through edit distance belong to this narrow list of shared items in the THESOC database (see list in [2.1], section 3.2 below). *Gabmap* stands as a witty label for the application online, conveying the idea of "talkative/eloquent maps".

### 3.1.2 Source of the THESOC data and processing of tokens

We analyzed 71 entries (word forms) from the Occitan THESOC database using Gabmap. Here are the basic tenets of the data processing on Gabmap:

- **Places**: 662 locations (dialect varieties, or lects) were studied.
- **Items**: we examined 71 specific items/phonological cognates (entries in the THESOC database).
- **Instances**: throughout our study, edit distance was applied to these items a total of 44,748 times.
- **Characters**: the THESOC corpus analyzed here contained 245,029 characters in total.
- **Unique characters**: there were 99 different characters used.
- **Tokens**: the total number of tokens counted was 240,562.
- **Unique tokens**: among these tokens, there were 203 distinct types used.

These results provide insights into the frequency and variety of language elements found within the Occitan phonological THESOC database, helping us understand the linguistic landscape of the region more comprehensively, as in the GIS maps below see Figure 4 (GIS map).

On the left of Figure 5: polygonization. On the right: localities.

We used Edit distance (function *string data, tokenized*) with Gabmap to analyze the phonological data from our THESOC database. This method helps us measure how similar or different words are by counting the minimum number of operations needed to transform one word into another. These operations include deleting ($x \leftrightarrow 0$), adding ($0 \leftrightarrow x$), or substituting ($x \leftrightarrow y$) characters. For example, if we compare the word "abeille" (French for 'bee') to different entries in the database, the numbers next to each pair of words show the edit distance scores. A lower score indicates that the words are more similar, requiring fewer changes to match each other. Here's concretely what each score means, applied to a sample of string tokens from the function "alignments" in Gabmap, according to the three basic operations of *Deletion* (removing a character), *Addition* (adding a character), *Substitution* (replacing one character with another):

Each entry in the database undergoes this comparison process against a set of standard words or patterns, helping us understand patterns of variation and similarity within the dialect network under scrutiny in a holistic way, far beyond limited sets of isoglosses or traits, as in Ronjat's or Bec's philological tradition. Applying these trivial computational techniques, we unravel both commonalities and unique features across the Occitan-speaking world–or, more technically, across its *diasystem*. In the grids below, various localities from the *Atlas linguistique de Provence* (ALP) are compared to other lects according to the data available in THESOC from e.g., Gascony (ALG: *Atlas Linguistique de Gascogne*), Eastern Languedoc (ALLOr: *Atlas Linguistique du Languedoc Oriental*) or Western Languedoc (ALLOc: *Atlas Linguistique du Languedoc Occidental*), on the basis of lexical item *abelha* 'bee'. Every subset of alignment may provide 0 or differential scores (see Figure 6), as here lects ALP 1 (Hauterives, Drôme) compared to ALLOc 12.06 (Onet-l'Eglise, Sébazac-Concourés, Aveyron) with a score of 4, out of substitutions $a \leftrightarrow ɔ$, $v \leftrightarrow \beta$, $ɛ \leftrightarrow e$, $ə \leftrightarrow ɔ$, corresponding to at least four phonological variables, as pretonic vowel raising, betacism, tonic vowel lowering and degrees of posttonic vowel reduction, summing 4 points for this single pair ALP 1 vs. ALLOc 12.06.

An edit distance, as with Gabmap therefore enables a comprehensive examination of both micro and major isoglosses throughout the (Occitan) dialect network. This method goes beyond focusing solely on specific variables like the six variables in Bec's Arverno-Mediterranean block I of neogrammarian rules (1.1) above, or the 10 variables in Bec's Aquitano-Pyrenean macrodialect in Block II (in 1.2). Instead, it covers a broad spectrum of linguistic features, providing insights into how language (i.e., the *diasystem*) varies across different areas. By systematically analyzing these linguistic features at the scale of the relation of every single lect to the rest of the dialect network, we can map out the intricate patterns of variation and identify significant as much as unexpected differences within the dialect continuum. We get a comprehensive distance matrix, out of the comparison of 71 phonological cognates across 662 locations, from which we can visualize how lects group together based on their holistic linguistic profiles. Ward's method, for instance, as in Ronjat's granularism, minimizes variance within clusters and maximizes variance between clusters, providing a basic hierarchical structure of lect relationships. Complete linkage and group average clustering seize broader similarities between lects and groups of lects, while MDS (Multidimensional Scaling) projects lects onto a lower-dimensional space to suggest embedded patterns of similarity and dissimilarity between *dialects*, *subdialects* and *singletons* (islets).

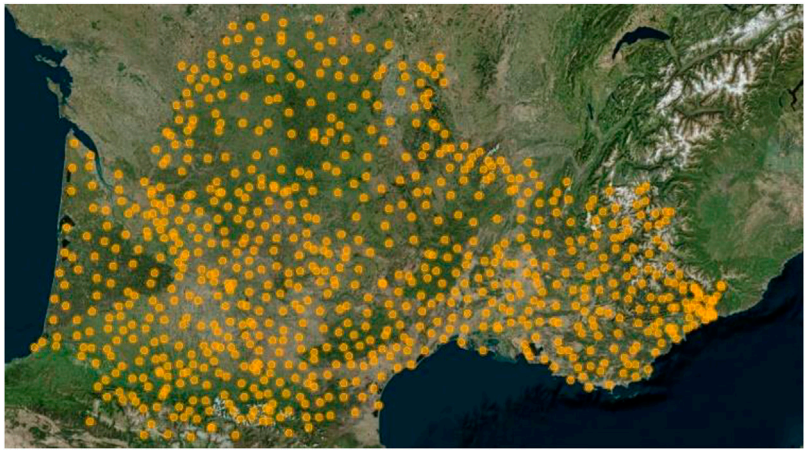


FIGURE 4
Tools (1): THESOC GIS map[12].

## 3.2 Overall structure of the phonological database

Figure 7 shows the distribution of differences, with mild positive skewness, accounting for the robustness of the base, and the density of the edit distance interactions network between varieties in the Occitan domain as represented by our THESOC data–a fine grained reading of patterns in this map allows a careful reader of Ronjat's intricate ODC to spot most of the dialects, subdialects and varieties identified by this author as "groups", "branches" and "subgroups". In our quantitative survey, we'll first use a set of hierarchical methods of classification applied to phonology (see dataset [2] below), starting with the canonical algorithms used in dialectometry, such as *Ward's Method* (Section 3.3) and *Complete Linkage* (Section 3.4), proceeding with additional tools provided by Gabmap, such as *Group Average* (Section 3.5) and *Weighted Average* (Section 3.6), before employing stochastic methods, such as *Multidimensional Scaling* (Section 3.9.1).

Dataset 2 displays the list of patterns based on Latin etyma, as previously done for several lists of phonological criteria relevant to the diasystem under scrutiny (set of rules 1.1 above, Section 2.2.2):

(2) List of lemmas analyzed through the THESOC 71 lemma database:[13]

(2.1) List of the 71 lemmas used for the treatment of phonological isoglosses in the THESOC corpus: *bee, magpie, lamb, to lamb (flock of sheep), needle, tree, wheat, beef, quail, hat, chimney, shirt, horse, chestnut tree, goat, sky, scissors, key, neck, knife, thigh, butt, sheet, dice, water, staircase, fire, leaf, gall, thread, liver, hay, make, pitchfork, cold, knee, sheaf, acorn, wasp, wool, milk, lye, hare, moon, honey, fly, mule, ripe, blackberry, nest, walnut tree, eye, egg, goose, bird, stone, to rain, to fold, meadow, well, sun, supper, to sweat, soot, cow, calf, wind, viper, donkey, ladder, star*. These lemmas cover the entire range of phonological isoglosses one can expect from this Gallo-Roman domain.

(2.2) List of etymological patterns analyzed through the THESOC 71 lemma database: Pretonic drop, tonic -AL, final -AL, *-arb/-amb*, -ARE, -ARIU, *-aticum*, -ATUS, AU-; -B-, b/B, BL-, CA-, -CA-, -CE- (*s/z*), CL-, final *-k*, final -P, final -R, final -S, final -T, final *-ts*, -CT- > *ch/jt*, CU-, -D-, -DIA-; diphthongs A + yod, E + L; diphthongs O, O + R, I + L, O + k, O + j, O + L, O + R, O + V, O + yod; U + L; -ELLUM; F-, -F-; feminine plural; -G-, *g/ɣ*, -GN-, cluster *-dr-/ch*, -IC(U)L (US), -IC(U)LA, *-js* final; L-, -L-, -LI-; masculine plural; -MB-, metathesis *form/frum*, *k-br/kr-b*; -N-, nasal ending, nasalization ending; PL-, -QU-, -RBR, *s-* (*s/ch*), -SK,

12 Main atlases used for the THESOC database with their corresponding acronyms: ALG: J. Séguy, *Atlas de la Gascogne* (1954–1973, 180 localities, 6 vol., 2,531 maps); ALMC: P. Nauton, *Atlas du Massif Central* (1957–1961), 55 pts, 3 vol., 1899 maps); ALAL: J.-Cl. Potte, *Atlas de l'Auvergne et du Limousin* (1975–1992, 76 loc, 3 vol., 1736 m.); ALP: J.-Cl. Bouvier & Cl. Martel, *Atlas de Provence* (1975–2016, 170 loc, 4 vol., 1,358 maps); ALLOc: X. Ravier, *Atlas du Languedoc Occidental* (1978–1993, 131 loc, 4 vol., 1,198 maps); ALLOr: J. Boisgontier, *Atlas du Languedoc Oriental* (1981–1986, 86 loc, 3 vol., 980 maps). Acknowledgements to Guylaine Brun-Trigaud for providing this reference list and especially for designing the map and compiling the 71 items of cognates to feed the database used here. This data recollection is part of a broader project by three authors (Léonard et al., 2024).

13 Aknowledgement: both the selection of items and the description of variables for historical phonology in dataset (2) were provided by Guylaine Brun-Trigaud, THESOC, Université Côte d'Azur/CNRS, see http://thesaurus.unice.fr/.

TABLE 3 Model of properties for the *ontology of dialects/clusters of lects*.

| Modeling/Models | External factors | | | Taxonomy | |
|---|---|---|---|---|---|
| | Geography | | Society | Clade/Dendreme | |
| Properties | Centrifugal | Endemic | Small World | Main | Outlier |
| Macrodialect | + | - | - | + | - |
| Dialect | + | +/– | - | +/– | - |
| Subdialect | + | +/– | + | - | +/– |
| Buffer zone | - | + | +/– | - | + |
| Default dialect | - | +/– | - | - | + |
| Hub | + | + | + | - | +/– |
| Singleton | - | + | + | - | + |

*-sk-*, -ST-, -T-, *tg*, -TR-; treatment of U; U + L final, V-, final feminine vowel *-e*, final masculine vowel; tonic vowel; *-z-*.

There are numerous hierarchical clustering algorithms that can yield very different results depending on the nature of the data (Prokić and Nerbonne, 2008). Gabmap encompasses four of them that we'll use to grasp Occitan dialects as CADS: Hierarchical Cluster Analysis, i.e., Ward's Method (WM), Complete Linkage (CL), Group Average (GA), Weighted Average (WA), These hierarchical/taxonomic methods will help us to explore THESOC data to fathom how Occitan dialects are related to each other in various *scales* and according to various types *of mutual or exclusive relations*. Ward's Method, also known as the *Minimum Variance method*, is more granular, and as such, is considered as the standard approach to start any survey of dialect cluster hierarchies, especially when *congruence* with canonical methods, as bundles of isoglosses (Ronjat's and Bec's method), is required. It merges groups at each step of the analysis in a way that minimizes the increase in the sum of squares of distances of each element from the mean of its group. This method uses a variance-based approach to calculate distances between groups and tends to create groups of equal size. Complete Link, also known as the Furthest Neighbor Method, starts by merging pairs of objects that are both the furthest apart and the most similar, doing so iteratively. Group Average belongs to a category of methods known as average-linkage clustering. In GA, the distance between two groups is calculated as the average of the distances between all members of the two groups. This average is weighted according to the group's size, assuming smaller groups have less weight, while larger ones have more. Weighted Average, similar to Group Average, calculates the distance between two groups as the average of the distances between all members of the groups. However, in Weighted Average, the merged groups have equal weight regardless of the number of members in each group. These explanations are based on the study by Prokić and Nerbonne (2008), which meticulously explores the recognition of groups among dialects. Table 2 provides a concise summary of the strengths and weaknesses of each hierarchical clustering method, according to the results of our Occitan THESOC data analysis, helping readers understand the trade-offs involved in choosing a clustering approach for dialectometric analysis.

It is important to note that we do not generate heat maps in the analysis below. Instead, we use RGB color mapping (Red-Green-Blue) to represent the results of clustering methods, with the aim of providing a clear-cut visual representation of the relationships between Occitan dialects based on their linguistic features, as processed by edit distance. This method enhances how dialects are distributed geographically and make the patterns of linguistic variation within the Occitan language continuum more obvious to the reader. We'll now delve into the details of what algorithmic resources, as sketched in this section can bring to us, and how the diversity of vicarious geolinguistic and taxonomic (through maps and dendrograms) configurations they generate can contribute to General Dialectology and Complexity Theory.

## 3.3 Hierarchical cluster analysis (WM)

With Hierarchical Cluster Analysis (i.e., WM: Ward's Method), two macrodialects emerge in an oblique or transversal way, from North-West to South-East, matching Bec's Model DS II (Aquitano-Pyrenean): Northern Occitan, with a North-Western pair of dialect areas (NW: Limousin and Auvergnat; SE: Provençal) on the one hand, and the Southern Occitan macrodialect, with two Gascon subdialects in the SW: Western vs. Eastern Gascon (GW vs. GE), as opposed to the central Languedocian western dialect, i.e., Lgd(W). In turn, in the Central-Southern part of the geolinguistic domain, Lgd is divided into two subdialects: Western (LgdW) and Eastern (LgdE), making up the CW branch of the dendrogram. The former clusters with the Guyenne (Guy) subdialect already mentioned above,[14] from Ronjat's ODC, as a buffer zone between Southern and Northern Occitan in the West, whereas the latter clusters with Provençal (Prov). Interestingly enough, the Southern Occitan Guyenne

14 About the ontology of this historical region in SW France, to the East of Gironde and embracing Perigord and a conspicuous part of SW Northern Occitan, see Dartigue (1950).

subcomponent mingles with SW Auvergnat, which projects itself through tiny enclaves of the splinter type embedded both within the western Languedocian dialect and within the Guyenne buffer zone–fragmentary overlapping of Vivaro-Alpine (Viv) in NW Provençal and of Eastern Languedocian in Western Provençal also happens. Noteworthy deep green splinters south of Auvergnat (Auv) are mostly scattered throughout what can be considered the Great Cevennes complex: a Piedmont of the Massif Central (called in Occitan *La Montanha* 'The Highlands'), with intricate hill and plateau (Causses) geological structures, from Mont Lozères and Sauveterre to Rouergue (Aveyron, i.e., Rouergue) and the Massif de l'Aigoual or Cevennes proper, north of Le Vigan and Montpellier (see Cabanel, 2021: 119). Capillary orography might explain this intricate embedding of Auvergnat splinters into Western Languedocian, although the *Cluster Validation* function in Gabmap points to some empirical inconsistencies here.

All these hierarchical patterns, with their intricated inner structures including splinters and islets, would not come out with traditional isoglottic or philological methods, which focused more on configurations (DS *patterns*) than on hierarchization (DS *constituents*). These fine-grained clusters can be considered a by-product of self-organization of the conspicuous flow of interactions through the multilateral comparison of tokens[15] on the grounds of the three basic operations of Edit distance (*addition, deletion, substitution*) instead of bold sets of neogrammarian rules. Most taxonomic claims by Ronjat and Bec are nevertheless confirmed, at this level of analysis (we chose 8 intervals or DS classes, taking into account the canonical ODC in seven dialects + 1 additional "joker").[16] In the next set of hierarchical models of Occitan dialects, we will raise the number of DS classes up to 10, in order to increase the granularity of hierarchies, aiming at reaching a compromise between Bec's reductionist view and Ronjat's detailed nomenclature.

## 3.4 Complete Linkage

Complete linkage (see Figure 9) neatly aligns with Bec's first map in Figure 2 above, according to the great divide opposing Northern Occitan and Southern Occitan. On the one hand, in the Northern Macrodialect, the Croissant buffer zone (with neat tropism towards a wide array of competing Oïl dialects to the North) clearly emerges at the top of the map (red area), as a subcomponent of a clade clustering it with another mixed dialect (with strong SE tropism): the Vivaro-alpine component. Both areas are interferential zones, which further cluster with a threefold complex of what Ronjat and Bec would consider Auvergnat, in terms of an eponym dialect. Northern Auvergnat. with Clermont-Ferrand (CF) as its main urban center appears in deep green, whereas the Eastern Monts Dore show up in light purple as the eastern rural zone around CF. Both components make up the Puy-de-Dôme (PdD) cluster, opposing a (former) urban dialect (CF) and a resilient rural one (Monts Dore). Not only do these components connect with the Crescent area in the North, they also cluster with Eastern Vivaro-Alpin (in light green). Vellave, which is usually considered a core-dialect of the eponymous Auvergnat dialect actually clusters in the next dendreme, with two subdialects of Limousin. As a result, Limousin makes up the core of North Occitan, while the Auvergnat component turns out to have one core, around PdD and CF, while all other components are peripheral mixed DS constituents: Crescent in the North, Vivaro-Alpine in the East–to some extent, these are heterogeneous "default dialects". Indeed, the fragmentation of Northern Occitan is bewildering, while the near-uniformity of the Southern Occitan macrodialect appears in sharp contrast, opposing Gascon (SW) to a huge area encompassing Languedocian and Provençal, further splitting into Alpine Provençal (in pink).

Unifying trends now compete in the South with dividing trends in the North, although clear patterns emerge in this turbulence area: buffer zones in peripheral areas (Crescent, Vivarais), robust core-areas in the Limousin dialect (in deep purple and light orange in the upper part of the map, on the right), among which the Auvergnat Vellave subdialect in the East (in deep blue on the map) as opposed to a Central Western zone, and what can be considered a very active diversification hub in Northern Auvergnat proper, around CF and the Eastern Monts Dore (EMD), in light purple and deep green on the map). These dynamics recall H. Simon's "near-decomposable components", questioning eponym dialects such as Limousin and Auvergnat, which were already viewed as an intricate complex since Ronjat's ODC–he conflated both into his D cluster under the label "Auvergnat-Limousin", opposing the former, subdivided into three subdialects (Northern Auv., Southern Vellave Auv., Southern Auv.), contrasting with no less than 11 varieties scattered from La Marche to Périgord, pertaining to the latter. Even if Ronjat did his best to enumerate discrete DS traits, the overall picture was still something of a mess. In fact, the whole region of the Northern macro-dialect is the realm of a huge mountain complex of volcanic origin, known in France as the Massif Central and linguistic diversity echoes the ecological complexity of the landscape). The extremely intricated orographic structure of this natural region, and the fact that this mountain also works as a powerful and dense hydrographic basin from where many major rivers flow down to the plains of Northern and Southern France, make the Massif Central a tension multiplier, prone to fostering hubs and small-worlds, or chains of both entities, from a DS and CADS standpoint. The pair CF/EMD could be considered a typical hub, while Velay makes up a small-world of its own,

15 662 places (villages, hamlets, small towns), 44,748 instances, 245,029 characters (among which unique: 99), 240,562 tokens (203 unique).

16 The Croissant mixed dialect would be the seventh Occitan dialect, although P. Bec was skeptical about whether it could be classified as an Occitan dialect, and he included it reluctantly on some of his maps, as in the right-hand map in Figure 1. Ward's Method with 8 intervals (WM-8) in Figure 8 does not select Croissant as a distinct area, although Complete Link and other more reductionist algorithms do (see below, Figures 9–11).

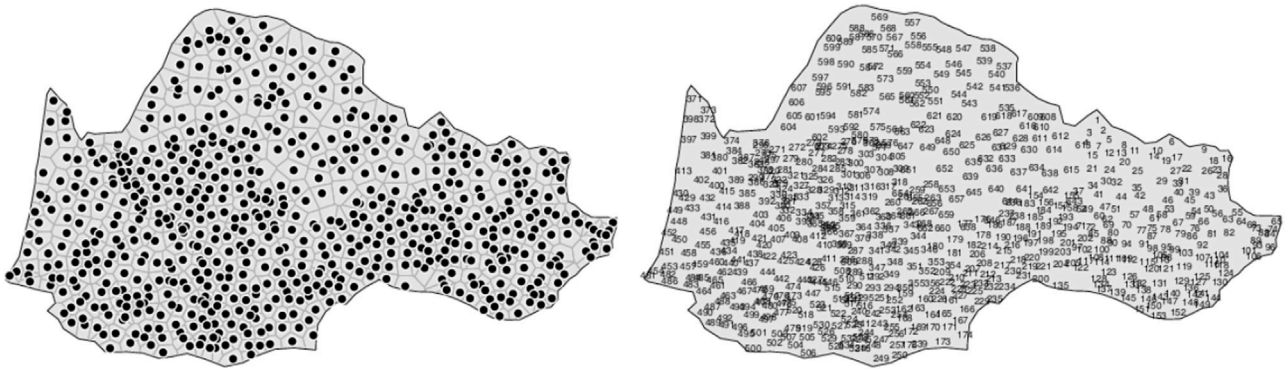


FIGURE 5
Tools (2): GIS maps from THESOC for data processing on Gabmap.

with connections in many directions within the Massif Central complex.[17] In many ways, the core of Highlands Auvergnat connects with Piedmont Limousin, in terms of orography. On the one hand, the four central entities of Central Limousin, Eastern Limousin (Velay), PdD and EMD make up the core of Northern Occitan, whereas the Croissant stands on its own as a buffer zone with Oïl dialects to the North. On the other hand, the Southern Macrodialect makes up a huge area, reaching from Gascon in the SW as a clear-cut dialect to a massive block combining Languedocian and Southern Provençal, with Alpine Provençal attached as an outlier within this major central clade in Figure 9 (on the right).

## 3.5 Group average (GA)

Group Average in Figure 10 tends to behave as a powerful centrifugal force as it searches for central tendencies within groups, preserving discrete clusters, yet including sets in a broader manner than WM (i.e., hierarchical cluster analysis)– it performs very well with enclaves and outliers in a dialect continuum, namely, with (*near-*)*singletons*. Here, with GA, both trends already observed with Complete Linkage are confirmed: intense heterogeneity of the Northern macrodialect vs. homogeneity of the Southern one. Yet, GA now conspicuously enhances the contrast between unifying trends in the Southern macro-area vs. diversification in the North: the Gascon vs. Languedocian and Provençal divide has now melted, and only a few splinters show up in the South, in the upper part of the map, to the left. Interestingly enough, two splinters in yellow correspond to enclaves of allogenous varieties (from the Oïl domain, i.e., Poitevin-Saintongeais),[18] which were not detected using the previous algorithms. Moreover, the Easternmost corner of the Niçard region (ALP 76: Brigue; 86: Saorge) pops out in the Easternmost corner of the gigantic Southern macrodialect, but is now considered as a Gallo-Italic Ligurian exclave (and a *singleton*, as such).

As to the intricated area in the North, the inner structure is now somehow quite different, and tells us more about hidden hierarchies and near-decomposable components for this segment of the DS. Now, Auvergnat is reduced to the Northern Highland hub with its two subcomponents (PdD vs. EMD). The Crescent now stands as an outlier of the Limousin cluster (in light green), instead of being associated "par défaut" to Auvergnat and, in a more abstract way, to Vivarais, as in the previous model provided by the CL algorithm. The Limousin dialect now patterns smoothly, splitting into two subdialects: Southern (in pink) vs. Central (in orange). It is no longer connected to Vellave, which has now become the most external outlier of the Southern macro-

17 Velay in SE Auvergne provides a unique opportunity to enhance the relevance of Simon's concept of *near-decomposable* entities in complex systems: on the one hand, the Vellave sub-dialect does indeed emerge as an entity on its own from our reductionaist algorithms; on the other hand, it can also be considered as a buffer zone, from Nauton (1974) and Guiter (1980) vicarious approaches–the former qualitative, the latter quantitative. This *Janus* effect of ambivalent properties can be explained by the fractal dimension of any geolinguistic entity. Nevertheless, it does not confirm the Paris & Meyer's *Continuist Hypothesis*. On the contrary, it challenges its methods, indicating the inability of the Paris & Meyer's Null Hypothesis to fathom taxonomic depth and intricacy.

18 The population from Saintonge settled in NE Gironde in the 15th Century after this area had been devastated by the plague. However, the original Oïl dialect has been under strong pressure from Gascon ever since. By the late 19th Century, it was already no longer spoken (although the *Atlas Linguistique de France* (ALF) by Gilliéron & Edmont, 1902–10 contains valuable data from the former Oïl dialect of the Montségur enclave, see http://lig-tdcge.imag.fr/cartodialect5/#/, ALF 635, Gironde). Instead, the variety recorded in the ALG (*Atlas Linguistique de Gascogne*, which is part of our THESOC database) had already become more of the Gascon type when elicited by Jean Seguy's team, though preserving some features of the Poitevin-Saintongeais substrate, making this Gascon variety a *singleton* on its own. See more data in Jagueneau (2014) as to the former Saintongeais variety spoken in this area.

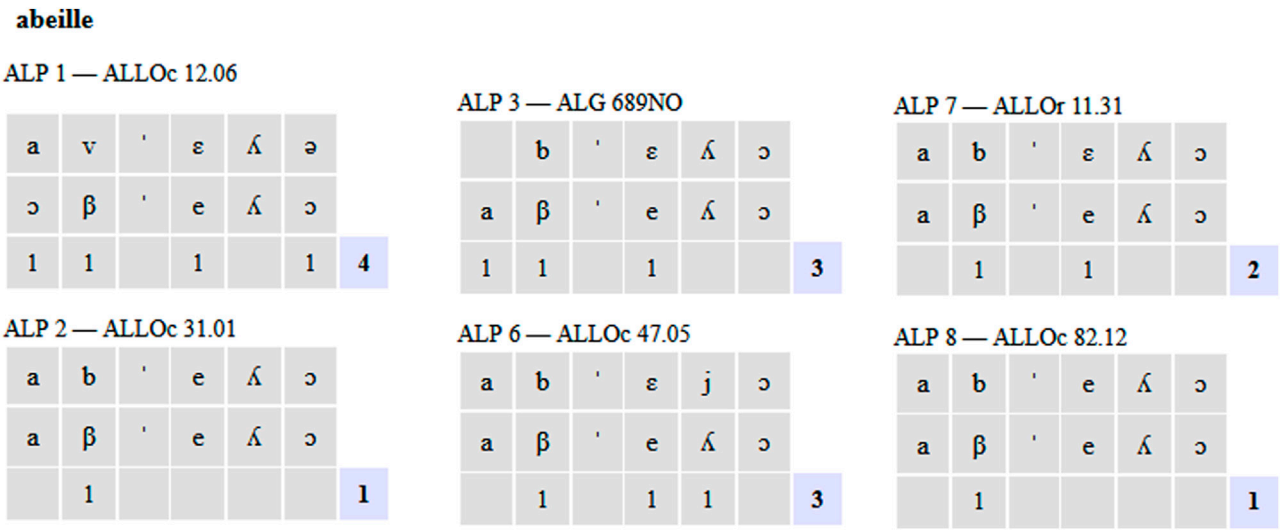



FIGURE 6
Tools (3): alignments on Gabmap for data processing according to edit distance (string tokens): *abelha* 'bee'.

area, next to Vivarais, in a "default" manner. These are interesting vicarious results obtained by GA as a reductionist algorithm.

Our definition of "Hubs" here differs from Duncan Watts's claim that "A separate development in the recent literature on networks has been the growing realization that in many real-world networks, the distribution of the number of network neighbors—the degree distribution—is typically right-skewed with a "heavy tail", meaning that a majority of nodes have less-than-average degree and that a small fraction of hubs are many times better connected than average." (Watts, 2004: 250). Gephi networks as in Section 3.2.2 below (see Figure 14) show that ALAL 10-11 do not belong to the most densely connected varieties within the cloud of ALAL localities. Nevertheless, Complete Linkage and Group Average (10 intervals) clearly show that these two varieties generate discrete variations of their own, embedded within the Auvergnat-Limousin network (see Figures 9, 10). We therefore consider them as entropic *hubs* or matrices of intense local variation. We therefore define *hubs* here by their *agentiveness* rather than by their *connectedness*. If a *hub* as defined by Watts is to be seen in the Auvergnat-Limousin network, such an entity does indeed show up at Gephi threshold 82.25, located in SW Limousin, corresponding to the pink area in the map, Figure 10.[19] Nevertheless, we rather consider it a "Small World" (Watts, 1999) in a broader sense, similar to the light blue area of what could be called a kind of extended Vellave major hub (in Auvergne) in the Central South part of the Northern macrodialect. As to *small worlds* from a qualitative standpoint in dialectology, see Léonard (1991, 2020b), Léonard and Dell'Aquila (2012).

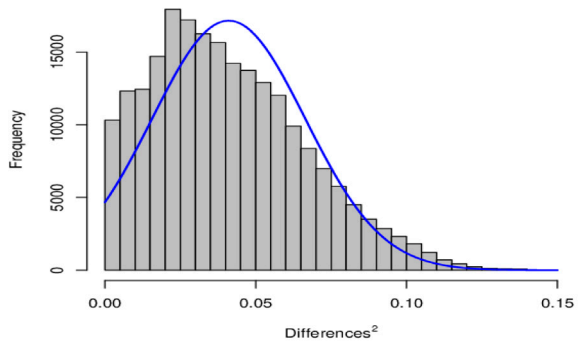
## 3.6 Weighted average (WA)

Weighted Average, in Figure 11, provides results at the crossroads between CL and GA: it enhances wider subgroups, yet with fewer consequences on the overall inner structure of the taxonomy, as we will see next. The Southern macrodialect turns out to be divided into two large areas: SW, with Gascon and Western Languedocian on the one hand vs. a wide complex embracing all the Massif Central areas to the North–including Southern Auvergnat, Eastern Languedocian and the bulk of Provençal, with the exception of Vivaro-Alpin in the NE part, which presents as an outlier of Southern Occitan.

As far as Northern Occitan is concerned, the overall inner structure is similar to the previous one (with GA in Figure 10 above), yet with a crucial difference: the Vellave subdialect in Eastern Auvergnat blurs and conflates into a wide Central-Northern area (in light purple), embracing all the Southern part of the Massif Central, as a sister constituent of the clade containing the two main SE components: the Provençal dialect and the Eastern Languedocian subdialect (both in orange). Interestingly enough, conflated Gascon and Western Languedocian in the Center-West (in light green) matches Bec's Aquitano-Pyrenean block analyzed above in Section 2.

The application of Complete Linkage (Figure 9) and especially Group Average (Figure 10) to our data resulted in flat hierarchies for the Southern macro-component of the DS, while Weighted Average (Figure 11) provides now far more balanced middle-sized DS units in the South. However, Northern groups preserve their patterns of strong diversity,

19 Namely, this Wattsian hub cluster ALAL29-30, ALAL 41–2, 44–6 in the Corrèze department, and have ALAL 43 (Queyssac) in Pays de Bergerac, Southern Périgord as a link to NW Languedocian localities, in the Western part of the Guyenne sub-dialect, in Gironde.



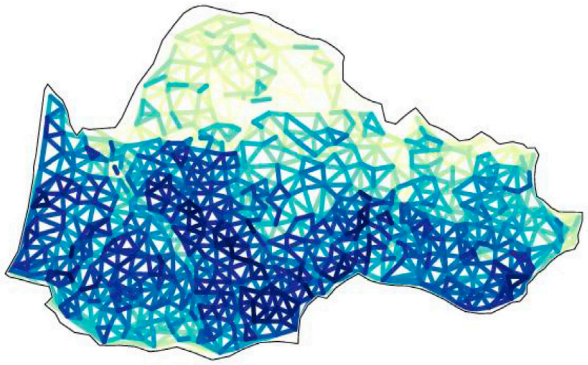




FIGURE 7
Structure of the THESOC phonological database[12].

with an intricate array of buffer zones (Crescent, hubs, dialects and subdialects blending into one another, etc.

## 3.7 A glimpse at external factors

The external factors explaining the diversity of structures through these models are many, and none should be too hastily enforced, as much research remains to be carried out, but we can already mention a few, which can contribute to unravel patterns of adaptive evolution between DS subsets and therefore, between dialectical entities of the Occitan dialect continuum in space and time:

(3) External Factors to take into account

(a) Patterns of Romanization of Southern Gaul: two bridgeheads (one in the SE corner of Bec's Aquitano-Pyrenean area, around Narbonne, the other in Provence, especially around Arles, at the intersection of roads such as Via Aurelia (to the East) and Via Domitia (to the West) – Leroy-Ladurie 2005: 274–283). Latinization proceeded from these strategic regions to the rest of the territory–through the Rhône Northwards in the East, through the wide plains of the SW from Narbonne to Aquitania and the hinterland of Languedoc toward the Massif Central piedmont of the Causses complex to the North.

(b) Late superposition of French as a dominant language, as convincingly demonstrated by Auguste Brun in his pioneering essay (Brun, 1923) inspired by the master Ferdinand Brunot on the spreading of written and spoken French in the Oc area (see Courouau, 2009 for an up-to-date survey of research on this topic since Brun's landmark essay). Bec (1970 : 401) points out that in the Middle Ages, the Langue d'oc had been a "major language of civilization", i.e., a prestigious literary language, with the Troubadour tradition, and a vehicular language, autonomous from Latin, in administrative settings, until the 15th-17th Century.[20] A revival started in the 19th Century under the leadership of Nobel Prize in literature laureate Frédéric Mistral and his movement, the Felibrige, and in the 20th Century with a renewed orthography rooted in Medieval Occitan and, to some extent, diasystemic insights, through the Institut d'Etudes Occitanes (IEO) –Bec (1970: 402). The langue d'oc thus formed an autonomous area to the south of Gallo-Romance for around a millennium, in a sociolinguistic context where the dominant language was Latin rather than French and its various forms from the Middle Ages to the Renaissance. This *de facto* sociolinguistic autonomy, despite the geopolitical effects of the so-called Crusade "des Albigeois" (1,209–1,229), led therefore to the emergence, through self-organization, of a major diasystem structured around large dialectal areas of relatively symmetrical proportions, between the regional powers of the southern part of the Kingdom of France. In some regions, such as Gascony in the SW, geopolitic autonomy was even greater than elsewhere in the South of France, as foreign powers had long prevailed (cf. the English period in historic Gascony, from 1,152 to 1,453).

20 Nevertheless, more recent research inspired by Jean-Pierre Chambon's reappraisal (Chambon, 2004; Chambon and Olivier, 2000), has shown that in many cases, the Langue d'oc alternated with Latin especially for short scripts in legal texts, see a major survey of Auvergnat written corpora since the Middle Ages in Velay, by Vincent Surrel (2022). Variation in monolingual or bilingual (Latin/Oc) was especially conspicuous until the 15th Century, but French was undoubtedly a late agent of linguistic contact, even in Northern Occitan, which happened to be located closer to centers of dissemination of spoken and written French. On the evolution of Occitan *scriptae* and literary traditions in Langue d'Oc through space and time, see Martel (2003) and Bernsen (2003).

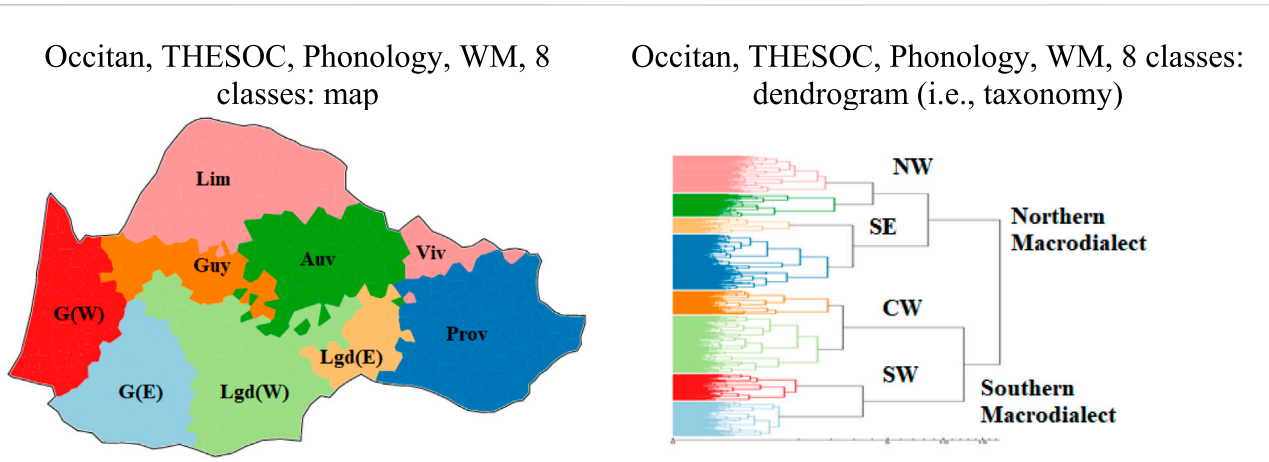



FIGURE 8
Hierarchical Cluster Analysis, 8 intervals, THESOC db (71 items), Phonology.

(c) Hydrographic and orographic factors: the geography of the Occitan domain can be described as a huge funnel, head downwards in the East, encapsulating the Rhône Basin, leading from Avignon to Lyon, with the Massif Central complex in the West, and the Alps on the Eastern side, on the one hand, and the Garonne/Dordogne great basin on the West, sealed by the Pyrenean chain in the South. The former configuration accounts for Bec's Arverno-Mediterranean zone, while the latter matches his Aquitano-Pyrenean macro-area. Charles Camproux in his essay on the linguistic geography of the Gévaudan Region (NE Languedoc, close to SE Auvergnat) is eager to explain variation in the Occitan highlands out of what he calls the "au fil de l'eau" pattern of settlement and condensation of dialectal norms (Camproux, 1962 : 762–3) – "Parce que le long de l'eau s'établissent et se succèdent les habitats humains".[21] Hydrographic basins happen to be extremely intricate in the central highlands (Massif central, i.e., "La montanha"), providing strong conditions for the spreading of and interaction between human aggregates (see Derruau-Boniol, 1970: 25–8, and 68–125 on the human geography of this highland complex).

(d) Ecological intercomplementarity: on the Western side, lowlands simplified interactions between populations, while the transversal chain of the Pyrenees fostered ecological and economic complementarity between pastoral and diversified agrarian economic models, benefiting from outlets to the Atlantic. On the Eastern side, a variegated pastoral-agrarian economy in the Massif Central and Alpine highlands crossed the Rhône as a logistics vector, in complementarity with the Mediterranean lowlands. These complex ecological settings fostered diversity of agrarian and economic models (wheat, wine production, livestock, etc.). As can be seen in Figure 12, the whole area where Occitan is spoken can be described as built up around (i) the Aquitano-Languedocian Isthmus in the West, with a major and capillary hydrographic basin covering a huge zone of plains and a Piedmont in its Eastern fringe, (ii) a massive twofold block with the Massif central to the West and the Western Alps to the East, separated by (iii) a wide corridor, dominated by the Rhone, which belongs to the category of major fluvial strands in Europe, like the Rhine and the Danube in Central and Eastern Europe (Lafont, 2003; Martel, 2019); (iv) last but not least, the third highland lattice in the South–the Pyrenees, where strong Medieval states emerged, such as the États du Béarn and Aragon, and powerful tropisms pulled this region towards the English world (Aquitania and Guyenne in the SW) or the Spanish and Catalan sphere, in the South-Eastern part.

(e) As a consequence of (i-iii) in the later point above (point d) and of the main civilizational trends in the socioeconomic framing of the whole area (Leroy-Ladurie, 2005: 270–85), intense and wide circulation of people and goods, through river transportation in the wide hydrographic basins, herding with "transhumance" i.e., seasonal migration of herds in Central and Alpine highlands, a dense network of small and middle-sized urban geopolitical centers (Pau, Agen, Toulouse, Narbonne, Carcassonne, Albi, Foix, Rhodez, Conques, Arles, Nîmes, Mende, Avignon, etc.) and their markets and arrière-pays and their location in relation to the pilgrim trail to Santiago de Compostela, etc. have contributed to trends of unification.

(f) Political divisions (Comtés de Provence, de Toulouse, "États du Bearn" in the SW, etc.). From this standpoint, a whole geopolitical history of the French Midi still awaits chronicling and a synthesis (see Leroy-Ladurie, 1977; Leroy-Ladurie 2005:

21 "Because human settlements follow one another along the water's edge." To the author, this rule sounds familiar, and mountains are not needed to find an application–Siberia, although hugely flat, shows the same patterns, with most of the linguistic and ethnic diversity spreading along hydrographic basins, to the point that linguistic subfamilies can be labeled according to the river they cling to, such as *Ob*-ugrian peoples (Uralic Hanti and Mansi), or in the Altaic linguistic stock, the so-called Ienissei Kirghiz (Khyagas or Khakas), etc.

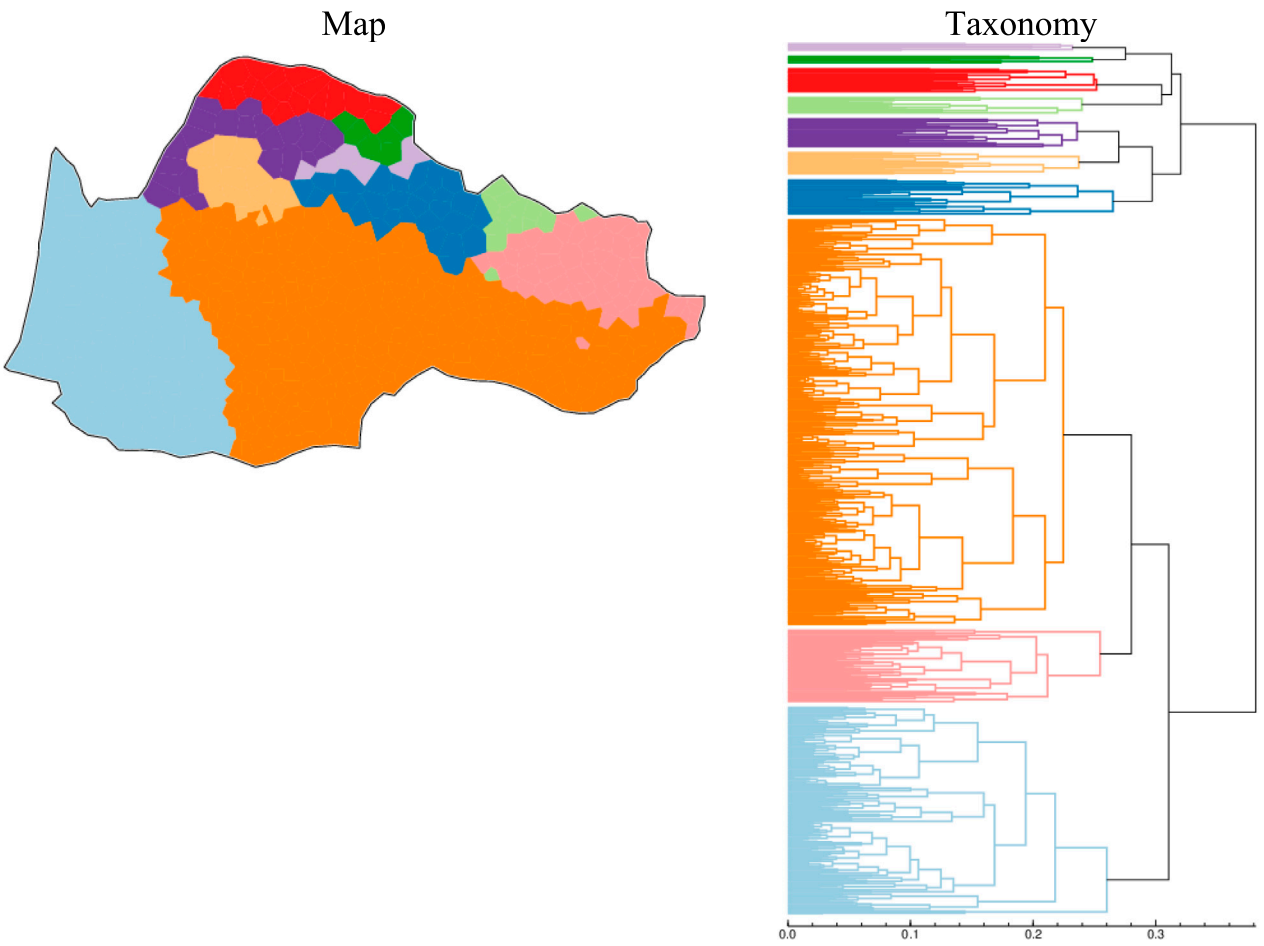



FIGURE 9
Complete Link, 10 intervals, THESOC data, Phonology, CL, 10 classes (71 items).

285–312) – among many other authors, see Wolff (1967: 121–353) for the central region, Languedoc and Lafont (2003: 85–159) for a broader view, within a European framework.

We can now sketch a model of articulation between our vicariant DS patterns in ODC and external factors, which belong to a higher degree in the description of the complexity of the Occitan dialect network.

## 3.8 Modeling near-decomposable components of the Occitan diasystem

As already suggested, any diasystem can be broken down into many parts, confirming–if necessary–the relevance of dialects. Beyond ruling out Meyer and Paris' Null Hypothesis (see section 1 above), our survey of vicarious qualitative (Ronjat, Bec) and quantitative (Gabmap processing) patterns highlights various ontological models, of which we will suggest a tentative overview in Table 1, based on physical and social coordinates (geography and society), and on pattern analysis. According to this schema and as a consequence of the patterns observed in the survey of our data, what we generically call "dialects" can be more precisely defined, as in Table 3, as *macrodialect, dialect, subdialect, buffer zone, default dialect, hub, singleton*.

The notion of *macrodialect* has already been illustrated here by Bec's transversal areas (Northern vs. Southern; Arverno-Mediterranean vs. Aquitano-Mediterranean, see Figures 2A, B above): it is necessarily centrifugal (from center to periphery), but hardly endemic if not in the broader sense, and it appears as a major or main clade in taxonomic models such as dendrograms (as in Figures 9–11). It can hardly surface as an outlier, due to its mass, its centrality and generality.

A *dialect* proper is a smaller entity subsumed by the former category of macrodialect, embedded within its mass. As a subcomponent of a macrodialect, it shares the same centrifugal properties (i.e., the spreading trend) as its supercoordinate, but is still too big to make up a small world. It may or may not rank as a main component in the taxonomy (+/–)– for instance, Languedocian often surfaces as a main dialect within taxonomic trees, whereas Auvergnat and Vivaro-Alpin hardly ever do. A dialect is generally too conspicuous to surface as an outlier in taxonomic trees.

A *subdialect* shares the centrifugal property with the former two entities, and it may be endemic or not (i.e., made up of interfering or

fluctuating local or sporadic bundles of traits). It may well make up a small world–to some extent, the Vellave subdialect matches this property, embedded as it is at the very core of the Massif central, as a result of highly condensed relationship patterns within its territory and a measure of autarchy. It cannot be a main constituent in dendrograms, and it is generally paired within a given dialect, although the affiliation between Limousin and Auvergnat fluctuated in our vicarious results. It can therefore be an outlier or not.

A *buffer zone*, in contrast, is more endemic than centrifugal, as a mixed variety open to many exterior influences, it can hardly make up a small world–it would need more inner condensation of relationships and autonomy. It seldom, if ever, surfaces as a main clade and most often presents either as an outlier alone in its branch, or paired with a *default dialect*. This later type shares most of the same properties with a buffer zone, but as its status as such is more a by-product of taxonomic comparison, it may gather enough idiosyncrasy to make up a small world.

A *hub* shares many properties with higher ranking categories, such as dialect and subdialect: it is densely centrifugal, it has inner patterns of endemicity and its inner growth dynamic is that of a dense small world too. But as a very active or tense core of differentiation within the DS, it hardly presents as a main clade, and the intensity of its development requires an inner contrast, like the PdD/EMD, to grow, so that the property of small world remains optional. Last, but not least, a *singleton* has no expanding charge, it is endemic (in its own spot, which may be a micro-region or an exclave/enclave as in the case of the South-Easternmost region) and it can only surface in the dendrogram as a taxonomic outlier.

In this survey of dialect ontologies in the Occitan domain summarizing the main types of DS entities observed through vicarious models, we have gone from splitting (with WM) to lumping (with CL, GA and WA), fulfilling one of the basic prerequisites of CT to consider that a complex system is more than the sum of its parts. We considered several approaches and scales of hierarchy for the Occitan DS and we handled the issue of components of the system as being near-decomposable categories, despite being concrete ones, to the extent we could model a table of entities and holistic properties as in Table 3 above, from the standpoint of *internal* (taxonomy) and *external* factors (geography, society). Nevertheless, all the tools implemented have relied on hierarchical methods. In the next step, we will explore further holistic aspects of this DS as a complex system, from two standpoints: first, *multidimensionality*, second, *reticularity*.

## 3.9 From dialect continua to dialect networks

### 3.9.1 Multidimensional scaling

Figure 13 shows two images of MDS for the same data as used so far. Most of the salient patterns already observed, with a full house of hierarchical models, are condensed here into maps rendering the complexity of the data visible. The traditional map, on the left, shows how dialectal entities (*macrodialect, dialect, subdialect, buffer zone, default dialect, hub*, etc.) as in Table 3 surface as a palimpsest–and yet they remain traceable, once identified using the preceding protocols.

NB: *r* stands for the coefficient of geolinguistic correlation.

Some areas are very dense and unified, while others blend into each other, revealing all patterns at once. At first sight, on the left side of the map, Figure 13, we can recognize massive blocks and their seams, which have much to do with external factors listed in Section 3.7 above. To point out just a few, we see the central bulk of Auvergnat mingling in the North with Central Languedocian (in various degrees of purple), corresponding to the Southern tropism of Massif Central ("La Montanha"), in terms of geographic and historical trends of interrelations between various agrarian economy models. Eastwards, we discern Provençal, also mingling in the West with the Massif Central; the seam made up by the deep Rhône valley further to the SW, and the intricate overlapping of Vivaro-Alpine with both Eastern Northern Occitan and Provençal. Further to the NW, we now better visualize what could be the domain of Limousin, separate from Auvergnat, presenting as a compact mass (in green) and showing strong inner complexity, and including the Crescent on top, as a kind of Limousin roof. Last, but not least, the Aquitaine zone surfaces with its two subdialects (Western vs. Eastern Gascon) –the latter being more intricate than the former, out of the interaction with a chain of piedmont or highlands dialect *hubs* (Bearn, Bigorre, Comminges, Couserans). The former (Western Gascon, or Maritime Gascon) shows more unity, through the plains and former marshes of the Landes, intersected in its Northern part by the Girondin subdialect. Even splinters and enclaves can easily be identified: the Petite Gavacherie in the North-Eastern fringe (spots in light green), and other splinters from Auvergnat and Vivaro-Alpin on the Southern outskirts of the Massif Central.

Gabmap provides up to six dimensions, but we will focus on one only, for the sake of concision: the 2$^{nd}$ dimension (Figure 13, to the right) enhances centrality (the gap in yellow at the center of the map) vs. peripherality (in various nuances of blue and green), unveiling deep properties of the Occitan DS. The most divergent dialects are therefore Gascon (especially Western) and what Ronjat dubbed "Auvergnat-Limousin" (group D in his nomenclature). Ranking third in idiosyncrasy stands the Provençal dialect, especially the Alpine and the South-Easternmost singleton. Splinters and other singletons are easy to spot all the same: the Petite Gavacherie at the outskirts of Eastern Gironde and a strong Vivaro-Alpine one in the East.

From the start, we have implicitly preferred the term "dialect network" to its more traditional counterpart "dialect continuum ". The former conveys the dynamic notion of interaction, and fits well with a CT approach, while the latter, as we have seen, while not altogether inappropriate, has strong connotations linked to G. Paris' and F. de Saussure's Null hypothesis. More concretely, various tools currently used in CT (GraphStream, Gephi, etc.) help to concretely implement the network dimension of dialect connectivity. Some results will be presented in the next section.

### 3.9.2 Networks streams–reticularity

The THESOC phonological data were first processed with edit distance, in order to obtain a matrix of values which was then converted into similarities, using normalized thresholds, with Gephi (Figure 14). This method of visualization makes it possible to detect another range of dialect differentiation, as the higher the value, the more scattered the varieties on the graph (abbreviations correspond to the linguistic atlas processed: ALAL, ALG, ALLOc, etc., followed

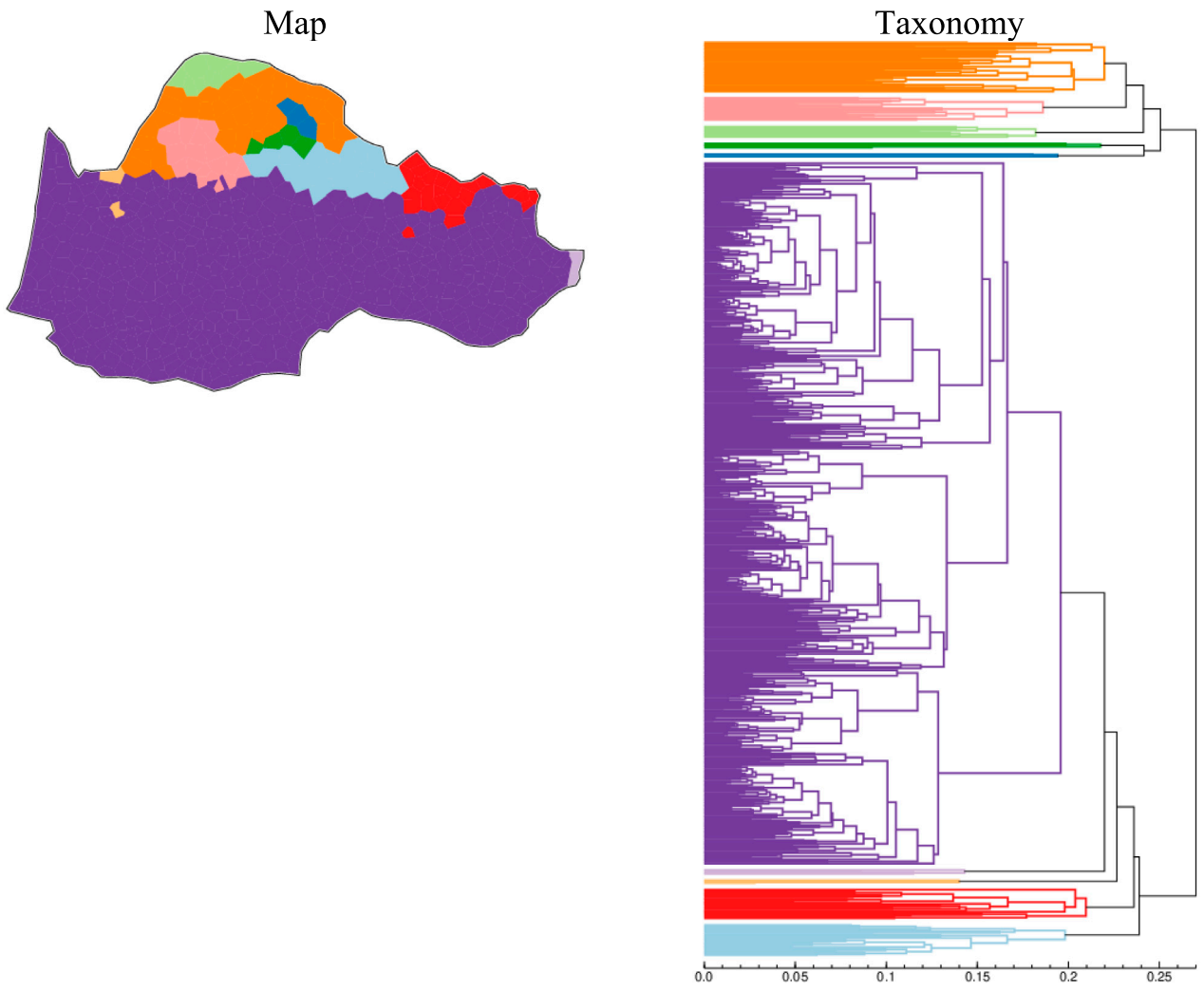



FIGURE 10
Group Average (GA), 10 intervals, THESOC data, Phonology, GA, 10 classes.

by the index of the locolect) in the graphs in Figure 14. At threshold 87.5, three dialects still show rather strong density, and can be considered to share to some extent the network property of compactness, i.e., consistency: Gascon (ALG), Provençal (ALP) and the two subdialects Western Languedocian (ALLOc) along with Eastern Languedocian (ALLOr) – all belonging to Bec's Southern macrodialect. The other two dialects belong to North Occitan, (Limousin: ALAL and Auvergnat: ALAL & ALMC). Nevertheless, a hierarchy of compactness can be determined: the more compact dialect being Gascon, while the less consistent or more scattered of this cohort is undoubtedly Provençal, suggesting a pattern ALG >> ALLOr >> ALLOc >> ALP. In contrast, the two Northern dialects, as presented in the corresponding atlases, have low network compactness, presenting the declivity ALMC >> ALAL (i.e., roughly Limousin >> Auvergnat). The graph at threshold 92.5 enhances the robustness of these patterns, making core entities easier to grasp.

Compactness is by no means the sole property made visible by these graphs: vertices between nodes also indicate interconnectivity of the varieties within the DS. As we can generate many graphs of this kind, these models may help us to detect more specific properties in the lower levels of our model in Table 3, especially *buffer zones* (connected to a diversity of different dialects), *hubs* (densely connected in local chunks), *default dialects* (sparsely or randomly connected), *singletons* (unconnected to the rest of the network). Of course, the evaluation of these properties cannot rely on cherry picking in a linear fashion, but should be performed step by step comparing various levels of aggregation at different thresholds, and comparing to hierarchical patterns and MDS results, as in the previous section.

## 3.10 Data mining applying our ontological model

Major isoglosses, such as T1 & T5 (AqP/CO) are handy: T1 (AqP/CO) for instance stipulates those velar stops do not undergo

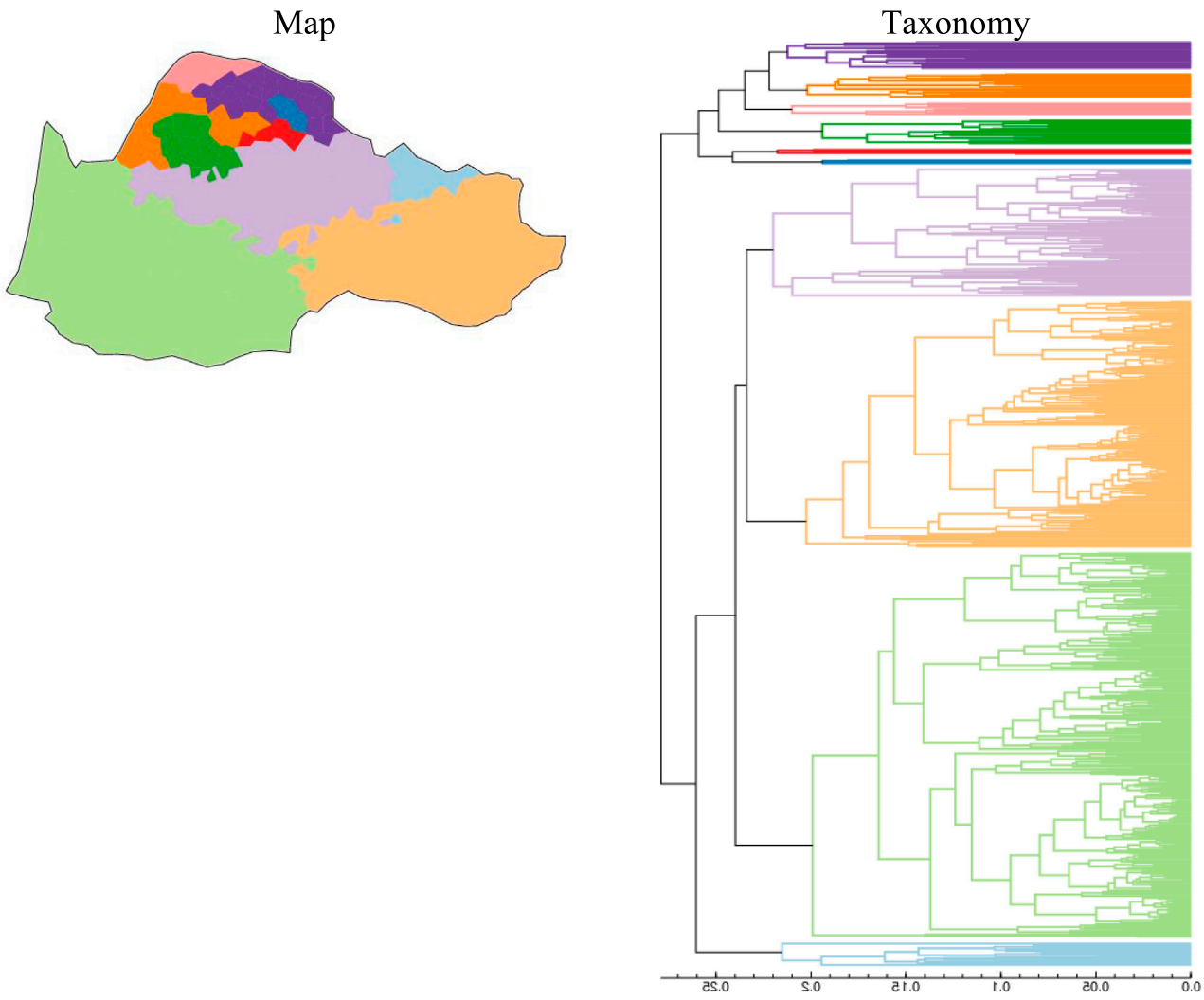



FIGURE 11
Occitan, THESOC, Phonology, 71 items, WA, 10 classes.

palatalization in the Aquitano-Pyrenean domain, as one can see in Figure 3 above. This type is so powerful that it even accounts for the North vs. South division of the entire Occitan domain, as in Figure 2 (left map), according to Pierre Bec. Bold characters show relevant reflexes, in terms of the structuration of spatial entities, such as those revealed by Group Average (10 intervals, Figure 10).

Nevertheless, the combinatorics of segments induce many divergent patterns, both in the diasystem itself and in the local grammars. For instance, a velar stop combined with a nasal sonorant T1 (AqP/CO), Old High German **agaza* 'magpie', widely disseminated in the Occitan and the Poitevin-Saintongeais domains, confirms this division, as the intervocalic voiced velar *-g-* is preserved in Southern Occitan, including in Central Vivaro-Alpine ALP 16, while the core of the AqP/CO zone has the so-called "Iberian" *lenis* reflex *-ɣ-* in Toulouse Languedocian (ALLOc 31.12) and Gascon ALG 699SE, including in the Gavache enclave ALG 635. In contrast, in the Northern part of the domain, palatalization emerges as much more than palatalization proper. The array of reflexes goes from genuine palatal reflexes (such as *ʒ* in the Crescent ALAL 66–68) to depalatalized tokens, such as *dz* in Central Limousin (ALAL 25–26 & 48) and in the N Auvergnat hub (ALAL 10–11), to even more obviously antipalatal reflexes, such as the very interesting case of the coronal voiced laminal *ð* in CW Limousin (ALAL 48). However, two varieties behave aberrantly in this respect: NW Rhodanian Provençal (ALP 3) developed a palatal approximant *j*, while a Central Limousin variety (ALAL 27) has the "Iberian" *-ɣ-* reflex. This points to a different scenario from the one entailed by classical dialectology: the endemic cohort of reflexes such as *j, ð, ɣ*, of the lenis or approximant type as opposed to the either preserved velar stop vs. palatalized stop–both of the *fortis* type.

Nevertheless, any major variable, such as velar onsets before low vowels, like T1 (AqP/CO), is merely the tip of the iceberg of the much complex web of combinatory interactions between sounds in the phonological component of a language. These second range traits will be listed here as Nx (*Nexi*), as they involve the same phonological material (as velar stops), embedded within different patterns, resulting in very different outcomes: Latin velar stops can undergo general palatalization, as in Nx.1 in AGNELLUM and Nx.2 AGNELLARE, or they can split their reflexes in two, as in Nx.3 resulting in Latin

TABLE 4 Data mining of specific diasystemic spots for T1 & T5 (AqP/CO) and CT (Contextual Traits), monitored by Group Average (GA 10 classes).

| Phonological variable | CA/GA | Pal_GN-, -ELLU# | T & N > V | Pal_CU/GU | Betacism |
|---|---|---|---|---|---|
| **Type** | T1 (AqP/CO) | Nx.1 | Nx.2 | Nx.3 | T5 (AqP/CO) |
| **Etyma** | Old HGerm **agaza* | AGNELLUM | AGNELLARE | ACUCULAM | APICULAM |
| **GA_10_clas** | 'magpie' | 'lamb' | 'lamb a flock of sheep' | 'needle' | 'bee' |
| **Crescent Lim** | | | | | |
| **ALAL 66** | ***ʒ***ˈ*as*$^{ə}$ | *aɲ*ˈ*e* | $^{a}$*ɲel*ˈ*a* | *g*ˈ*yj*$^{ə}$ | *b*ˈ*œj* |
| **ALAL 67** | *a****ʒ***ˈ*as* | *aɲqˈɔ* | | *aɟ*ˈ*yj*$^{ə}$ | *abej*ˈ*e* |
| **ALAL 68** | *a****ʒ***ˈ*as* | *aɲ*ˈ*ɔ* | | *ag*ˈ*yj* | *ab*ˈ*œj* |
| **Auv-Lim (ALAL)** | | | | | |
| **10_N.Auv_Hub_Mt.Dore** | ***dz***ˈ*asɔ* | *aɲ*ˈ***i*** | $^{a}$*ɲi****l***ˈ*a* | *g*ˈ*yʎa* | *b*ˈ*ɛʎa* |
| **11_N.Auv_Hub_PdD** | ***dz***ˈ*asɔ* | ɑɲˈ*e* | *nel*ˈ*ɒ* | ***i****gɥ*ˈ***i****jɔ* | *abij*ˈ*i* |
| **ALAL 48_CW Lim** | ***ð***ˈ*aʃɔ* | | *aɲɛl*ˈ*a* | *gyʀ*ˈ*ɔ* | *bɛʀ*ˈ*ɔ* |
| **ALAL 25_CN. Lim** | ***dz***ˈ*asɔ* | *aɲ*ˈ*a*$^{ə}$ | *ɲel*ˈ*a* | *ag*ˈ*yʎɔ* | *b*ˈ*œʎɒ* |
| **ALAL 26_CLim** | ***dz***ˈ*asa* | *aɲ*ˈ*e* | | *aɟ*ˈ*øja* | |
| **ALAL 27_CLim** | *aɣ*ˈ*asə* | *iɲ*ˈ*ɛ* | ***i****ɲə*ˈ*e* | *aɣ*ˈ*yʎə* | *aβ*ˈ*øʎə* |
| **Vivarais & Provence** | | | | | |
| **ALP 16_Cvi-A** | *ag*ˈ*asœ* | *anj*ˈ*ɛw* | *aɲel*ˈ*a* | *agɥ*ˈ***i****jɔ* | *ab*ˈ*ejɔ* |
| **ALP 3_NW Prov_Rhod** | *a****ʃ***ˈ*asɔ* | *anj*ˈ*ɛ* | | *eg*ˈ***i****ʎɔ* | *b*ˈ*ɛʎɔ* |
| **Western Languedocian** | | | | | |
| **ALLOc 33.10** | *ag*ˈ*asœ* | *aɲ*ˈ*ɛw* | *aɲel*ˈ*a* | *ag*ˈ*yʎə* | *aβ*ˈ*eʎœ* |
| **ALLOc 31.12 (Tls)** | *aɣ*ˈ*asɔ* | *aɲ*ˈ*ɛl* | *aɲel*ˈ*a* | *aɣ*ˈ*yjɔ* | *aβ*ˈ*eʎɔ* |
| **ALLOc 11.02** | *ag*ˈ*asɔ* | *aɲ*ˈ*ɛl* | *aɲel*ˈ*a* | *ag*ˈ*œʎɔ* | *aβ*ˈ*eʎɔ* |
| | | | | | |
| **ALG 699SE (CASAU)** | *aɣ*ˈ*asa* | *aɲ*ˈ*ɛt*$^{a}$ | *aɲer*ˈ*a* | *aɣ*ˈ*yʎə* | *aβ*ˈ*eʎa* |
| **ALG 635 Gav** | *aɣ*ˈ*asə* | ***i****ɲ*ˈ*ɛ* | ***i****ɲə*ˈ*e* | *aɣ*ˈ*yʎa* | *aβ*ˈ*øʎə* |

[a]Final *-t* here results of complementary distribution (with delateralization in final position vs. rhotacism of the lateral in intervocalic position, as in *aɲer*ˈ*a* in the next cell of the table.

unvoiced velar intervocalic onset before a high back vowel in ACUCULAM, opposing resilient velar stops or approximants to palatalized stops of the *ʃ* type, as in ALAL 26 C Lim. or Crescent ALAL 67. Items matching the Occitan derivational pair noun <*agnel*> 'lamb' vs. denominal verb <*agnelar*> 'lamb a flock of sheep' are all the more interesting as they trigger reflexes embedded in word formation, highly dependent on context, e.g., in Gascon (ALG 699SE) Nx.1 and Nx.2 oppose final *-t* < -LL- in Pyrenean Gascon to consonant dropping in the Gavache enclave (ALG 635) and rhotic vs. lateral reflexes in each variety.

The Latin nominal diminutive suffix -ELLU is a nexus endowed with strong variational properties, as it displays a variegated list of secondary diphthongs (ALAL 25 *aɲ*ˈɑ$^{ə}$) and contrasting stem vowels (ALAL 66 *aɲ*ˈ***e***, ALAL 10 *aɲ*ˈ***i***, ALAL 48 *aɲ*ˈ***ø***; apophonic ALAL 27 *iɲ* ˈ***ɛ***, etc. vs. Languedocian *aɲ*ˈ***ɛl*** with preservation of the final lateral). Such nexi tightly bound to affixal combinatoric can be called "phonolexical", as they induce specific sets of fusional reflexes. Unlike regular "neogrammarian-like" traits unfolding in huge areas to make up macrodialects, they tend to feed hubs and singletons in a geolinguistic space. Interestingly enough, the Latin unvoiced velar stops in $AC_1UC_2$(U)LA involved in Nx.3 behave very differently from their counterparts before low vowel in T1 (AqP/CO): $C_1$ undergoes palatalization only sporadically and superficially in the North (ALAL 67 *aɟ*ˈ*yj*$^{ə}$), whereas $C_2$ makes up an organic nexus with the last syllable lateral, resulting in a palatal lateral after syncope of the second *u*: Late Lat. ACUC'LA > ALAL 25 *ag*ˈ*yʎɔ*. ALAL 11 ***i***gɥˈ***ij***ɔ shows apophonic interaction of the stem vowel with the initial one, while ALAL 66 gˈ*yj*$^{ə}$ combines apheresis (initial vowel dropping) and delateralization of the -C'L-nexus. Posttonic vowel reduction tends to split in two: vowel raising, as in Languedocian or schwa reflexes in Crescent and Central Limousin (ALAL 27) and Pyrenean Gascon, but low vowel preservation in Gavache, Central Limousin (ALAL 26) and the Mont Dore hub.

This diversity of local grammars strongly advocates for a categorical bipartition between Typological Traits proper ($T_n$), as in (1) above in Section 2.2.2 vs. Nexi ($Nx_n$) or Phonolexical

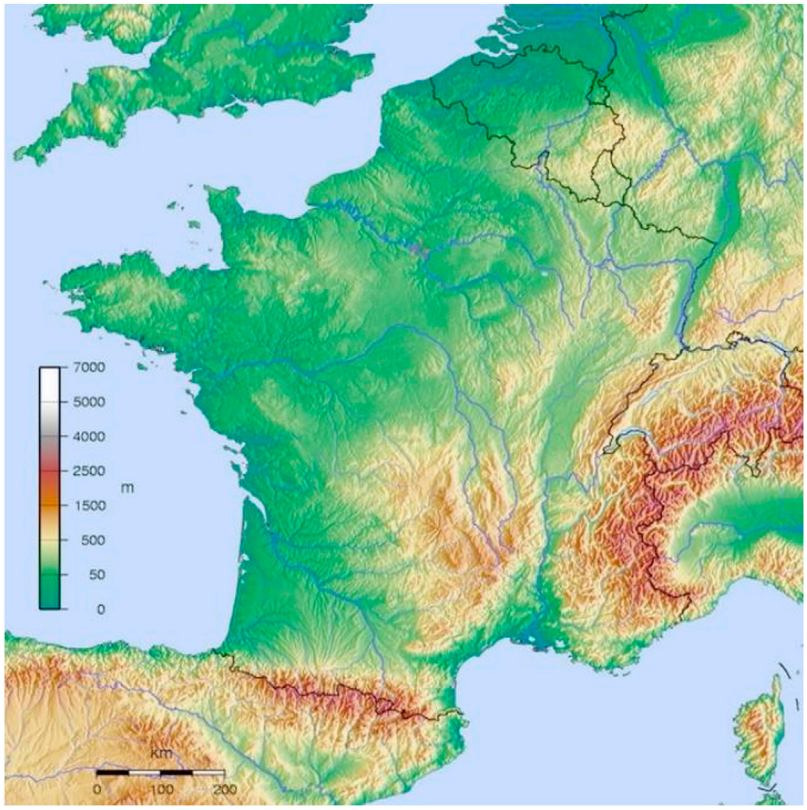



FIGURE 12
Distribution of some endemic reflexes (velar and coronal secondary approximants).

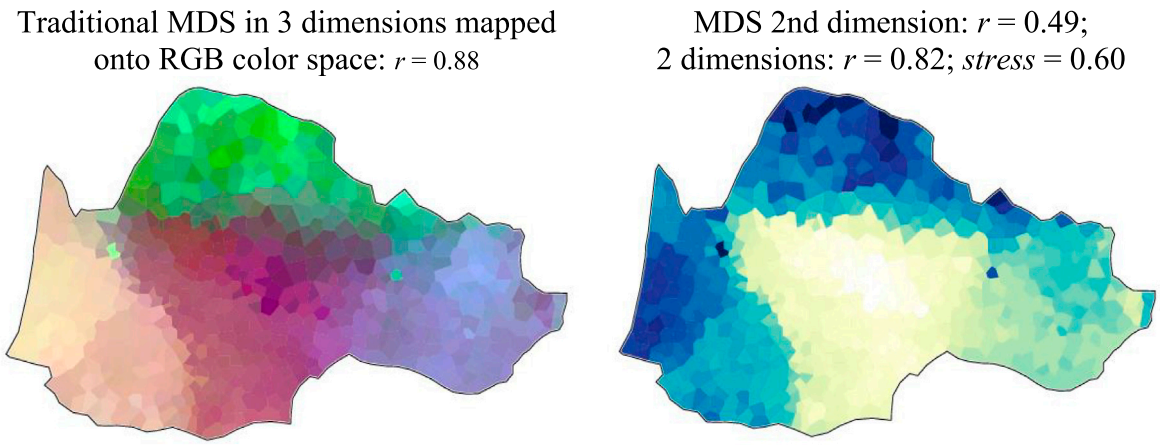



FIGURE 13
Multidimensional Scaling, THESOC database, phonology.

variables, as in Table 4 above, in this section. Last, but not least, T5 (AqP/CO) stands for *betacism*–another so-called "Iberian feature" inside the Occitan network, according to Pierre Bec. Although the *b* reflex shows up in Languedocian and Gascon (including in the Gavache splinter), it also appears sporadically in Northern Occitan, as in ALAL 27 (Clim) *aβ'øʎə*. Strengthening of the labial stop through apheresis of the low vowel occurs densely in the North (see ALAL 66, 10, 48, 25 and even in ALP 3 in the East).

These distribution patterns do not invalidate entities such as Bec's macro-areas or macrodialects and Ronjat's dialects and subdialects. On the contrary, they pay tribute to the agentiveness of the emerging entities (such as *buffer zones, small worlds, hubs* and *singletons*, listed in Table 3 above) within the diasystem, according to

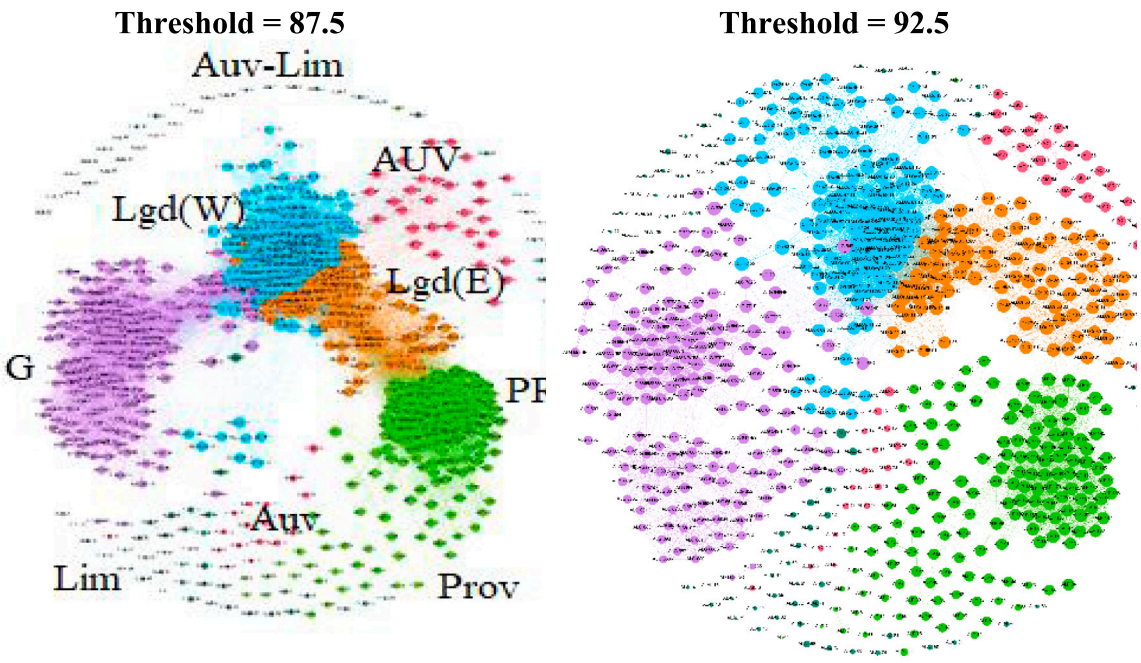



FIGURE 14
Multidimensional Scaling, THESOC database, Gephi graph, phonology (71 items)[22]. Abbreviations for dialects as in map Figure 8 above (with different colors).

ecological patterns of individuation and interactions in their local context. The gradation of the distribution of the velar approximant on the left of Figure 15 matches the two main subdivisions of the Gascon dialect (Western vs. Eastern) and even enhances the buffer zone of the NW Girondin area (to the West of Bordeaux) and the Gavache singleton. In other words, this pattern looks like a fractal of the main divisions of Gascon.

Other two independent buffer zones (with Guyennais on the one hand and Western Languedocian on the other) appear on the NW and CE fringes of the Gascon block, and a Southern Auvergnat hub can also be grasped right in the middle of the map, represented by green spots embedded in yellow. On the right of Figure 15, the overall pattern still agrees with the AqP/CO *macrodialect*, while enhancing Gascon. But this time, *hubs* and *buffer zones* predominate: two in the Gascon dialect (Girondin-Guyennais in the NW vs. SW Languedocian in the Eastern fringe of Gascon) and one in the Eastern part of Velay (Auvergnat). A *singleton* spot shows up in SC Limousin.

## 4 Discussion and prospects

In Léonard (2022), we advocate a multifold model for examining dialect variation in space and time, on the basis of Georgian data (Kartvelian, Great Caucasus), using edit distance to test an array of variables hierarchized according to a model inspired by Gudschinsky's (1958) on Mazatec–an Otomanguean language spoken in South-Eastern Mexico. Gudschinsky proposed a model articulated on ten fundamental notions from General Dialectology (items with index* point at similar categories as in our model, Table 3 above): *The Dialect Split Layer* (matching our macrodialect* layer here), *The Buffer Zone* Effect*; *Variable Bleeding* (in smallworlds*), i.e., *Relative Chronology* (Scalar Change); *The Feature Pool*[23] *Effect* (in hubs*); *Emerging Isolates* or singleton*, surfacing as dendrographic outliers; *The Center-Periphery Effect*, i.e., the Bartolian center/periphery mechanism; *Phonolexical Endemic Patterns* (endemicity*); *Word Geography*, i.e., lexical dissemination; *Local Semantic Shifts*; *External Factors* (as in Section 3.7 above). Although we did not apply Gudschinsky's stratified model here, it underlies much of our interpretation of results in the previous section, as suggested by our indexation. It takes a multidimensional approach to dialect variation, from a qualitative standpoint. However, the opposition between major Typological Traits ($T_n$) and Nexi ($Nx_n$) in block (1) in Section 3.10 above flows directly from this previous modeling, and accounts for much of the disarray provoked since the very beginnings of modern dialectology by the imbrication of isoglosses on geolinguistic maps. Any phenomenon may indeed be messy so long as one does not have an efficient and consistent, yet simple, hierarchy at hand. The *T* vs. *Nx* distinction may help here: some units of the sound system or the grammar

22 We thank Flore Picard, postdoctorate, Sorbonne University, for the implementation of the Gephi graphs in Figure 14.

23 See Mufwene (2001).

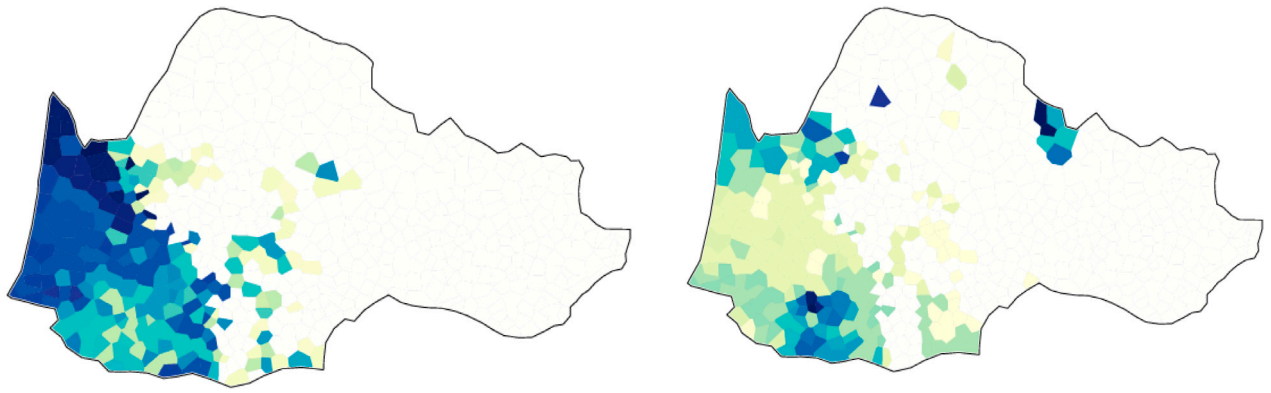



FIGURE 15
Tools (2): GIS maps from THESOC for data processing on Gabmap.

(or the lexicon, syntax, etc.) in a language resort more to the simplex type than the complex type, and this may make a big difference in how we use them to build up units of knowledge on natural or social phenomena (such as languages, dialects, etc.). Figure 16 shows how these two qualitative categories may contrast within a diasystem, in correlation with quantitative results, as surveyed previously.

Here again, we come across the same fractal (i.e., iterative and metonymic) property of phonological variables, for which cumulative processing may often release patterns very similar to the overall picture emerging from computing hierarchical outputs. In both maps provided by stochastic clustering, macrodialectal divisions clearly appear in the South, especially with T1 (velar stops before low vowel: CA/GA), patterning as a relatively flat hierarchy, except for the inner split of Gascon in the SW as opposed to Languedocian-Provençal in the center and in the South East. In contrast, the Northern Occitan macrodialect landscape is intensely variegated in both cases. The Crescent emerges as a roof organically rooted in Western Limousin–instead of a mere buffer zone with Oïl dialects, as usually depicted–, and Vellave Auvergnat, in the central left-hand part of this macro area unfolds widely with bridgeheads on both sides, westward (into Limousin) and eastward (into Vivaro-Alpin), surfacing as a major "Wattsian hub". Vivaro-Alpin however shows a much more scattered profile, especially for T1 and seems here–unlike for the Nx.1 feature–to qualify as a "default dialect". Strikingly enough, *singletons* neatly emerge, such as the Gavache enclave in NE Gascon, South Eastern Provençal (Nice), the Briançon spot in the NE Alpine area, as do occasional splinters close to Auvergnat and Vivaro-Alpine in Languedoc and Provence. All these details stress the metonymic property of these variables. If major typological *traits*, such as T1 differ from *nexi*, such as Nx.1, by the complexity of their inner structures (major *traits* are simplex and evenly distributed in geolinguistic space, while nexi are complex and tend to have a more labile spatial behavior), the Nx.1 map on the right of Figure 16 shows an interesting property: it seems to be efficient at unveiling *dialects* and *hubs* (in the agentive sense of this term): as to the former, Gascon, Languedocian and Provençal clearly appear in the South, as opposed to Limousin-Auvergnat and Vivaro-Alpine to a smaller extent. As to the latter, in the Gascon dialect, Bearnese surfaces along with a chain of Pyrenean hubs on the central and eastern fringe of the Pyrenees (i.e., Bigorre and Comminges-Couserans). Both T1 and Nx.1 (Figure 16, right) unearth asymmetric agentive hubs: the former in Eastern Limousin, the latter in Western Auvergnat (Puy-de-Dôme and Monts Dore). Buffer zones (in grey on the Nx.1 map, Figure 16) are also enhanced all around the central area made up by Languedocian.

To conclude, Complexity Theory greatly enriches General Dialectology (Rusu, 1985; Léonard, 2012) as well as the study of linguistic diversity in time and space (Nichols, 1992; Goebl, 2005; Dunn, 2023), combining both qualitative and quantitative methods, and not only separately, but organically, as we have attempted to do here. Since the beginnings of dialectology as a modern science, the development of this paradigm has been severely hindered by ideologies and social prejudices about the very conception of dialects in their relationship to referent languages considered as superordinates–national languages. The compass needle has gone from denying their existence (as in Gaston Paris and Paul Meyer's doctrine at the end of the 19$^{th}$ Century) to the use of euphemisms of all kinds, such as Matteo Bartoli's "areal norms" (Bartoli, 1945) or dysphemisms, such as "patois", eagerly used by many French dialectologists and elsewhere. All these models tended to handle dialectal entities as mere objects of study, instead of as ontological subjects, endowed with *agentivity*. The only hierarchy assumed in this prospect was the national language as opposed to the local "smaller languages" or underspecified "varieties", which could hardly be considered organized complexes. Even the division of a linguistic domain aiming at defining a discrete scalar hierarchy such as *macrodialects, dialects, subdialects* and smaller, yet consistently organized entities (such as *singletons* and *splinters*) often remain fuzzy in descriptions carried out by dialectologists, as if such hierarchies should be the privilege of national languages only. Occitan Dialect classifications, from Lamouche (1901) to Sumien (2009), often tended to be redundant, if not teleological, mainly

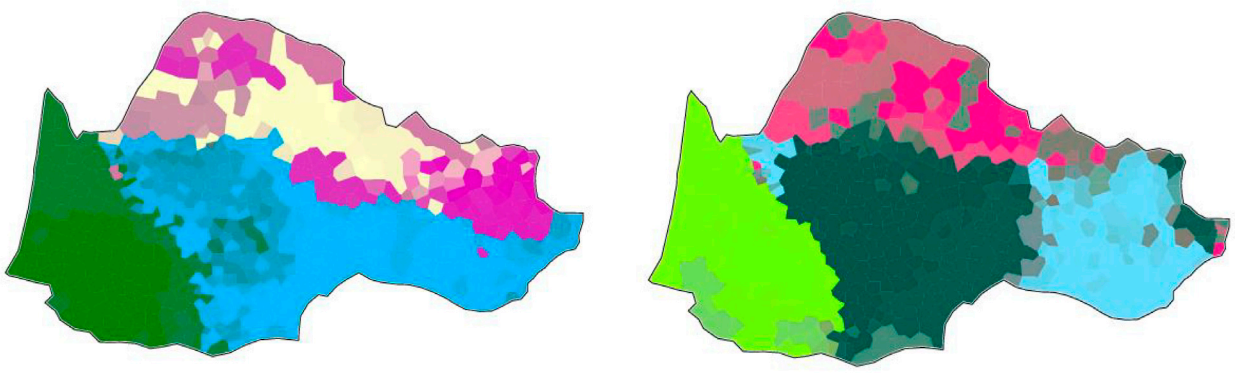



FIGURE 16
Stochastic clustering showing the major cluster divisions for T1 (CA/GA)[24] and Nx.1 (-ELLU$_{suf}$)[25].

based on eponymous dialects. In such a context, deeper level entities (Nerbonne & Kretzschmar's "invisible dialects") were not deemed equally important–nor were they considered at all relevant in formal studies on Gallo-Romance dialect classification. Do macrodialects, dialects, subdialects and varieties exist as empirical entities–not only abstract and historical ones, as Charles Bruneau suggests, attempting a compromise with the Continuist Approach by Meyer & Paris (Bruneau, 1913)? They do, indeed, even if they may surface as near-decomposable categories or entities, as with many phenomena in the physical and social sciences. The same categorical skepticism could be projected over social classes or historical periods, as in Le Goff's recent essay (2014).

Nevertheless, this debate about continuity/discontinuity is somewhat circular and misleading: all categories in the sciences (atom, molecule, cell, planetary systems, galaxies; language, dialect, subdialect, variety; historical periods such as Antiquity, Middle Ages, Modern Ages, etc.) are contrived as heuristic. They serve as building blocks and grids to explore the complexity of the world which surrounds us. This complex world can be grasped in mainly two ways: through qualitative properties (such as philology, isoglosses and typological traits in dialectology) and through quantitative methods (such as dialectometry), as we have done here. Both approaches feed into one another to better grasp underlying, deep patterns evolving below the surface of the grids, and what arises tends to be emerging patterns of competing realities and hierarchies, from a *diachronic* as much as from a *synchronic* standpoint. *Hierarchies* follow *dynamic patterns* of diversity, but they always tend to have *cores* and *peripheries*, *intermediate zones*, *centers of gravity*, more or less *near* vs. *distant relationships*, diverse *degrees* and *manners* of *interaction*, at different *thresholds* and *degrees*. As Veny i Clar (1985: 25–7) reminds us, we now have ways of measuring dialect differentiation in a dialect network–language at threshold 80% of differences, dialect between 50 and 79%, subdialect from 30 to 49%, variety between 20 and 29% and finally, subvariety between 15 and 19%. This grid was initially inspired by Henri Guiter's proposal (1973), but seems too categorical today, in the light of new tools for calculating distances vs. similarities, thresholds and emerging networks, multidimensional scaling of geolinguistic space, etc., as here. Yet, even though categories may still fluctuate, progress is undeniable, since the Meyer & Paris Continuum hypothesis: in a geolinguistic world as mathematically flexible as can be grasped now with algorithmic complexity, dialectology has been freed from *Flatland* and the monodimensional space posited by the Null hypothesis–and which Ferdinand de Saussure adhered to in his *Cours de Linguistique Générale* (Saussure, 1913). CT is particularly effective at challenging common-sense hierarchies and concepts, and one of its main concerns is the agentivity of interacting units and aggregates

24 Items/lemma (Occitan orthography, with English translation): initial onset (unvoiced) ***c****aval* 'horse', ***c****astanhièr* 'chestnut tree', ***c****apèl* 'hat', *camisa* 'shirt', ***c****abra* 'goat', (voiced) ***g****arba* 'sheaf'; postconsonantic unvoiced onset *for****c****a* 'fork', *mos****c****a*, 'fly'; intervocalic (unvoiced) *va****c****a* 'cow', (voiced) *a****g****aça* 'magpie', *bu****g****ada* 'laundry', *ple****g****ar* 'fold'. Places: 662, items: 12, instances: 7,634, characters: 7,716, unique characters: 25, tokens: 7,653 unique tokens: 34.

25 Items in Occitan orthography: *anh****èl*** 'lamb', *cap****èl*** 'hat', *cis****èl*** 'scissors', *cot****èl*** 'knife', *auc****èl*** 'bird', *ved****èl*** 'calf'. Places: 662, items: 6, instances: 3,798, characters: 7,343, unique characters: 44, tokens: 7,168, unique tokens: 57. NB: Gascon has preserved the coda and therefore has the [et/etʃ] ending instead of the original *-el* ending as in Languedocian, which evolved into a secondary *-ew* diphthong in Provençal. In Northern Occitan, this diphthong resulted in a variegated array of reflexes–as in Table 4, column Nx.1.

within dynamic adaptive systems. Exactly what dialectology has been striving for since its very beginning, as we have attempted to suggest in this paper.

## Data availability statement

The data analyzed in this study is subject to the following licenses/restrictions: Data not publicly available but can be made available if requested. Requests to access these datasets should be directed to guylaine.brun-trigaud@unice.fr.

## Author contributions

JL: Conceptualization, Formal Analysis, Investigation, Methodology, Project administration, Supervision, Validation, Visualization, Writing–original draft, Writing–review and editing.

## Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This research has been partly funded within the framework of the ViSoMAC project (*Sociolinguistic Vitality of Languages in Mountain Massifs: Alps and Caucasus*, 2022), supported by the Scientific Council of Paul-Valéry University Montpellier 3.

## Acknowledgments

To Guylaine Brun-Trigaud for creating the THESOC database from the linguistic atlases of the Occitan domain (71 entries) and designing the phonological variables processed for this paper; to Flore Picard for creating the connectograms using Gephi; to Louise Esher and Jean Sibille for editing a more reader-friendly chapter on next insights on Occitan dialectometry, co-authored by JL, Guylaine Brun-Trigaud and Flore Picard, for the *Handbook of Occitan Linguistics* (Esher and Sibille 2024). Gotzon Aurrekoetxea (UPV/EHU) for having insisted so much on the heuristic power of dialectal taxonomies of today and the past and about what can be done with them in the light of Complexity Theory. I occasionally turned to ChatGPT-4 to add additional explanations requested by reviewers, whether in tabular form (Tables 1 and 2) or in prose. This tool was particularly helpful in designing comparative tables on the "advantages and disadvantages of hierarchical methods," distilling my explanations into a clearer, structured format. These tables were crafted in response to a reviewer's feedback, which suggested that the initial argumentation might benefit from more explicit contrastive templates. Additionally, ChatGPT-4 assisted in rephrasing the passage on edit distance scores (beginning of Section 3.1.2), drawing on insights from the initial Gabmap report to make the computational process more accessible for the discerning reader. ChatGPT-4 has proven invaluable in refining such editorial details to enhance the readability of complex data.

## Conflict of interest

The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

## Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

## Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fcpxs.2024.1429114/full#supplementary-material

## References

Allières, J. (2001). *Manuel de linguistique romane*. Paris: Honoré Champion.

Ascoli, G. I. (1875). Schizzi franco-provenzali. *Arch. Glottol. Ital.* 3, 61–120.

Bartoli, M. (1945). *Saggi di linguistica spaziale*. Turin: Università di Torino.

Bec, P. (1963). *La langue occitane*. Paris: Presses Universitaires de France.

Bec, P. (1970). *Manuel pratique de philologie romane, tome 1 (italien, espagnol, portugais, occitan, catalan, gascon)*. Paris: Éditions A. & J. Picard.

Bec, P. (1972). Per una dinamica novèla de la lenga de referéncia: dialectalitat de basa e diasistèma Occitan, *Annales de l'Institut d'Erudes Occitanes: Orientation d'une recherche occitaniste. 4è série* II-6, 39–61.

Bec, P. (1973). *Manuel pratique de linguistique occitane*. Paris, Picard.

Bernsen, M. (2003). "Geschichte des Literatursprache in der Romania: Okzitanisch," in Romanische Sprachgeschichte/Histoire linguistique de la Romania (Berlin-New York: Walter de Gruyter), 1981–1996.

Berrendonner, A., Le Guern, M., and Puech, G. (1983). *Principes de grammaire polylectale*. Lyon: Presses Universitaires de Lyon.

Brun, A. (1923). *Recherches historiques sur l'introduction du français dans les provinces du midi*. Paris (repr. Genève 1973 & EDR/Edition des Régionalismes, Cressé, 2019).

Bruneau, C. (1913). *La Limite des dialectes wallon, champenois et lorrain en Ardenne*. Thèse complémentaire présentée en Sorbonne. Paris, Champion.

Cabaznel, H., *Patrick [1998]-2021,* Histoire des Cévennes, Paris: Presses Universitaires de France.

Camproux, C. (1962). Essai de géographie linguistique du Gévaudan, *tome II*. Paris: Presses Universitaires de France.

Chambon, J.-P. (2004). Les centres urbains directeurs du Midi dans la francisation de l'espace occitan et leurs zones d'influence: esquisse d'une synthèse cartographique, *Revue de linguistique romane*. Romane, Strasbg: Société Linguist 68, 5–13.

Chambon, J.-P., and Olivier, P. (2000), L'histoire linguistique de l'Auvergne et du Velay: notes pour une synthèse provisoire, *Travaux de linguistique de Philol.*, Klincksieck 38, 83–153.

Courouau, J.-F. (2009), L'introduction du français en domaine occitan (XVe-XVIIe siècle): bilan provisoire et perspective de recherche, in *Annales du Midi: revue archéologique, historique et philologique de la France méridionale*, 121–267. Les députés de Montpellier aux États de Languedoc, 317–344.

Dartigue, C. (1950). *Histoire de la Guyenne*. Paris: Presses Universitaires de France.

Derruau-Boniol (1970). Simone & Fel, André *[1963]-1970, le massif central*. Paris: Presses Universitaires de France.

Diller, T. (2006). "Polylectal grammar and Royal Thai," in *Catching Language* (Berlin: Mouton de Gruyter), 565–608.

Dubert, F., and Sousa, X. (2016). On quantitative geolinguistics: an illustration from Galician dialectology. *Dialectologia* 6, 191–221.

Dunn, J. (2023). Syntactic variation across the grammar: modelling a complex adaptive system. *Front. Complex Syst.* 1, 1–23. doi:10.3389/fcpxs.2023.1273741

Esher, L., and Sibille, J. (2024). *Manuel de Linguistique Occitane*. Berlin/New York: Mouton de Gruyter.

Evans, N. (2003). "Bininj Gun-Wok: a pan-dialectal grammar of Mayali Kunwinjku and Kune," in *Pacific linguistics publishers*.

Goebl, H. (2005). La dialectométrie corrélative: un nouvel outil pour l'étude de l'aménagement dialectal de l'espace par l'homme. *Rev. Linguist. Romane* 69, 321–367.

Grassi, C., and Telmon, T. (1979). *Teoria del dialetto. Dialetto e spazio – dialetto e tempo. Corso di dialettologia italiana*. Turin: G. Giappichelli editore.

Gudschinsky, S. C. (1958). Mazatec dialect history: a study in Miniature. *Language* 34, 469–481. doi:10.2307/410694

Guiter, H. (1973) *Atlas et frontières linguistiques*. Paris: Colloque international du C.N.R.S, 930, 61–107.

Guiter, H. (1980) "Limites linguistiques du Velay méridional," in Bulletin historique scientifique, littéraire, artistique & agricole publié par la Société académique du Puy et de la Haute-Loire, 56, Congrès du Puy-en-Velay 19-20 mai 1979, *Le Puy-en-Velay, Editions de la Société Académique*, 116.

Hartmann, F. (2023). *Germanic phylogeny*. Oxford: Oxford University Press, 51.

Jagueneau, L. (2014). *Glossaire de langue gabache. Charles Uguel (1876-1947)*. Bordeaux: Maison des Sciences de l'Homme d'Aquitaine.

Lafont, R. 2003, *Petita istòria europèa d'Occitània*, Canet, Trabucaire.

Lamouche, L. (1901). *Note sur la classification des dialectes de la langue d'oc*. Montpellier: Hamelin Frères.

Le Goff, J. (2014). *Faut-il vraiment découper l'histoire en tranches ?* Paris: Seuil.

Leinonen, T., Çöltekin, Ç., and Nerbonne, J. (2016). Using Gabmap. *Lingua* 178, 71–83. doi:10.1016/j.lingua.2015.02.004

Léonard, J. L. (1991). Variation dialectale et microcosme anthropologique: l'île de Noirmoutier (Vendée, Fr). PhD Diss. Aix-en-Provence: Univ. Provence.

Léonard, J. L. (2012). Éléments de dialectologie générale, *Paris, Michel Houdiard*.

Léonard, J. L. (2016). La valorisation des données dialectales d'oïl du liseré frontalier wallon recueillies par Ferdinand Brunot en 1912: enjeux contemporains. *Diachroniques* 2016, Location: Paris, publicher Sorbonne Université Presses (see https://sup.sorbonne-universite.fr/catalogue/diachroniques) -a bizarre calque from English for "Presses de l'Université de la Sorbonne". . .87–120.

Léonard, J. L. (ed.) (2020a). *Modélisation diasystémique,* Verbum*, Tome XLII*, 1–2. Nancy: Presses Universitaires de Nancy.

Léonard, J. L. (2020b). "Le mazatec: un terrain-monde," in *Du terrain à la théorie : les 40 ans du Lacito, Paris, Editions du Lacito*. Editors I. Leblic and L. Souag, 307–342.

Léonard, J. L. (2022). *Hints at Georgian dialect history: a study in Miniature. Makharoblidze, Tamar*. Wilmington-Malaga-Sevilla: Vernon Press, 3–28.

Léonard, J. L., and Albinet, L. (2023). *Ecol. linguistique des milieux littoraux étude de cas dialectométrique sur les langues variétés dialectales du Golfe du Lion, selon les données de Louis Michel*. 7th of July 2023 hal-04229391.

Léonard, J. L., Brun-Trigaud, G., and Picard, F. (2024). "Atlas linguistiques et perspectives dialectométriques," in *Sibille, Jean & Esher, Louise (dir.)x*, Manuel de Linguistique Occitane (Berlin & NY: Mouton de Gruyter).

Léonard, J. L., and Dell'Aquila, V. (2012). "Mazatec (popolocan, Eastern otomanguean) as 1212 a multiplex sociolinguistic small world in Bereczki, Urmas (Dir.): *The languages of smaller populations: risks and possibilities*, Lectures from the Tallinn Conference, Tallinn, Ungarian, March 16–17, 2012 (Miscellanea Hungarica: Institute's Series), 27–55.

Leroy-Ladurie, E. (1977). Occitania in historical perspective. *Rev. Fernand Braudel Cent.* 1, 20–30.

Leroy-Ladurie, E. (2001) "Histoire des régions de France," in *La périphérie française, des origines à nos jours*. Paris: Seuil.

Levenshtein, V. I. (1966). Bin. Codes Capab. Correcting Deletions, Insertions, Reversals, Soviet Phys. – Doklady 10/8, 707–710.

Martel, P. (2003). Histoire externe de l'occitan, in Romanische Sprachgeschichte/ Histoire linguistique de la Romania (Berlin-New York: Walter de Gruyter), 829–839. doi:10.1515/9783110146943.1.7.829

Martel, P. (2019). *Histoire de l'Occitanie. Le point de vue occitan*. Fouenant: Yoran.

Michel, L. (1964), *La langue des pêcheurs du Golfe du Lion. Introduction: I. L'homme et la mer, II*. Dialectologie côtière, Paris, d'Artrey.

Miestamo, M. (2017). Linguistic diversity and complexity. *Lingue linguaggio XVI*. 2, 227–253.

Mufwene, S. (2001). *The Ecology of Language evolution*. Cambridge: Cambridge University Press.

Mühlhäusler, P. (1992). *Polylectal grammar*. In W. Bright (ed.). International Encyclopaedia of Linguistics, 3. Oxford: Oxford University Press. 243–245.

Nauton, P. (1974). *Géographie phonétique de la Haute-Loire*. Paris: Les Belles Lettres.

Nerbonne, J., and Kretzschmar, W. (2003). Introducing computational techniques in dialectometry. *Comput. Humanit.* 37 (3), 245–255. doi:10.1023/a:1025064105053

Nichols, J. (1992). *Linguistic diversity in time and space*. Chicago: Chicago University Press.

Paris, G. (1888). Les parlers de France; lecture faite à la réunion des sociétés savantes, le samedi 26 mai 1888. *Rev. Des. patois gallo-romans* 2, 161–175.

Polian, G. L., Léo, J., Heinsalu, E., and Patriarca, M. (2014). "Variación dialectal de la morfología tseltal (Maya occidental) en los ámbitos morfológico, fonológico y léxico: un enfoque holístico del diasistema," in *Patterns in Mesoamerican morphology*. Editors J. L. Léonard and A. Kihm (Paris: Michel Houdiard Editeur), 280–303.

Prokić, J., and Nerbonne, J. (2008). "Recognize groups among dialects," in *International Journal of Humanities and Arts computing, special issue on language variation*. Editors J. Nerbonne, C. Gooskens, S. Kürschner, and R. van Bezooijen, 153–172.

Puech, G. (1979) *Les parlers maltais. Essai de phonologie polylectale*, 2. Lyon: University. PhD Dissertation.

Ronjat, J. (1930). *Grammaire istorique [sic] des parlers provençaux 1267 modernes*. Montpellier: Société des Langues Romanes Vol. 4.

Rusu, V. (1985). *Dialettologia generale*. Roma: Zanichelli.

Saussure, F. de (1891), *Troisième conférence à l'Université de Genève (novembre 1891)*, CLG: *édition critique par Rudolf Engler*, fascicule 4, *Wiesbaden, Otto Harrassowitz*.

Saussure, F. de (1975). *Cours de Linguistique générale, éd*. Paris, Payot: Tullio de Mauro.

Sauzet, P. (2016). "Jules Ronjat: La syntaxe et la langue occitane," in Autour des travaux de Jules Ronjat, 1913-2013. Unité et diversité des langues. Théorie et pratique de l'acquisition bilingue et de l'intercompréhension, *Éditions des archives contemporaines*. Editor P. Escudé 9782813001917. hal-02275830.

Simon, A. H. (1962). The architecture of complexity. *Proc. Am. Philosophical Soc.* 106 (6), 467–482.

Sumien, D. (2009). *Classificacion dei dialèctes occitans*. Linguistica Occitana, 7. Available at: www.revistadoc.org.

Surrel, V. (2022). *Les textes occitans de l'ancien Velay. Problèmes d'élaboration philologique et d'exploitation linguistique d'un corpus*. PhD dissertation. Paris: Université de Paris 8 University & École nationale des chartes.

Veny, i C, J. (1985). "Introducció a la dialectologia catalana,". Barcelona: Enciclopèdia Catalana.

Watts, D. J. (1999). *SmallWorlds: the dynamics of networks between order and Randomness*. Princeton, NJ: Princeton University Press.

Watts, D. J. (2004). The "new" science of networks. *Annu. Rev. Sociol.* 30, 243–270. doi:10.1146/annurev.soc.30.020404.104342

Weinreich, U. (1954). Is a structural dialectology possible? *Word* 10, 388–400. doi:10.1080/00437956.1954.11659535

Wolff, P. (1967). *Histoire du Languedoc*. Toulouse: Privat éditeur.

Wüest, J. (2003). "Evolution des frontières des langues romanes: la Galloromania," in *Romanische Sprachgeschichte/Histoire linguistique de la Romania* (Berlin-New York: Walter de Gruyter), 646–657. doi:10.1515/9783110146943.1.5.646

Zufferey, F. (2008). "Histoire interne de l'occitan," in *Romanische Sprachgeschichte/Histoire linguistique de la Romania* (Berlin-New York: Walter de Gruyter), 2998–3020.