Isogloss
Open Journal of Romance Linguistics

Isogloss
Open Journal of Romance Linguistics

Received: 02-03-2024
Accepted: 02-11-2024
Published: 22-03-2025

Computer modelling of innovations relative to Latin in contemporary Romance dialects

Philippe Boula de Mareüil
Université Paris-Saclay, CNRS, LISN, Orsay, France
philippe.boula.de.mareuil@lisn.fr

Marc Evrard
Université Paris-Saclay, CNRS, LISN, Orsay, France
marc.evard@lisn.fr

Alexandre François
LATTICE, CNRS, École normale supérieure, Paris, France
alexandre.francois@ens.psl.eu

Antonio Romano
Università degli Studi di Torino, LFSAG, Turin, Italy
antonio.romano@unito.it

How to cite: Boula de Mareüil, Philippe, Evrard, Marc, François, Alexandre & Antonio Romano. 2025. Computer modelling of innovations relative to Latin in contemporary Romance dialects. In *Romance minority languages: A challenge for linguistic theory*, eds. Michela Russo & Shanti Ulfsbjorninn. Special issue of *Isogloss. Open Journal of Romance Linguistics* 11(3)/8, 1–32.
DOI: https://doi.org/10.5565/rev/isogloss.423

Abstract

This study relies on a corpus illustrating several dozen Romance dialects from France, Belgium, Switzerland, Italy, Spain and Portugal, for which 145 innovations relative to Latin have been encoded in the form of 1 (presence) or 0 (absence). Based on contemporary recordings (translations of Aesop's 100-word fable "The North Wind and the Sun" and another 100-word list), following the principles of dialectometry, the Comparative method and especially historical glottometry, we propose computational

tools to address the relationships and classifications amongst these Romance varieties. Results of data-mining techniques confirm the robustness of a North/South divide — with the *Oïl* area being, by far, the most innovative — and, secondarily, an opposition between the South-West (mainly Ibero-Romance) and the South-East (mainly Italo-Romance, more conservative). Among the most important/discriminant features are the palatalisation of Latin CA, which characterises the majority of northern Gallo-Romance dialects, and the simplification of geminates north of the La Spezia-Rimini line. Most innovations relate to phonetic/phonological traits. However, we also consider morphosyntactic and lexical features, such as non-null subject in northern Gallo-Romance varieties, and the substitution of CUM 'with' by APUD > *amb* in Occitano-Romance varieties. By retaining only morphosyntactic innovations, we still find a North vs. South-East- vs. South-West tripartition.

Keywords: phylogenetic classification, Wave model, computational dialectometry, historical glottometry, Romance varieties.

1. Introduction: identifying innovations among Romance dialects

Stammbaums (i.e., phylogenetic trees) inspired by biology have had some success in linguistics to account for language diversification. For Romance languages derived from Latin, in particular, in which abundant written sources exist, they have been used by Neogrammarians since the late 19th century. Nevertheless, the model has long been criticised for its simplistic assumptions. "One of the main limitations of the tree model of language evolution is the underlying assumption that the protolanguage develops independently in each branched subcommunity. Such an idealised situation rarely occurs; usually, innovations are born in one community and spread to other adjacent communities" (Patriarca *et al.*, 2020: 79). The scenario portrayed by repeated splits with complete loss of contact is therefore quite unrealistic. Proponents of the Wave theory (Schmidt, 1872; Saussure, 1916), observing that the tree model is unsuitable for representing the genealogy of modern languages, addressed this difficulty by spinning a new metaphor: innovations are like waves, which happen independently, imprinting their marks on the landscape; and crucially, these successive waves of innovations often define intersecting patterns, which cannot be represented with a tree (François, 2014). The geography of these innovations can be reconstructed by historical linguists, provided they have access to enough relevant data.

Indeed, whichever framework one wants to adopt, an important issue for unwritten dialects is the availability and dispersion of existing data. Along the first half of the 20th century, linguistic atlases documented the dialects spoken in France (Gilliéron & Edmont, 1902–1910), Switzerland (Jaberg & Jud, 1928–1940), Wallonia (Haust *et al*., 1953–2011), Italy (Bartoli *et al.*, 1995) and the Iberian Peninsula (García Mouton *et al*., 2016).[1] After World War I, Ronjat (1930) listed 19 traits believed to be characteristic of Occitan, most of which are phonetic traits. Some of these features were taken up by linguists like Bec (1995), and later Sumien (2008), to structure the so-called Occitano-Romance supradialectal space, which includes Catalan. Later, quantitative approaches were carried out by Pei (1949), covering seven Romance languages (including Occitan) but considering only the accented vowel system; and a comparable study was conducted by Blanchet (1996), with about forty phonetic features individuating the Poitevin and Gallo dialects, in the west of France. Morphosyntactic features are seldom included.

Concerning Italo-Romance varieties, scholars have attempted to draw up various classifications since Dante's (1303–1304) *De vulgari eloquentia*, principally based on ethno-geographic criteria, pointing out the effects of Celtic, Etruscan or Italic substrates (Cugno, 2023). Ascoli (1882/1885) adopted a genealogical approach, measuring the affinity to Latin, translated into a comparison with Tuscan — the reference dialect of Florence being considered as the root of the Italian national language. Rohlfs (1937) proposed a classification based on the areal diffusion of 18 linguistic phenomena and the identification of bundles of isoglosses, facilitated by the publication of the first linguistic atlases. Pellegrini (1973, 1977) used an approach based on the contextual application of sociolinguistic and geolinguistic criteria — dozens of features partly inspired by Rohlfs' (1937) model. Among these features aiming at measuring dialect similarity, one can cite phonetic features such as the presence of velar nasals or the palatalisation of CT; morphosyntactic and lexical peculiarities are rarer or excluded.

Dialectological studies, usually, take a synchronic perspective and thus do not attempt at identifying which features are innovative. In particular, edit distances (ED), popular in dialectometry (Séguy, 1973; Goebl, 2002, 2003; Heeringa, 2004; Gooskens, 2005a, 2005b; Nerbonne *et al*., 2007; Beijering *et al*., 2008; Scherrer, 2021; Brun-Trigaud, 2021), are based solely on observables in synchrony. They most often use the Levenshtein algorithm, which computes the minimum number of edit operations (insertion, deletion, substitution) necessary to turn a character string into another one. This is the case, in particular, of the Gabmap web application for dialectometry (Leinonen *et al*., 2016).

In the case of Romance languages, most features can be compared, fortunately, with the ancestral state of Latin — thus making it possible to identify whether phonetic, morphosyntactic and lexical features represent innovations. This has an important advantage: namely, the possibility to analyse the data within the framework of the Comparative method, for which the distinction between retentions and innovations is crucial (François, 2014: 164). The Comparative method, which emerged in the 19th century (Schleicher, 1861 [2010]), gave rise to the cladistic paradigm (Gaillard-Corvaglia *et a*l., 2007). Unlike ED-based methods, the application of such a paradigm requires the intervention of linguists, who invest their knowledge about the history of individual changes. The same method also underlies the more recent approach of historical glottometry (François, 2014; Kalyan & François, 2018; François & Kalyan, f/c), inspired by the Wave model. In sum, it should be possible to analyse modern data from Romance languages through the lens of the historical linguist's comparative method — whether this results in visualisation efforts in terms of trees (or dendrograms), wave diagrams, two-dimensional planes or choropleth maps, taking into account geographical space.

[1] What was spoken in earlier stages — or what is spoken today in some undocumented dialects — is less well known.

Recent computational methods usefully complement each other to highlight dialectal phenomena (Léonard *et al*., 2024); however, because they build on traditional linguistic atlases, the results are mostly based on isolated words. Richer information, particularly grammatical, can be gathered from collections of parallel texts. Since the early 19th century, a whole tradition has consisted in having the parable of the Prodigal Son translated into a number of Romance dialects (Coquebert de Montbret, 1831). Another story, "The North Wind and the Sun", one of Aesop's fables, has been used by the International Phonetic Association (IPA), for more than a century, to illustrate many languages and dialects spoken in the world (Romano, 2016). On the basis of this Aesop fable, a speaking atlas of the regional languages of France was designed (Boula de Mareüil *et al.,* 2018), and later extended to other European countries (Boula de Mareüil *et al*., 2021). The linguistic atlas, available at https://atlas.limsi.fr, allows visitors to hear and read thisone1-minute story in hundreds of versions, in minority languages or dialects (the difference between a language and a dialect being ill-defined). Most speakers of the atlas, recorded in the field, also translated a list of a hundred words into their varieties: some of these items may be read and listened to on the site of this online atlas (Knyazeva *et al*., 2022).

Based on these digital data, the present article proposes to use computational tools to address the following questions. Is there more variation between northern and southern Romance dialects, or between the west and east of the domain? How can we quantify dialectal variation? To what extent do the groupings depend on the levels considered (phonetic, morphosyntactic or lexical)? The present study relied on a sub-corpus illustrating several dozen Romance dialects from France, Belgium, Switzerland, Italy, Spain and Portugal. We assessed each of these Romance varieties for more than a hundred innovations relative to Latin; we then applied a range of classification techniques in order to visualise the emerging clusters (e.g., in the form of trees, or projections into a two-dimensional plane) and draw the principal isoglosses. The results of the different methods of analysis and calculation were confronted in order to propose a synthesis, making it possible to reassess the location of the main dividing lines between dialect groups. This study sheds new light on Romance dialectology, contributing to model the dynamics of territorial expansion since the breakup of Latin.

The remainder of this paper is organised as follows. We will first present our corpus and methods, explaining our parsimonious selection of survey points in the Romance domain and innovations with respect to Latin. Then, we will report on the results obtained using various data-mining techniques: they suggest that even a limited-size corpus is adequate to accurately predict the topology of Romance dialects.

Table 1. Survey points located in France, Belgium, Switzerland, Spain and Italy, with the 3-letter abbreviations (label) of the corresponding dialects/languages.

label	dialect / language	France	Belgium (B), Switzerland (S)	Spain	Italy
ang	*Angevin*	Athée			
bou	*Burgundian*	Montsauche-les-Séttons			
cen	*Central Oïl*	Paris (= French)			
cha	*Champenois*	Les-Hautes-Rivières	Vresse-sur-Semois (B, cha1)		
frc	*Franc-Comtois*	Banvillars	Cœuve (S, frc1)		
gal	*Gallo*	Sérent			
lor	*Lorain*	Labaroche	Virton (B, lor1)		
mai	*Mainiot*	Neufchâtel-en-Saosnois			
nor	*Norman*	Réville			
pic	*Picard*	Lille	Dour (B, pic1)		
poi	*Poitevin-Saintongeais*	Saint-Pardoux			
wal	*Walloon*	Vireux-Molhain (wal1)	Liège (B)		
frp	*Franco-provençal*	Thollon-les-Mémises	Treyvaux (S, frp1)		Introd (frc2)
cro	*Crescent*	Fursac			
gas	*Gascon*	Momas		Vielha (gas1)	
lan	*Languedocian*	Najac			
noc	*Northern Occitan*	Charmensac			Pomaretto
pro	*Provençal*	Sanary-sur-Mer			
cat	*Catalan*	Perpignan (cat1)		El Vendrell	Alghero (cat2)
cor	*Corsican*	Corte			
lig	*Ligurian*	Bonifacio (lig1)			Varazze

2. Corpus and methods

2.1. Selection of survey points

The study relies on a speaking atlas laid out in Boula de Mareüil *et al*. (2018, 2021), which involves more than 200 survey points in Romance varieties, most of which were collected between 2014 and 2020. In the context of the present study — meant as a proof of concept — we selected 61 among these survey points. Our choice fell on "average" varieties, rather unmarked or in the middle of known linguistic areas. As far as possible, we also favoured the speakers who could be contacted, so as to be able to ask them questions in case of doubt or missing information; such follow-up dialogues are not possible when the corpus is based on surveys dating back more than 100 years.

These informants, well anchored in their community, representing various socio-professional backgrounds and usually retired, had a good knowledge of their dialect. We retained one speaker for each of the Romance dialectal areas mapped in https://atlas.limsi.fr: 21 in France, 4 in Belgium, 4 in Switzerland, 6 in Spain, 1 in Portugal, 25 in Italy.

Table 2. Survey points located in Italy, Switzerland, Spain and Portugal, with the 3-letter abbreviations (label) of the corresponding dialects/languages.

label	dialect/language	Italy	Switzerland	Spain	Portugal
emi	*Emilian-Romagnol*	Bologna			
lom	*Lombard*	Milan	Lugano (lom1)		
pie	*Piedmontese*	Corio			
ven	*Venetian*	Padua			
gsa	*Gallurese-Sassarese*	Sassari			
tos	*Tuscan*	Florence			
laz	*Laziale*	Rome			
mac	*Marchigiano*	Macerata			
sab	*Sabine*	L'Aquila			
umb	*Umbrian*	Umbertide			
abr	*Abruzzese*	Pescara			
cal	*Calabrian*	Gizzeria			
cam	*Campanian*	Naples			
luc	*Lucanian*	Policoro			
mol	*Molisan*	Isernia			
pug	*Apulian*	Bitonto			
sal	*Salentinian*	Lecce			
sic	*Sicilian*	Palermo			
sal	*Sardinian*	Nuoro			
fri	*Friulian*	Ragogna			
lad	*Ladin*	Badia			
rum	*Romansh*		Zuoz		
ara	*Aragonese*			Uesca	
ast	*Asturian*			Oviedo	
cas	*Castilian*			Madrid	
gle	*Galician*			Vivalda	
por	*Portuguese*				Lisbon

Some dialects, spread across several countries, are represented several times in our selection:

between France and Belgium: Walloon, Picard, Lorrain and Champenois;
between France and Switzerland: Franc-Comtois;
between France and Spain: Gascon (Béarnese, Aranese);
betweenFrance and Italy: Northern Occitan (Vivaro-Alpine) and Ligurian (Bonifacien, Genoese);
between Switzerland and Italy: Lombard.

Some languages/dialects are even spread across three countries:

between France, Italy and Switzerland: Francoprovençal;
between France, Italy and Spain: Catalan (including Algherese, in Sardinia).

Table 1 and 2 list the survey points located in France, Belgium, Switzerland, Spain and Italy. The first rows of Table 1 (from Angevin to Walloon) correspond to northern (*Oïl*) Gallo-Romance varieties, whereas the following rows (from Gascon to Provençal) are southern (*Oc*) Gallo-Romance dialects. Between the two domains, we have Francoprovençal (named after Ascoli 1877); as well as the linguistic Crescent, a transition area between *Oïl* and *Oc* languages — so named by Ronjat (1913) because of its half-moon shape — in the centre of France. As for Corsican and Ligurian, they are Italo-Romance varieties.

Other Italo-Romance varieties are listed in Table 2: northern varieties in the first rows (from Emilian-Romagnol to Venetian), followed by central (from Laziale to Umbrian) and southern varieties (from Abruzzese to Sicilian). Gallurese-Sassarese, Tuscan and Sardinian play a separate role, because they do not pattern with the other dialect groups (see Pellegrini, 1977). As for Friulian, Ladin and Romansh (spoken in Switzerland and illustrated here by the Puter dialect), they belong to Rhaeto-Romance. The bottom rows (from Aragonese to Portuguese) form the Ibero-Romance set.

Due to practical considerations in data collection, we have chosen to focus on Western Romance varieties spoken in Europe. As a result, we are aware that our corpus lacks data from Balkan-Romance varieties of Romania and neighbouring countries. Our selection may also show a certain imbalance to the disadvantage of Ibero-Romance varieties. In Spain, Castilian (or Spanish) is relatively homogenous, despite peculiarities in Andalusia (Herrero de Haro & Hajek, 2020). The same goes for the Portuguese spoken in Portugal, of which it is the official language — the only other recognised language being Mirandese, an Astur-Leonese variety. The Portuguese spoken in Brazil has well-known specificities, much like the Spanish spoken in other Latin American countries or the French spoken in Africa (e.g., Cámara Jr, 1970; Lipski, 1996; Ploog, 2002; Boutin, 2003); however, despite their demographic weight, we did not include them in this study — largely due to practical considerations in carrying out fieldwork. Nor will we consider neo-Romance Creoles, which show profound typological differences regarding morphosyntax compared to European languages but superficial phonetic adaptations (e.g., Kihm, 2005). In sum, we chose to concentrate on a compact space, that of Western Romance varieties spoken in continental Europe, for which we were able to meet individual speakers and collect new data.

2.2. Description of the collected data

Observations can be made about pronunciation, grammar and lexicon, based upon our corpus. Let us start with traits shared by several Gallo-Romance varieties. Table 3 provides samples of the fable "The North Wind and the Sun", showing versions in standard French[2] and two dialectal varieties, respectively collected in Fursac (Crescent

[2] The English version of the fable reads as follows: "The North Wind and the Sun were disputing which was the stronger, when a traveller came along wrapped in a warm cloak. They agreed that the one who first succeeded in making the traveller take his cloak off should be

Limousin, central France) and Treyvaux (Francoprovençal, Gallo-Romance Switzerland). The recordings for these texts were — together with those of vocabulary lists — the empirical basis for our analyses.

Table 3. Orthographic transcripts of the fable "The North Wind and the Sun", showing versions recorded in French (Paris), a Crescent variety (Fursac) and Francoprovençal (Treyvaux)

French (Paris)	La bise et le soleil se disputaient, chacun assurant qu'il était le plus fort, quand ils ont vu un voyageur qui s'avançait, enveloppé dans son manteau. Ils sont tombés d'accord que celui qui arriverait le premier à faire ôter son manteau au voyageur serait regardé comme le plus fort. Alors, la bise s'est mise à souffler de toute sa force mais plus elle soufflait, plus le voyageur serrait son manteau autour de lui et à la fin, la bise a renoncé à le lui faire ôter. Alors le soleil a commencé à briller et au bout d'un moment, le voyageur, réchauffé, a ôté son manteau. Ainsi, la bise a dû reconnaître que le soleil était le plus fort des deux.
Crescent (Fursac, France)	La bise e le solelh se disputavan. Chascun assurave qu'eu ère le pus fòrt. Quand ilhs an veüt un voiatgeur que s'avançave, engonçat dins son mantel, ilhs son tombats d'acòrd que queu-qui qu'arriverí le premier a li far enlevar son mantel serí gaitat come le pus fòrt. Alòr, la bise s'a mese a bufar de tote sa fòrce, mas mai 'la bufave, mai le voiatgeur sarrave son mantel autorn de se. E, a la fin, la bise a renonçat a le li far pausar. Alòr le solelh a començat a brilhar, e, au bot d'un moment, le voiatgeur, reschaufat, a enlevat son mantel. De mèsme, la bise a degut reconeistre que le solelh ère le pus fòrt de los dos.
Francoprovençal (Treyvaux, Switzerland)	La bije è le chèlà chè kontrèyivan ; tsakon achurin k'irè li, le pye yô ; kan l'an yu on voyadyà ke ch'avanhyivè, inbortoyi din chon mantô. I chon tseju d'akouâ ke chi k'arouvèrè le premi a fére trére chon mantô ou voyadyà, cherè yu kemin le pye yô. Adon la bije chè betâye a choyâ dè totè chè fouâchè, ma, mé y chohyavè, mé le voyadyà charâvè chon mantô outoua dè li, pu po fourni, la bije l'a pyakâ dè l'i fére ôthâ. Adon le chèlà l'a keminhyi a èhyiri, è ou bè de na vouêrbèta, le voyadyà, rètsoudâ, l'a tré chon mantô. Dinche, la bije, l'i a fayu rèkonyèthre ke le chèlà irè le pye yô di dou.

The comprehensive list of innovations analysed as present (1) or absent (0), in different dialects, is available online (see note 10).

2.2.1. Observations on Gallo-Romance dialects

In the phonological domain, Latin /k/ before /a/ has been maintained in Catalan, in Southern Occitan as well as in Norman — to the north of what is known as the Joret (1881) line — and in Picard, resulting in such forms as *recauf(f)é* 'warmed up' (< EXCALFATUM).[3] In other varieties, /k/ before /a/ evolved:

considered stronger than the other. Then the North Wind blew as hard as he could, but the more he blew, the more closely did the traveller fold his cloak around him; and at last the North Wind gave up the attempt. Then the Sun shined out warmly, and immediately the traveller took off his cloak. And so the North Wind was obliged to confess that the Sun was the stronger of the two." (IPA, 1999).

[3] While Norman and Picard do not exhibit the change CA > [ʃ], they have developed a [ʃ] in another context — namely, in reflexes of Latin CE, CI or TI. Thus, Late Latin FORTIA 'strength' (a nominalised form of the neuter plural adjective *fortis*) is reflected as *forche* [fɔʁʃ].

into [ʃ] around central *Oïl* dialects, resulting in forms like French *réchauffé* [ʁeʃofe];

into an interdental [θ] (spelled <sh>) in several Francoprovençal varieties, resulting in forms like *resheudô* [rə**θ**ødo];

into an affricate [ts] in Northern Occitan (as well as in Aosta Valley Francoprovençal), resulting in forms like *eschaufat* [ej**ts**awfa].

Another feature found in southern Gallo-Romance varieties and Walloon (north-eastern France and Belgium) is the retention of Latin /s/ in syllable coda: e.g., Occitan/Catalan *estar* 'to be', Walloon *estant* 'being'; this contrasts with other *Oïl* varieties, which lost that /s/, and only retained the epenthetic [e] (e.g., French *étant* 'being').

Perceptively very salient, an [h] can be heard in several varieties, albeit with different origins, and consequently treated differently in our annotation scheme (see § 2.3):

in Poitevin-Saintongeais, the laryngeal [h], spelled <jh>, corresponds to French [ʒ] (e.g., *najhe* [nah] 'swimming', French *nage* [naʒ]);

in Gascon, the same phoneme, spelled <h>, results from the debuccalisation of Latin /f/ (e.g., *hòrt* [hɔɾt] 'strong').

The rhotic phoneme surfaces as an apical [ɾ, r] throughout southern varieties as well as in Mainiot (*Oïl* domain), but as a dorsal [ʁ] in other northern Gallo-Romance dialects. Regardless of its realisation, the rhotic /R/ correlates with word-initial epenthesis/metathesis in forms like Picard *arnonché* or Norman *arnounchit* 'renounced'. As for the lateral /l/, it can be palatalised into [j] after a consonant, a yod which may be spelled <ll> in western varieties (e.g., *soufflét* 'was blowing') and in Burgundian. In the Occitan domain, the vocalisation of /l/ in coda position, widely observed in Provençal, is also attested in Gascon and Northern Occitan.

Rhotacism of L is found in Gascon, Northern Occitan and sporadically elsewhere in the word *sorelh* < *SOLICLU(M) 'sun'. Still in the Occitan domain, betacism (merger between [b] and [β]/[v]) is found in Gascon, Languedocian and sporadically elsewhere, and is shared with Catalan (e.g., *arribaria* 'would arrive'). Also, while Catalan and Gascon preserve the labiovelar /kʷ/ in *quan* 'when', other Gallo-Romance varieties simplified it to a mere velar [k], at least in this word.

Regarding vowels, nasal vowels and mid front rounded vowels do not belong to the Occitan system, except in the north of the domain with the Crescent, but they are found in all the systems of the *Oïl* zone. For example, Latin -EN- gave [ɛ̃] in Picard and Walloon, and [ɑ̃] in the other *Oïl* varieties (e.g., *momint* vs. *moment*).

Regarding morphophonology, infinitives in [a] in Catalan and Occitan for verbs of the 1st group continue Latin /a/. We also find this [a] in Poitevin-Saintongeais, resulting from a reduction of the diphthong /ai/ (e.g., *bufâe* 'to blow'), but the archiphoneme /E/ in the rest of the *Oïl* domain.

The agent suffix corresponding to French *-eur* is raised to [u] in Burgundian, Franc-Comtois, Lorrain and Gallo (e.g., *vayaijou* 'traveller'), but it is fronted in other *Oïl* varieties. In the Occitan domain, we have [u] in Gascon in *viatjador* 'traveller' — with a specialisation of the suffix *-ador* < -ATŌRE(M), reflecting the oblique case (*cas régime*) of the medieval language — while other varieties selected the form *viatjaire*, reflecting the subject case. Regarding another very productive suffix, reflexes of the Latin diminutive -ELLU(M) in Picard, Champenois and Burgundian exhibit a yod (e.g., *mantiau* 'coat').

In the domain of inflectional morphology, the verbal system calls for some remarks. The simple past preterite is in use throughout the Occitan domain, in Norman and Poitevin-Saintongeais (with forms in [i] for verbs of the 1st group) as well as in Walloon (with forms in [a] even for verbs of the 3rd group).

In the imperfect, the verbs of the 1st group have a /v/ in Catalan, Occitan, Francoprovençal and Walloon. We thus have forms like *avançava* 'was moving forward' in Occitan (see *avançave* in the Fursac text of Table 3), *avancive* in Walloon, and *avensive* in Francoprovençal (compare *avanhyivè* in the Treyvaux text of Table 3). In *Oïl* varieties, the ending of the imperfect ultimately reflects Latin -EBA-, which was extended to all verbs and merged with imperfects in -ABA- (Haust *et al*., 1953–2011; Zink, 1989; Guérin *et al*., 2021).

In the realm of morphosyntax, the definite article is used in possessives like 'his coat' in Catalan (*el seu manto*) and Gascon (*lo son manto*). Like Catalan, most Occitan varieties are pro-drop, but the Crescent uses a subject personal pronoun, as do *Oïl* varieties. Most often, a 3rd person pronoun anaphorically repeats the subject noun phrase in Angevin, Mainiot and Picard. The double marking of negation (or its marking with a single postposed adverb of the type *pas* 'not' or *nén* in Walloon) can also be observed throughout France, including in northern Catalan. Other grammatical features are specific to certain dialects, even if they do not extend to the entire corresponding domains: e.g., Gascon expletives or pronouns *ac* 'it' and *eth* 'him'. We will not dwell on these features, which have been the subject of significant dialectological documentation (Chambon & Greube, 2002; Chevrier & Gautier, 2002; Poplineau, 2006; Abalain, 2007; Sumien, 2007; Martin, 2011; Tillinger, 2016; Legeard, 2020; Quint, 2023, to cite just a few recent publications).

In the lexical domain, we note some peculiarities, which are reported in the *Französisches Etymologisches Wörterbuch* [FEW] (Wartburg, 1922–2002). Let us cite *d'assent* to express agreement and *pal(e)tot* to translate 'coat' in western Gallo-Romance varieties.

2.2.2. Observations on Ibero-Romance dialects

In the phonological domain, Aragonese has some traits in common with Catalan — and Portuguese. Some are shared retentions: for instance, these varieties preserve Latin /f/ (e.g., FACERE > *fer* 'to do'), where contemporary Spanish has lost it — with the reflex *hacer* [aθɛr]. The interdental [θ] — spelled <z> in Aragonese, regardless of the following vowel (as in *prenzipió* 'began') — and the betacism phenomenon can also be heard in Spanish, Astur-Leonese and Galician. The diphthongisation of Latin stressed E and O is shared by Spanish, Astur-Leonese and Aragonese: e.g., *viento* 'wind', *fuerte* 'strong', where we note the preservation of the final vowel ([o] and [e], respectively), whereas Catalan *vent* and *fort* reflect regular apocopes. Traits shared by Galician and Portuguese include the rhotacism from former /l/, found for instance in *obrigar* 'to oblige'; the palatalisation of GN > [ɲ] in RECOGNOSCERE > *recoñecer* 'to recognise'; the reduction of the unstressed vowel inventory to three units (Regueira, 1996) as well as the preservation of the final vowel of the Latin suffix -ITATE(M) (e.g., *superioridade* 'superiority'), unlike most varieties. Other Galician innovations are original compared to its neighbours: the emergence of the unvoiced fricative [ʃ] (spelled <x>, as in *viaxeiro* 'traveller') where Portuguese has /ʒ/ and Spanish has [x];

the reflex of Latin QU > [k] instead of *[kʷ] (e.g., QUANDO > *cando* 'when') and the velar nasal in *unha* [uŋa] (vs Portuguese *uma*) < UNA(M) 'a'.

In the morphological domain, the systems of determiners (e.g., *o sol* 'the sun') and possession (in the form of determiner + possessive adjective) are similar in Aragonese, Galician and Portuguese (Nagore Laín, 1989). Nevertheless, verb inflection in Aragonese presents some originalities in comparison with its neighbours (e.g., ERAT > Aragonese *yera* 'was', Spanish *era*). Also, Aragonese preserved everywhere the consonant B of the Latin imperfect (e.g., VENIEBAT > Aragonese *beniba* 'vas coming') — whereas other Ibero-Romance varieties lost that B in the 2nd and 3rd conjugations (e.g., VENIEBAT > Spanish *venía*, Portuguese *vinha*).

At the lexical level, let us mention Galician *e mais* (literally 'and more') compared with Portuguese *e* 'and'.

2.2.3. Observations on Italo- and Rhaeto-Romance dialects

As far as pronunciation is concerned, our data confirms that geminates were simplified in northern Italo-Romance varieties (hence the spelling *acòrd* 'agreement', from the Latin verb *ACCORDARE, in Lombard and Piedmontese). Conversely, the so-called *raddoppiamento fonosintattico* 'phonosyntactic doubling' phenomenon only applies to central and southern varieties: this phenomenon triggers the gemination of a word-initial consonant after certain words — the list of which, though, is dialect-dependent (Loporcaro, 1997a).

Noteworthy cases of phonetic change include the loss of intervocalic L in Venetian (e.g., *soe* < SOLE(M) 'sun') and that of intervocalic R in Ligurian (e.g., *ëa* < ERAT 'was'); Ligurian may also have dropped L in specific contexts like *cädo* [ka:du] < CALIDU(M) 'hot', with compensatory lengthening (Garassino & Filipponio 2021).

While the Latin cluster PL is usually preserved in Rhaeto-Romance, most Italo-Romance varieties reflect its palatalisation to [pj]. That cluster is further palatalised in Ligurian: PLUS > *[pju] > *ciù* [tʃy] 'more'; the same word became *(c)chiù* [(k)kju] in Sabine and all southern Italo-Romance dialects. On the other hand, Sardinian shows rhotacism here: PLUS > *pru*. Sardinian is also original in retaining Latin /k/ before /e/ (e.g., *riconnoschere* < RECOGNOSCERE 'to recognise'), unlike most varieties.

In southern Italo-Romance dialects and Sardinian, we note the cacuminalisation of Latin LL > *dd* (e.g., *manteddu* < MANTELLU(M) 'coat'). Locally, Sicilian, Campanian and Lucanian show the tapping or rhotacism of D- (e.g., DE > *r(i)* 'of'). The assimilation of ND into *nn* (e.g., in QUANDO 'when') is shared by central varieties (Marchigiano, Laziale) and southern varieties (Sicilian, Lucanian, Campanian and Apulian). The voicing of NT into *nd* is shared by Marchigiano, Abruzzese, Campanian and Apulian, for instance in Marchigiano *tramondana* < TRANSMONTANU(M) 'north wind'. Other notable consonantal features include:

- the debuccalisation of F into a laryngeal [h] in Calabrian (e.g., *fhorta* 'strong'), encoded '1' like in Gascon (even though this is evidently a parallel innovation);
- the merger of some voiced/voiceless stop consonants, resulting in the emergence of forms such as *te* for DE 'of' in Salentine;
- the well-known *gorgia toscana*, that is, the lenition of voiceless stop consonants (Marotta, 2008) in Tuscan (e.g., *il suo calore* pronounced [il suo halore] 'its/her heat');

a back rhotic [ʁ] (‘r’ *francese*) across north-eastern Italy (Piedmont and Liguria, especially), albeit variably, encoded ‘1’ as in most of *Oïl* varieties and Portuguese;

the simplification of clusters ST, RT or LD into a voiceless lateral fricative (Contini, 1987) in Sassarese (e.g., FORTE(M) > *forthi* [foɬ:i] ‘strong’, CALIDU(M) > *caldhu* [kaɬ:u] ‘hot’).

As far as vowels are concerned, the innovation Ū > [y] is characteristic of Gallo-Italic varieties like Ligurian (e.g., *ùn* [yn] < UNU(M) ‘a’), Piedmontese and Lombard. The outcome of pretonic O is [u] in southern varieties such as Sicilian, Salentine and Calabrian (e.g., *tramuntana* ‘north wind’). In upper-southern varieties like Campanian and Molisan, unstressed vowels are strongly reduced: we thus have forms like *solë* ‘sun’ and *viendë* ‘wind’ — the latter showing diphtongisation of the stressed E.[4] Final unstressed vowels (except /a/) are most often dropped in Gallo-Italic varieties like Emilian-Romagnol where, in particular, the post-tonic vowel of infinitives of the 1st group has been elided (e.g., *SUFFLARE > *supièr* ‘to blow’). Finally, our corpus shows apheresis in the masculine singular indefinite article UNU(M) > *nu* ‘a’ in central and southern varieties.

In the domain of morphology, the system of clitic determiners and pronouns is subject to the so-called *Lex Porena* — loss of the lateral in weak contexts — affecting Ligurian and various southern Italo-Romance dialects: e.g., the definite article in Sicilian (masculine ILLU > *u*, feminine ILLA > *a* ‘the’). While (almost) all other Romance languages derive their definite article from ILLE/ILLA/ILLU, the one in Sardinian is original insofar as it reflects IPSE/IPSA/IPSU, yielding *su*/*sa* (masculine/feminine) in the singular and *sos*/*sas* in the plural. Another morphological originality of Sardinian lies in its plurals in *-s*, when all other Italo-Romance varieties have endings in *-i*/*-e* — the latter being considered as innovations, encoded 1 (see Table 5). As for possession, it is expressed in several varieties such as Molisan, Sabine, Abruzzese, in the form determiner + noun + possessive adjective (e.g., *la forza su(j)a* ‘his/her strength’).

In the morphosyntactic domain, northern dialects use the present perfect more than southern dialects like Sicilian and Salentine, which tend to prefer the preterite. Imperfects of the 3rd conjugation, in some dialects like Calabrian and Salentine, are in *-ìa* (e.g., *stringìa* ‘was squeezing’) and may be used for the conditional.[5] At a more syntactic level, subject doubling by a proclitic pronoun, at least in the 3rd person singular, is observed in all northern Italo-Romance dialects as well as in Friulian (Madriz & Roseano, 2006): e.g., *Il soreli, alore, al à tacât a scjaldâ* ‘the sun, then, he began to shine’).

Romansh shows a tendency to place the verb in the second position of the sentence (V2), under the influence of Germanic syntax (e.g., *Uossa ho il sulagl s-chudo* ‘now has the sun heated’). In a different area of grammar, southern dialects

[4] In some Apulian dialects, stress-retraction on /i/ and /e/ deletion caused the diachronic change Ĕ > [je]/|ie] > i: e.g., VĔNTU(M) > *viendë* > *vindë* ‘wind’.

[5] The Italian conditional differs from that of other Romance languages in that its endings go back to a cliticised form of the verb ‘have’ in the perfect tense, whereas in other Romance languages with a synthetic conditional, the endings go back to a cliticised form of the verb ‘have’ in the imperfect tense (Goyette, 2000).

such as Salentine and Calabrian tend to avoid infinitives and prefer subordinate clauses with complementisers, like the Balkan languages (Ledgeway, 2011). By contrast, infinitive periphrases are widespread in Sicilian: compare Salentine *riuscìa cu ffasçe* ‘managed that [he] would make’ with Sicilian *fussi arrinesciutu a fari* ‘would have managed to make’.

In the lexical domain, the nouns used to designate the traveller vary: *piligrin* in Bonifacian Ligurian, *cristianu* ‘person’ (literally ‘Christian’) in Salentine, Apulian, Lucanian, Campanian and Friulian. As for the word DIE(M) ‘day’, it gave way to the nominalised adjective DIURNU(M) > *giorno* or akin forms in many varieties; varieties preserving DIE(M) include Rhaeto-Romance and Gallo-Italic varieties (with masculine forms like *un dì*), as well as Sardinian (with feminine *una die* ‘one day’) and Apulian.

These observations, like the ones related to pronunciation and grammar, inspired our annotations, presented in the next subsection. Note, however, that we did not use all the features proposed by Rohlfs (1937) and Pellegrini (1973, 1977) to provide a classification of Italo-Romance dialects, because some traits — such as ‘today’ forms reflecting Latin HUNC HODIE — do not occur in our data.

2.3. Selection of innovations compared to Latin

In order to come up with a list of linguistic innovations compared to Latin, we exploited the double corpus that was recorded: on the one hand, the different versions, recorded and transcribed, of the fable “The North Wind and the Sun” (100–120 words per version, see Table 3); on the other hand, a supplementary list of isolated words (another 120 words) focusing particularly on fauna and flora.

As a corollary, we did not retain linguistic features related to the first or second persons of verbs, because there were too few of them in our texts — even though some speakers, who were free in their translations of the fable, chose to have the two protagonists, the north wind and the sun, dialogue. Certain morphological innovations were also discarded, because they are pan-Romance and would not help differentiate the dialects under investigation: e.g., the loss of synthetic forms of the passive, future, comparative and superlative (Coseriu, 1973; Goyette, 2000). The bottom-up approach we espoused, combining Aesop’s fable and the isolated word list, enabled us to narrow the scope of the work, amongst thousands of potential innovations. In return, this allowed us to rely on currently attested data, possibly confirmed by the speakers.

In total, more than 140 innovations were encoded, in a binary way, in the form of 0 (absence) or 1 (presence). For instance, if the apical /r/ of Latin has shifted to a dorsal phoneme ([ʁ]-like, as in French), at least word-initially (#_), the innovation is noted as 1. This example shows the importance of specifying the context in the encoding. For instance, the innovation /r/ > [ʁ] may not be categorical, or may be context-determined. It would be time-consuming to count the proportion of dorsal [ʁ]s — or nasal vowels, to take another example — as has been done in previous studies (Boula de Mareüil *et al*., 2013; Premat & Boula de Mareüil, 2018). Another example is the diphthongisation of stressed E and O, which may not be systematic. It was therefore important for the annotators (authors of this work PBM, AF and AR, who met regularly to find consensus) to agree on which observations to base their decisions.

For this purpose, we based our decisions on the most recurrent words of the fable — such as 'wind' (Spanish *viento*) or 'strong' (Spanish *fuerte*).

In summary, we encoded 145 innovations:
96 phonetic innovations,
like the regular sound change E > [wa];
27 morphological innovations,
like the merger of Latin imperfects in -ABA- and -EBA-;
8 syntactic innovations,
like the use of the present perfect as a narrative tense;
14 lexical innovations,
like the substitution of CUM 'with' with APUD HOQUE > *avec*.

The innovations cited here as examples are rather characteristic of northern Gallo-Romance dialects, but they are also attested in the linguistic Crescent. The substitution of CUM 'with' by APUD > *amb* is typical of the Occitano-Romance domain. In addition, the phonetic innovation E > [wa] is not observed regularly throughout the *Oïl* area; it tends to be avoided in Poitevin-Saintongeais, Norman and Lorrain.

Table 4 provides some examples taken from the list of isolated words complementing Aesop's fable. It illustrates three sound changes:
the change G > [h] (GALLU(M) > *jhàu* [hɑw] 'rooster') in Poitevin-Saintongeais,
the innovation E > [wa], in Picard, especially;
the vocalisation of the Latin diminutive -ELLU(M) in both these varieties.[6]

Table 4. Excerpt from the list of isolated words. Innovations are annotated (1).[7]

Sample of varieties			G > [h]	E > [wa]	Vocalisation of -ELLU(M)
Country	variety	Survey point	e.g., GALLU(M) 'rooster'	e.g., SERU/A(M) 'evening'	e.g., MARTELLU(M) 'hammer'
Belgium	pic	Dour	*co*	*swâr* (1)	*martiô* (1)
France	poi	Saint-Pardoux	*jhàu* (1)	*sér*	*martea* (1)
Italy	pie	Corio	*gal*	*sèira*	*martel*
Spain	ara	Uesca	*gallo*	(*veilada*)	*martiello*

Other innovations are linked to the La Spezia–Rimini line, a kind of natural barrier represented by the chain of the Apennines, which distinguishes Gallo-Italic varieties (north of it, for simplicity) from central and southern Italian varieties (Chambers & Trudgill, 2004). That isogloss represents, for varieties north of the line, the voicing, further weakening or even loss of voiceless consonants (mainly intervocalic stops). An example is Latin SECURU(M) > Ligurian *segûo*, Provençal *segur*, French *sûr* 'sure'.

Only in a few cases (affecting no more than three features) did we annotate 'not available' or 'not applicable' (NA), when the evidence we had was inconclusive with respect to a particular innovation. For instance, while the etymon SUFFLARE 'to blow' sometimes undergoes a sporadic (lexically specific) sound change F > [p] (e.g., Venetian *supiare*), that particular change cannot be assessed in neighbouring dialects where the equivalent is a non-cognate verb (e.g., Lombard *bufà* 'to blow'). Likewise, the innovation DEBITU(M) > **debutu* 'due' in Rhaeto- or Italo-Romance varieties cannot be assessed in those dialects where the deontic is rather expressed using a periphrase such as '(to) have to'. While such NA fields are well processed by historical glottometry (Kalyan & François, 2018: 77), this is however not the case of all the implementations of the algorithms we used (see § 2.4), which is why we made sure to keep them to a minimum in our dataset.

[6] The outcome of Latin -ELLU(M) was usually found in Aesop's fable, with reflexes of MANTELLU(M) 'coat'. When speakers from Belgium or France rendered it using a different word (e.g., *paltot* 'overcoat', *hardes* 'cloak'), we were able to assess this sound change based on reflexes of MARTELLU(M) 'hammer', present in the word list.

[7] The etymon of Aragonese *veilada* < VIGILATA(M) is not SĒR- 'evening', but in *veila* < VELU(M) 'sail', the stressed E did not diphthongise into [wa] either.

Table 5. Excerpt of the annotation table with examples of regular sound change (RSC), morphological and lexical innovations

Sample of varieties			RSC	Morph	Lex
Country	Variety	Survey point	R > [ʁ] / # _	Plurals in *-i*/ *-e*	ERA- replaced with STABA-
Belgium	wal	Liège	1	0	1
Italy	lig	Varazze	1	1	0
Portugal	por	Lisbon	1	0	0
Switzerland	rum	Zuoz	0	0	0

Table 5 displays an excerpt from the annotation table. Some innovations — such as the regular sound change (RSC) to a back [ʁ] — are subject to variation; in order to avoid ambiguities in the annotation, we specified the context — here, word-initial. The next feature in the table, which refers to the plural of nouns and adjectives in *-i*/*-e*, is considered an innovation compared to the plural accusative case of Latin in -S: this is what is observed in Italo-Romance varieties — except in Sardinian. The last innovation in Table 5 refers to the substitution of Latin imperfects in ERA- 'was/were' with STABA- (originally 'stood'), which is observed in almost all *Oïl* varieties (e.g., STABAT > Walloon *èsteût,* French *était* 'was').

2.4. Algorithms

It would be interesting to know if certain features are more relevant than others, whether due to their frequency in speech or their indexical role in speakers; if this could be assessed empirically, it could inspire methods for assigning different weights to these features. For example, the innovation Ū > [y], traditionally considered to distinguish Occitan from Catalan (Pai, 1949, among others), would receive a higher weight than another innovation characteristic of Catalan, namely the palatalisation of the initial L into [ʎ]. Due to the lack of criteria for assessing their relative importance, we did not attempt to weight innovations at this stage. The techniques we applied to disentangle the most important/discriminating attributes were thus based on structural features and not statistically established features. Several attribute selection algorithms were used, among which decision trees provide a readily readable representation, like single-access keys in the biological taxonomy.

The first approach we applied was historical glottometry, which crosses the Comparative method and the Wave model (François, 2014; François & Kalyan, f/c).

As sketched in the introduction, this quantitative method, inspired by dialectometry, focuses on diachrony without forcing a tree structure. The subgrouping it provides is based on shared innovations, which are indicative of stronger social relations: the more innovations are shared by a set of varieties, the greater their "subgroupiness". Our first analysis thus consisted in identifying the best-supported subgroups in our Romance sample, using a dedicated algorithm (Kalyan & François, 2018).

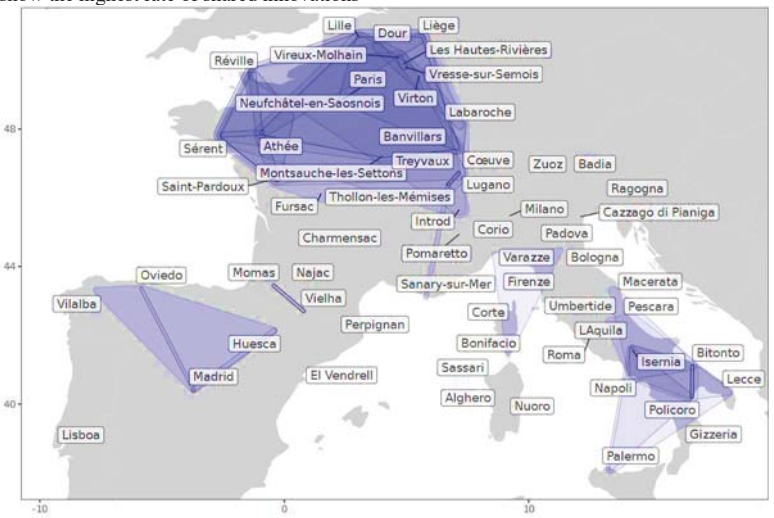
Agglomerative hierarchical clustering (also known as Ward's method) on the one hand, multidimensional scaling (MDS) and *t*-distributed stochastic neighbour embedding (*t*-SNE) on the other, provide different representations, in the form of a dendrogram or a projection in a two-dimensional plane, respectively. The former may be tuned with different thresholds to set the number of clusters desired. The latter techniques allow us to visualise which are the closest and furthest varieties, in the sense of a predefined distance — in our case, the Manhattan distance (also called city blocks). For these algorithms, we used the Python libraries Scikit-learn, SciPy and Plotly. Scikit-learn offers various attribute selection and classification algorithms, SciPy was used to process and visualise the hierarchical clustering, and Plotly allowed us to visualise the results in the form of variable-size points and choropleths (i.e., coloured surfaces, which may correspond to our dialectal areas). The base maps were designed in GeoJSON format and then vectorised as polygons representing the dialect areas.

Attribute selection is an important part of the work, as some machine learning algorithms, like Linear Discriminant Analysis (LDA), do not allow the number of features to be greater than the number of observation vectors (in our case, survey points). Two types of attribute selection algorithms exist and are proposed in the Scikit-learn library: model inspection by recursive feature elimination (RFE) and intrinsic feature importance analysis of tree-based classifiers. RFE is based on randomly altering the feature values, one at a time, and observing the resulting degradation of the model. However, it is problematic when the features are numerous and possibly correlated, as is the case here. Removing some of the colinear features does not degrade the model, which would suggest that they are not important, even though their overall impact may be. We chose the alternative approach, which is based on the Gini impurity index used within the decision tree classifier. It measures the probability of misclassifying a randomly chosen element if it were randomly labelled according to the distribution of labels in the subset. To improve the robustness of the results, we used Random Forest classifiers instead of single decision trees. These classifiers train several decision trees on random subsets of the data and average their predictions to reduce their tendency to overfit. The importance of a feature is also the average of the importance of the feature in each tree. To further improve the robustness of the results, we ran the algorithm 10,000 times and again averaged the feature importance for each classifier.

3. Results

The first results of the historical-glottometric approach took the form of a map, generated using François & Kalyan's (forthcoming) algorithm[8]. That map, shown in Figure 1, indicates the areas displaying the most consistent innovation patterns. The most prominent among these are: (1) the *Oïl* area, where the strongest subgroups include varieties represented by Picard, Lorrain, Champenois and Walloon, especially; (2) Ibero-Romance Spain; (3) southern Italy. While Occitan varieties also share a lot together, their similarity is mostly due to cases of shared retention; by contrast, the three areas shown in Figure 1 are characterised by intense rates of shared innovations. This outcome is interesting because it reveals innovation waves in three opposite directions from Rome.

Figure 1. A historical-glottometric map of the Romance domain, displaying which areas show the highest rate of shared innovations

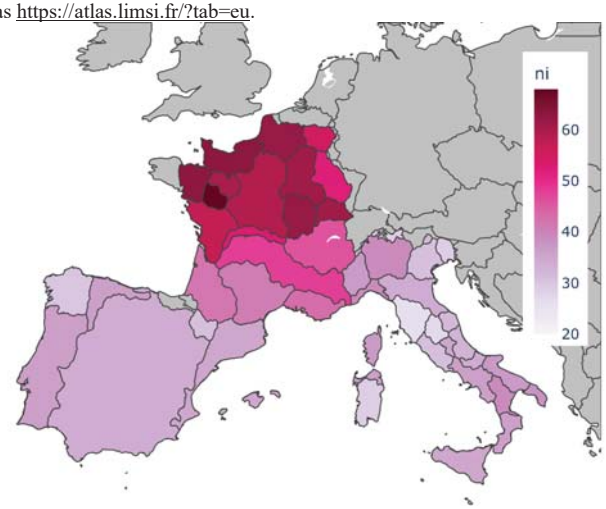


The varieties under study can also be ranked according to the number of innovations they reflect, and this number can be associated with a colour scale on a map. Figure 2 shows that "rate of innovativeness", and confirms that the *Oïl* area (northern Gallo-Romance) is the most innovative. This is particularly true of the Angevin dialect, in the north-west of the domain, with more than 65 innovations (including the palatalisation CL > [Cj], unknown in standard French). Within France, the region with the fewest innovations is Corsica, affiliated to the Tuscan group (Dalbera-Stefanaggi, 2002). Within Europe, Tuscan proper is the least innovative (i.e., the most conservative compared with Latin) according to our metric, with less than 30

[8] Link: https://tiny.cc/HGOA (accessed 10/11/2024).

innovations.[9] Ibero-Romance and southern Italo-Romance varieties rank between the extremes of Angevin and Tuscan, with around 35 innovations each.[10]

Figure 2. Linguistic map featuring the Romance areas that are the most innovative (dark red) vs. the least innovative (light pink). The contours of the linguistic areas are those of the speaking atlas https://atlas.limsi.fr/?tab=eu.



These scores concord partly with Pei (1949) who, on the basis of phonological changes to Latin stressed vowels, concluded that Sardinian and Italian are the most conservative, and French the most innovative. On the other hand, an author like Walter (1994: 119), on the basis of vocabulary, proposes that more peripheral languages, like Spanish and Portuguese, tend to be more conservative than more central languages like Italian and French: thus, to take just one example from our list of isolated words, Spanish *yegua* and Portuguese *égua* have preserved EQUA(M) 'mare', while Italian and French have innovated other etymons. Following the same principles as those presented by Cugno (2023), the present work provides a more precise picture and enables us to distinguish between Tuscan (which served as the prestige norm for the Italian language) and other Italian dialects. What is more, it allows us to distinguish between different linguistic domains.

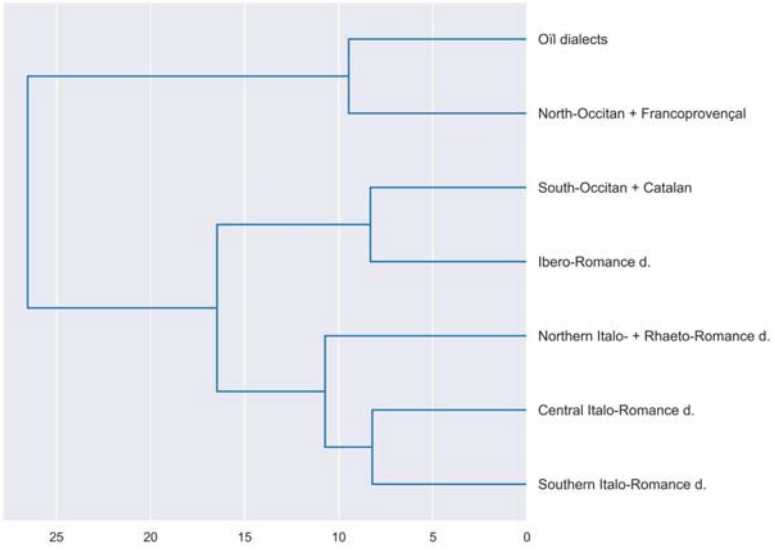
Our next computations involved techniques of hierarchical clustering, whether using all features, or only a subset of them, to guarantee the robustness and parsimony

[9] This recalls Cugno (2023: 203): "Tuscan is defined by the lack of common innovative traits from the other dialect areas and […] by a closer affinity with Latin".

[10] Note that these numbers are relative to the 145 innovations we identified in our dataset. Had we annotated more innovations, we would obtain higher numbers — but probably in the same proportions as our current observations. A full list of linguistic features on which the study is based is provided online at https://tiny.cc/Romance_innovations, thus making this study reproducible, interpretable and qualitatively evaluable by the historical linguist.

of the approach. The results provide heuristic answers to the questions raised in the introduction. In the simplified dendrogram of Figure 3 (reflecting the full set of innovations modelled), pruned with 7 leaves, cluster analysis corroborates a North/South division. The main divide runs through the middle of the Occitan area (southern Gallo-Romance), so that *Oïl* and intermediate dialects (Northern Occitan, Francoprovençal) cluster together in the North, the rest in the South. In the second branch of the dendrogram, the main division is between Southern Occitan, Catalan and Ibero-Romance varieties on the one hand, Rhaeto- and Italo-Romance varieties on the other. Within the latter group, the next split corresponds to the La Spezia–Rimini line: in the north, Rhaeto-Romance together with northern Italo-Romance varieties; in the south, central Italo-Romance (+ Sardinian), and southern Italo-Romance.

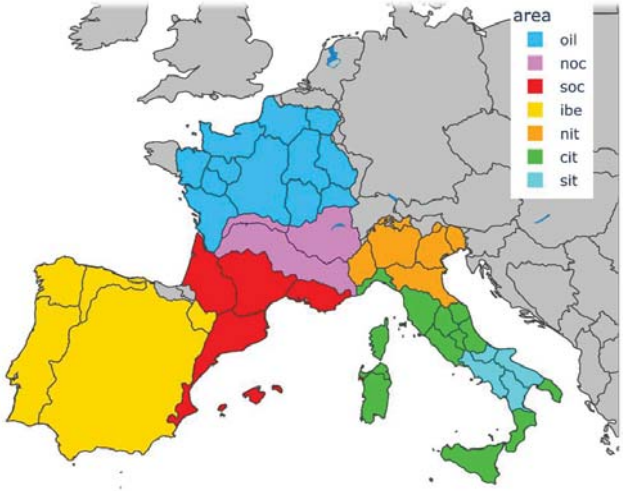
Figure 3. Simplified dendrogram resulting from a cluster analysis with 7 classes



The attachment of Northern Occitan to *Oïl* varieties rather than to Southern Occitan may seem counter-intuitive; it would displease both activists of Occitan (defenders of a single language, "one and plural") and activists of *Oïl* varieties, for whom Gallo, Picard and other varieties are distinct languages from French. Yet, it is undeniable that northern Occitan does share quite a few innovations with the Linguistic Crescent, Francoprovençal and (most) *Oïl* varieties, as exemplified by Table 3: for instance, the palatalisation of Latin CA and the dropping of consonants that had become final in Gallo-Romance, in particular in past participles (e.g., Auvergnat *eschaufat* [ejtsawfa] < EXCALFATU(M) 'heated'). As for southern Occitan, in particular Languedocian, it has developed other innovations: for instance, V > [b] (betacism) and CT > [tʃ] (e.g., *fach* < FACTU(M) 'done/fact'). Many (though not all) of these innovations are shared with Catalan.

We then associated different colours with the 7 classes resulting from the dendrogram of Figure 3 and projected them onto the map of Figure 4. Interestingly, languages like Spanish and Portuguese appear closer to each other than, say, Milanese Lombard and Neapolitan Campanian — which, possibly, are less mutually intelligible. Under the green colour, Ligurian (Gallo-Italic group), Abruzzese, Sardinian, Sicilian, Calabrian and Salentine are grouped with central Italo-Romance dialects — together with Gallurian-Sassarese, Corsican and Tuscan. In our data, Ligurian shows greater divergence from its Lombard neighbour than Lombard does from Piedmontese and Emilian-Romagnol (orange cluster). For example, Ligurian has maintained post-tonic vowels more than its neighbours have: compare Ligurian *forte* < FORTE(M) 'strong' with Piedmontese, Lombard or Emilian-Romagnol *fòrt*/*fort*.[11] Conversely, Venetian is grouped with Gallo-italic varieties as in some previous studies (Cugno, 2023: 205).

Figure 4. Map corresponding to the 7-cluster dendrogram



In the south of Italy, we have to distinguish between upper-southern dialects, grouped under the label "southern" in Pellegrini's (1977) classification (Apulian, Lucanian, Campanian, Molisan), and extreme-southern dialects, related to Sicilian by Pellegrini, based on data from Parlangèli's *Carta dei Dialetti Italiani* (Sicilian, Calabrian, Salentine). In upper-southern Italo-Romance dialects (the cyan area in Figure 4), particularly in Apulian, the "unstressed vowel system is strongly reduced" (Loporcaro, 1997b: 341): the usual outcome is a schwa-like vowel, transcribed <ë> in the atlas, which can be deleted under certain conditions (Boula de Mareüil *et al*., 2021): compare *fort(ë)* 'strong' with Sicilian *forti*, Calabrian *fhorta* and Salentine *forte*. This

[11] In terms of rhythmic metrics, the Ligurian samples also depart from the Piedmontese samples analysed with the methods used by Romano *et al*. (2010), falling into the area where syllable-timed languages are usually located.

deletion is associated with considerable shifts in vowel quality and lengthening phenomena, which frequently cause vowel diphthongisation or instability (Avolio, 1995; Romano, 2013). With regard to consonants, the voicing of postnasal voiceless plosives (e.g., NT > *nd*) and the assimilation of voiced stops in the same position clearly emerge as a shared feature, which is one of the most iconic upper-southern characteristics (Avolio, 1995; Pellegrini, 1977). Let us note *quannë* < QUANDO 'when', among forms with an original ND, and *tramëndanë* (Apulian *tramëndeunë*) < TRASMONTANU(M) 'north wind', among forms that originally had NT. Based on these innovations (and others), both historical glottometry and hierarchical clustering grouped upper-southern Italian dialects together. Extreme-southern Italian dialects, which are also closely related but share other innovations, have been clustered with central Italian dialects.

Sardinian is a special case that deserves particular attention. This conservative language is the only one in the Italo-Romance domain to have kept the *-t* ending of 3rd person of verbal forms (e.g., *incuminzat* < *INCOMINITIAT 'begins') and to have (pro)noun plurals in *-s* (Pei, 1949; Walter, 1994: 174–175; Goyette, 2000). It is also the only Italo-Romance language to have developed definite articles in *sa*/*su* < IPSE/IPSA/IPSU rather than ILLE/ILLA/ILLU. In our annotation, the distance between Sardinian (Logudorese) and central Italian varieties may be greater than the one between central and southern Italian varieties. In the 61-leaf dendrogram, the Sardinian branch splits from another branch which groups 13 "central" Italian varieties. This is challenging to represent in the form of a tree, illustrating the difficulty of classifying Sardinian amongst Italo-Romance languages (Contini, 1987; Adams, 2007: 576).

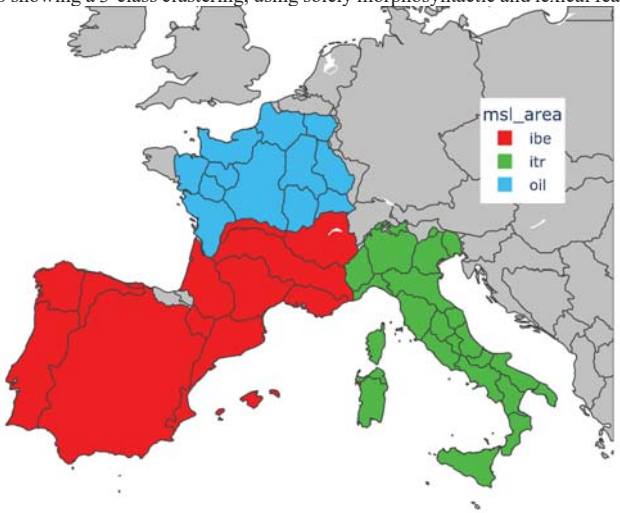
Figure 5. Plane resulting from the t-SNE algorithm (see Tables 1 and 2 for label abbreviations, as well as Figure 4 for the point colours)

A *t*-SNE modelling, shown in Figure 5, is better suited for this situation. The *t*-SNE algorithm isolates Sardinian at the bottom right of the plane, whereas most of the Italo-Romance dialects are grouped at the top right, with two rather compact subgroups corresponding to northern Italian varieties (here including Ligurian) and upper-southern Italian dialects. Romansh is also isolated, its closest neighbour being Friulian — another Rhaeto-Romance variety. Other groupings, to the left and bottom centre of Figure 5, correspond to Gallo- and Ibero-Romance, respectively. Northern and southern Gallo-Romance are well separated; yet, no subgroups for northern Gallo-Romance varieties were found, with western *Oïl* on one side and eastern *Oïl* on the other side: this corroborates their intermingled character, as noted above. The *t*-SNE algorithm globally gives an image less faithful to reality than MDS does; but locally, it highlights the specificity of certain points, which is particularly relevant for Sardinian — and Rhaeto-Romance varieties.

Finally, we sought to identify the most important/discriminant features that underlie the 7-class clustering reported in Figure 4, using several algorithms such as Random Forests. Attribute selection (see § 2.4) is not an easy task, due to the collinearity of many features in our data. Among the most important features are the palatalisation of Latin CA, which characterises the majority of northern Gallo-Romance dialects, and the simplification of geminates north of the La Spezia–Rimini line. Most innovations relate to phonetic/phonological traits, but we also have morphosyntactic and lexical features, such as non-null subjects in northern Gallo-Romance varieties, and the substitution of CUM 'with' by APUD > *amb* in Occitan varieties. Such features are often neglected in dialectometrical visualisations (Oliviéri, 2015). The 20 or 30 most important features[12] (e.g., the drop of final vowels other than A, past participles of verbs of the 1st group in /E/) give similar results in terms of dendrograms and maps, very close to the map of Figure 6.

[12] Their list is not provided due to space limitations and because, depending on random seeds, it varies slightly.

Figure 6. Map showing a 3-class clustering, using solely morphosyntactic and lexical features



If we retain only morphosyntactic innovations (which are 35 in our dataset) and use the same distance threshold as the one employed to generate Figures 3–4, we identify a North vs. South-East- vs. South-West tripartition. This holds true when lexical and morphosyntactic features are included, yielding Figure 6. Just like previously, the main border runs through the middle of France, except this time, what we see in Figure 6, is a large cluster that includes northern and southern Occitan (including the Crescent), Francoprovençal, Catalan, and Ibero-Romance. This comes in contrast with Figures 3 and 4, which were based on all the features, predominantly phonetic innovations. Even if non-null subject, double negation and periphrastic past tense have been encoded in our list of innovations, we acknowledge that others have not been annotated: direct object marking, clitic climbing, binary auxiliary selection based on transitivity/person (Ledgeway, 2022). Without denying the importance of such syntactic features, they are only a few among more than a hundred innovations addressed here, which will definitely have to be taken into consideration in future work. This is both a limit and an advantage of the methodology based on a closed corpus: we risk overlooking certain phenomena; but at least, those which are represented in our samples can be compared systematically across all varieties.

4. Conclusion

This study highlighted the diversity of minority languages/dialects across the Romance space, which still need to be documented. Despite the sheer size of the domain, the fieldwork collection of fine-grained data — even from a small corpus — has proven useful to accurately uncover the distribution of Romance varieties. The tree

representation has been advantageously supplemented by geolinguistic maps and other modelling, to bridge the gap between historical and computational linguistics. Therefore, the main contribution of this paper is not so much towards our understanding of phylogenetic similarity between Romance varieties and their distance from Latin, but rather demonstrating how this variation could be encoded and mapped, so as to integrate the principles and laws of the Neogrammarians' long-established Comparative method.

The use of computational methods of supervised or unsupervised learning offers grounding as well as new insight into classifications of Romance varieties. Results confirm the robustness of a North/South divide — with the *Oïl* area being, by far, the most innovative — and, secondarily, an opposition between the South-West (mainly Ibero-Romance) and the South-East (mainly Italo-Romance, more conservative); Occitano-Romance varieties occupy an intermediate position. The resulting maps are reminiscent of dialectometric maps such as those on the *dialektkarten.ch* website (Goebl, 2003; Scherrer, 2021). To some extent, this is reassuring, knowing that the latter maps are based on data that dates back more than a century. Edit distance-based approaches, however, do not account for linguistic changes, at least sound changes: they are agnostic of which side of an isogloss is innovative. For example, the palatalisation /ka/ > [ʃa] is a common innovation, which we encoded as such in our annotation scheme. An approach based on edit-distances would have grouped [ʃa] vs. [ka] varieties symmetrically, whereas our approach, based on the Comparative method principles, recognises the [ʃa] grouping as the only one with phylogenetic relevance — since it reflects shared innovation rather than shared retention. We have also manipulated morphosyntactic features, which are more difficult to address, such as oscillations between auxiliaries, reflexes of HABERE 'to have', TENERE 'to hold' and ESSE 'to be'. Interestingly, a well-known phenomenon such as the substitution in Portuguese of HABERE by TENERE > *ter* 'to have' was found in southern Italian dialects as well. Such innovations are marginally represented in traditional taxonomy.

Some of the annotations we have encoded can be debated. Likewise, any classification is somewhat arbitrary (Sumien, 2008) and since Ascoli (1877), most proposals have been controversial. In this respect, the great phylogenetic similarity between Spanish and Portuguese varieties may seem unexpected or surprising, while our results correspond for the most part to traditional intuition on the evolution of Romance languages. Another outcome of our empirical work is to have evidenced the fragmentation of southern Italo-Romance dialects, often underestimated by linguists (see Cugno, 2023). We do not claim that the splits shown in Figure 3 should be interpreted literally as successive stages in the diversification of Latin — as though, say, Gallo-Romance had "separated" first: this would be a simplistic reading of our results. Indeed, it is well established that the first branches of a Romance tree, interpreted chronologically, should be Sardinian and Lucanian, corresponding to the first regions conquered by the Romans (Lausberg, 1939; Goldstein, 2023). If we wanted to pursue our quantitative approach with machine-learning techniques, however, divergence-time estimation would require dating a hundred innovations over the long term.

The innovations we identified do not necessarily include the isoglosses that are perceived, in folk linguistics (Preston, 2005), as being the most important (such as the Benrath line *maken-machen* in the Germanic area). An innovation that affects the whole lexicon, such as the transition from the plural in -S to plurals in -*i*/-*e*, undoubtedly has greater impact upon speakers' spontaneous perceptions than, say, an irregular sound change or a lexical innovation affecting only one word; however, they all constitute different innovations which, in spite of perception, deserve equally to be incorporated into our results. Weighting or ranking innovations with respect to Latin would go beyond the scope of this article, and would require large amounts of dialect data. Likewise, the present work did not attempt to take into account the many detailed (mostly phonetic) isoglosses put forward — but not quantified — by Pellegrini (1973) for Italy. To our defence, preliminary experiments with 50 or 60 randomly selected innovations already give consistent results.

We are currently considering semi-automated ways to extend the present approach, in the future, to a greater number of survey points, including Istria and Romania. In all cases, the results will have to be compared with those provided by edit distances and acoustic distances: this is another research avenue which would have geolocation applications (Goldman *et al*., 2018).

Acknowledgements

This work is part of the project "Computational Language Documentation by 2025" (CLD2025) funded by the French National Research Agency (ANR) and the German Research Foundation (DFG). We are grateful to Frédéric Vernier for his help with the cartography, to Cédric Jr Tonga for the data processing, and to Siva Kalyan for his advice on glottometric results. We would also like to thank Lorenza Brasile, Lucia Molinu, Michela Russo, Paolo Roseano and Philippe Maurer for their advice regarding some Italo- and Rhaeto-Romance varieties. Finally, we express our gratitude to two anonymous reviewers of an earlier version of this article for their insightful comments.

References

Abalain, Hervé. 2007. *Le français et les langues historiques de la France*. Paris: Éditions Jean-Paul Gisserot.

Adams, James Noel. 2007. *The regional diversification of Latin 200 BC–AD 600*. Cambridge: Cambridge University Press.

Ascoli, Graziadio Isaia. 1877. Schizzi franco-provenzali. *Archivio glottologico italiano* 2: 61–120.

Ascoli, Graziadio Isaia. 1882/1885. L'Italia dialettale. *Archivio glottologico italiano* 8: 98–128.

Avolio, Francesco. 1995. *Bommèsprë. Profilo linguistico dell'Italia centro-meridionale*. San Severo: Gerni.

Bartoli, Matteo G., Pellis, Ugo, Massobrio, Lorenzo. 1995. *Atlante Linguistico Italiano*. Rome: Istituto Poligrafico e Zecca dello Stato.

Bec, Pierre. 1995. *La langue occitane*. Paris: Presses Universitaires de France.

Beijering, Karin, Gooskens, Charlotte, Heeringa, Wilbert. 2008. Predicting intelligibility and perceived linguistic distance by means of the Levenshtein algorithm. *Linguistics in the Netherlands* 25: 13–24.

Blanchet, Philippe. 1996. Problématique de la situation ethnolinguistique du pays de Retz (L.-Atl.) : pratiques linguistiques et identité en zone de marches. *Cahiers de socio-linguistique* 1: 45–80.

Boula de Mareüil, Philippe, Rilliard, Albert, Vernier, Frédéric. 2018. A speaking atlas of the regional languages of France. *Proceedings of the 11th International Conference on Language Resources and Evaluation*, 4133–4138. Miyazaki.

Boula de Mareüil, Philippe, Bilinski, Éric, De Iacovo, Valentina, Romano, Antonio & Vernier, Frédéric. 2021. For a mapping of the languages/dialects of Italy and regional varieties of Italian. In A. Thibault, M. Avanzi, N. Lo Vecchio & A. Millour (eds), *New Ways of Analyzing Dialectal Variation*, 267–288. Strasbourg: Éditions de linguistique et de philologie.

Boutin, Béatrice Akissi. 2003. La norme endogène du français de Côte d'Ivoire : mise en évidence de règles différentes du français de France concernant la complémentation verbale. *Sud Langues* 3: 33–46.

Brun-Trigaud, Guylaine. 2020. Études dialectométriques sur le lexique des parlers du Croissant. In L. Esher, M. Guérin, N. Quint & M. Russo (eds), *Le Croissant linguistique entre oc, oïl et francoprovençal : des mots à la grammaire, des parlers aux aires*, 173–181. Paris: L'Harmattan.

Câmara Jr., Joaquim Mattoso. 1970. *Estrutura da língua portuguesa*. Petrópolis: Vozes.

Chambers, Jack K. & Trudgill, Peter. 2004. *Dialectology*. Cambridge: Cambridge University Press.

Chambon, Jean-Pierre & Greub, Yan. 2002. Note sur l'âge du (proto)gascon. *Revue de linguistique romane* 66: 263–264.

Chevrier, Jean-Jacques & Gautier, Michel. 2002. *Le poitevin-saintongeais : langue d'oïl méridionale*. La Crèche : Gestes Éditions.

Contini, Michel. 1987. *Étude de géographie phonétique et de phonétique instrumentale du sarde*. Alessandria: Edizioni dell'Orso.

Coquebert de Montbret, Charles. 1831. Essai d'un travail sur la géographie de la langue française. In S. Bottin (ed.), *Mélanges sur les langues, dialectes et patois: renfermant, entre autres, une collection de versions de la parabole de l'enfant prodigue en cent idioms ou patois différens [sic], presque tous de France ; précédés d'un essai d'un travail sur la géographie de la langue française*, 5–29. Paris: Bureau de l'Almanach du commerce.

Coseriu, Eugen. 1973. *Sincronía, diacronía e historia. El problema del cambio lingüístico.* Madrid: Gredos.

Cugno, Federica. 2023. Italian dialect classifications. *Dialectologia* 10: 197–230.

Dalbera-Stefanaggi, Marie-José. 2002. *La langue corse*. Paris: Presses Universitaires de France.

François, Alexandre. 2014. Trees, waves and linkages: Models of language diversification. In C. Bowern & B. Evans (eds), *The Routledge Handbook of Historical Linguistics*. 161–189. London: Routledge.

François, Alexandre & Kalyan, Siva. Forthcoming. Subgrouping: Trees vs. Waves. In C. Bowern & B. Evans (eds), *The Routledge Handbook of Historical Linguistics* (Second edition). New York: Routledge.

Gaillard-Corvaglia, Antonella, Léonard, Jean Léo & Darlu, Pierre. 2007. Testing Cladistics on Dialect Networks and Phyla (Gallo-Romance Vowels, Southern Italo-Romance Diasystems and Mayan Languages. *Proceedings of the 9th Meeting of the ACL SIG in Computational Morphology and Phonology*, 23–30. Prague.

Garassino, Davide & Filipponio, Lorenzo. 2021. The impact of information structure on the phonetic implementation of vowel length. A contrastive analysis of two Ligurian Dialects. In A. Teixeira Kalkhoff, M. Selig & C. Mooshammer (eds), *Prosody and conceptional variation*. 213–236. Bern: Peter Lang.

García Mouton, Pilar, Fernández-Ordóñez, Inés, Heap, David, Perea, María Pilar, Saramago, João & Sousa, Xulio. 2016. ALPI-CSIC [www.alpi.csic.es], digital edition of Navarro Tomás, in T. Navarro Tomás (dir.), *Atlas Lingüístico de la Península Ibérica*. Madrid: CSIC.

Gilliéron, Jules & Edmont, Edmond. 1902–1910. *Atlas linguistique de la France*. Paris: Champion.

Goebl, Hans. 2002. Analyse dialectométrique des structures de profondeur de l'ALF. *Revue de linguistique romane* 66(261–262): 5–63.

Goebl, Hans. 2003. Regards dialectométriques sur les données de l'Atlas linguistique de la France (ALF) : relations quantitatives et structures de profondeur. *Estudis Romànics* 25: 59–121.

Goldman, Jean-Philippe. Scherrer, Yves., Glikman, Julie., Avanzi, Mathieu, Benzitoun, Christophe & Boula de Mareüil, Philippe. 2018. Crowdsourcing regional variation data and automatic geolocalisation of speakers of European French. *Proceedings of the 11th International Conference on Language Resources and Evaluation,* 3336–3342. Miyazaki.

Goldstein, David M. 2023. Divergence-time estimation in Indo-European: The case of Latin. *Diachronica* 40(3): 1–53.

Gooskens, Charlotte. 2005a. How well can Norwegians identify their dialects? *Nordic Journal of Linguistics* 28(1): 37–60.

Gooskens, Charlotte. 2005b. Travel time as a predictor of linguistic distance. *Dialectologia et Geolinguistica* 13: 38–62.

Goyette, Stéphane. 2000. *The emergence of the Romance languages from Latin: a case for creolization effects*. Ph.D thesis. Ottawa: University of Ottawa.

Guérin, Maximilien, Esher, Louise, Léonard, Jean Léo, Loiseau, Sylvain. 2021. Comparing inflectional structure across a dialect group: verb morphology in the "croissant linguistique". *Revue roumaine de linguistique* 66: 299–316

Haust, Jean, Remacle, Louis, Legros, Élise, Lechanteur, Jean, Boutier, Marie-Guy & Baiwir, Esther. 1953–2011. *Atlas linguistique de la Wallonie*. Liège: Vaillant-Carmanne.

Heeringa, Wilbert. 2004. Measuring dialect pronunciation differences using Levenshtein distance. PhD thesis. Groningen: Rijksuniversiteit.

Herrero de Haro, Alfredo & Hajek, John. 2020. Eastern Andalusian Spanish. *Journal of the International Phonetic Association* 52(1): 135–156.

IPA [= International Phonetic Association]. 1999. Handbook of the International Phonetic Association. A Guide to the Use of the International Phonetic Alphabet. Cambridge: Cambridge University Press.

Jaberg, Karl & Jud, Jacob. 1928–1940. *Sprach- und Sachatlas Italiens und der Südschweiz*. Ringier: Zofingen.

Joret, Charles. 1881. *Essai sur le patois normand du Bessin: suivi d'un dictionnaire étymologique*. Paris: Friedrich Vieweg.

Kalyan, Siva & François, Alexandre. 2018. Freeing the Comparative Method from the tree model: A framework for Historical Glottometry. *Senri Ethnological Studies* 98: 59–89.

Kihm, Alain. 2005. Les langues créoles. In J.-M. Hombert (ed.), *Aux origines des langues et du langage,* 390–427. Paris: Fayard.

Knyazeva, Elena, Boula de Mareüil, Philippe & Vernier, Frédéric (2022), Aesop's fable "The North Wind and the Sun" used as a Rosetta Stone to extract and map spoken words in under-resourced languages. *Proceedings of the 13th International Conference on Language Resources and Evaluation,* 2072–2079. Marseilles.

Lausberg, Heinrich. 1939. *Die Mundarten Südlukaniens*. Niemeyer: Halle (Saale).

Ledgeway, Adam. 2011. La sopravvivenza del sistema dei doppi complementatori nei dialetti meridionali. In P. Del Puente, (ed.), *Dialetti: per parlarne e parlare*. 239–262. Rionero in Vulture: Caliceditori.

Ledgeway, Adam. 2022. Residues and Extensions of Perfective Auxiliary BE: Modal Conditioning. *Languages* 7(3): 160.

Legeard, Fernand. 2020. *Et si on caôsait patois ! Histouéres d'mon Pays mainiot*. Éditions de Igé: l'Étrave.

Leinonen, Therese, Çöltekin, Çağrı & Nerbonne, John. 2015. Using Gabmap. *Lingua* 178: 71–83.

Léonard, Jean Léo, Brun-Trigaud, Guylaine & Picard, Flore. 2024. Atlas linguistiques et perspectives dialectométriques. In L. Esher & J. Sibille (eds), *Manuel de linguistique occitane*, 473–520. Mouton de Gruyter: Berlin.

LIPSKI, JOHN. 1996. *EL ESPAÑOL DE AMERICA*. MADRID: CATEDRA.

Loporcaro, Michele. 1997a. Lengthening and Raddoppiamento Fonosintattico. In M. Maiden, & M. Parry (eds), *The Dialects of Italy*, 41–51. London–New York: Routledge.

Loporcaro, Michele. 1997b. Puglia & Salento. In M. Maiden, & M. Parry (eds), *The Dialects of Italy*, 338–348. London: Routledge.

Madriz, Anna & Roseano, Paolo. 2006. *Scrivere in friulano*. Udine: Società Filologica Friulana.

Martin, Jean-Baptiste. 2011. Le francoprovençal. *Langues et cité* 18: 1–12.

Marotta, Giovanna. 2008. Lenition in Tuscan Italian (gorgia toscana). In J. Brandão de Carvalho, T. Scheer & P. Ségéral (eds), *Lenition and Fortition*, 235–270. Berlin: Mouton de Gruyter.

Nagore Laín, Francho. 1989. *Gramática de la lengua aragonesa*. Zaragoza: Mira Editores.

Nerbonne, John, Kleiweg, Peter, Heeringa, Wilbert & Manni, Franz. 2007. "Projecting dialect differences to geography: bootstrap clustering vs. noisy clustering. In C. Preisach, L. Schmidt-Thieme, H. Burkhardt & R. Decker (eds), *Data analysis, machine learning, and applications,* 647–654. Berlin: Springer.

Oliviéri, Michèle. 2015. Le statut des pronoms sujets dans les dialectes du nord de l'Occitanies. In S. Retali-Medori (ed.), *Paroddi varghji, mélanges offerts à Marie-José Dalbera-Stefanaggi*, 289–302. Alessandria: Edizioni dell'Orso.

Patriarca, Marco, Heinsalu, Els, Léonard, Jean Léo. 2020. *Language in Space and Time*. Cambridge: Cambridge University Press.

Pei, Mario A. 1949. A New Methodology for Romance Classification. *Word* 5(2): 135–146.

Pellegrini, Giovan Battista. 1973. I cinque sistemi dell'italo-romanzo. *Revue roumaine de linguistique* 18: 105–129.

Pellegrini, Giovan Battista. 1977. *Carta dei dialetti d'Italia*. Pisa: Pacini.

Ploog, Katja. 2002. *Le français à Abidjan. Pour une approche syntaxique du non-standard*. Paris: Éditions du CNRS.

Poplineau, Bernard. 2006. Méthode d'apprentissage du champenois. *Lou Champaignat* 28: 5–31.

Premat, Timothée & Boula de Mareüil, Philippe. 2018. Le /R/ « roulé » en français et dans quelques langues régionales de France. *Actes des 32es Journées d'Études sur la Parole*, 55–63. Aix-en-Provence.

Preston, Denis R. 2005. What is folk linguistics? Why should you care? *Lingua Posnaniensis* 47: 143–162.

Quint, Nicolas. 2023. Les parlers du Croissant : un aperçu des actions actuelles de documentation et de promotion d'un patrimoine linguistique menacé. In A. Rialland & M. Russo (eds), *Les langues régionales de France. Nouvelles approches, nouvelles méthodologie, revitalisation*, 213–245. Paris: Éditions de la Société de Linguistique de Paris.

Regueira, Xosé Luís. 1996. Galician. *Journal of the International Phonetic Association* 26(2): 119–122.

Rohlfs, Gerhard. 1937. *La struttura linguistica dell'Italia.* Leipzig: H. Keller.

Romano, Antonio, Mairano, Paolo & Pollifrone, Barbara. 2010. Variabilità ritmica di varietà dialettali del Piemonte. In S. Schmid, M. Schwarzenbach & D. Studer (eds), *La dimensione temporale del parlato*, 101–112.Torriana: EDK.

Romano, Antonio. 2013. Osservazione e valutazione di traiettorie vocaliche su diagrammi formantici per descrivere il polimorfismo e la dittongazione dialetti pugliesi. In F. Sánchez Miret, & D. Recasens, (eds), *Experimental Phonetics and Sound Change (with special reference to the Romance languages)*, 121–143. Munich: LINCOM.

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

Romano, Antonio. 2020. Vowel reduction and deletion in Apulian and Lucanian dialects with reference to speech rhythm. *Italian Journal of Linguistics* 32(1): 85–102.

Ronjat, Jules. 1913. *Essai de syntaxe des parlers provençaux modernes*. Paris: Imprimerie nationale.

Ronjat, Jules. 1930. *Grammaire istorique* [sic] *des parlers provençaux modernes*. Macon: Protat.

Scherrer, Yves. 2021. Les cartes dialectométriques interactives de dialektkarten.ch. In A. Thibault, M. Avanzi, N. Lo Vecchio & A. Millour (eds), *New Ways of Analyzing Dialectal Variation*, 137–152. Strasbourg: Éditions de linguistique et de philologie.

Saussure, Ferdinand de. 1916. *Cours de linguistique générale*. Paris: Payot.

Schleicher, August. 1861[2010]. *Compendium der vergleichenden Grammatik der indogermanischen Sprachen*. Cambridge: Cambridge University Press.

Schmidt, Johannes. 1872. *Die Verwandtschaftsverhältnisse der indogermanischen Sprachen*. Weimar: Hermann Böhlau.

Séguy, Jean. 1973. La dialectométrie dans l'Atlas linguistique de la Gascogne. *Revue de linguistique romane* 37: 1–24.

Sumien, Domergue. 2007. Preconizacions del Conselh de la Lenga Occitana. *Lingüistica Occitana* 6: 1–157.

Sumien, Domergue. 2008. Classificacion dei dialèctes occitans. *Lingüistica Occitana* 7: 1–55.

Tillinger, Gabor. 2016. Étude des frontières linguistiques à l'intérieur de la zone d'interférence appelée "Croissant". *Atti del XXVIII Congresso internazionale di linguistica e filologia romanza*, 1037–1052. Rome.

Walter, Henriette. 1994. *L'aventure des langues en Occident*. Paris: Éditions Robert Laffont.

Wartburg, Walther von. 1922–2002. *Französisches Etymologisches Wörterbuch. Eine darstellung des galloromanischen sprachschatzes*. Bonn/Leipzig/Bâle: Teubner/ Klopp/Zbinden.

Zink, Gaston. 1989. *Morphologie du français médiéval*. Paris: Presses Universitaires de France.