A Database of Romance Clitics With Speech Samples

Abed Qaddoumi, Francisco Ordóñez, Lori Repetti, Owen Rambow
Linguistics and IACS, Stony Brook University, New York, NY, USA

Abstract

We present a new database of Romance clitics across nine varieties. The database includes speech, transcriptions, and linguistic annotations. The database concentrates on clitics, and includes varieties of Romance with stressed clitics. Specifically, the database includes data from the regions of Corsica, Pyrénées-Atlantiques, Sardinia, Liguria, Basilicata, Campania, Mallorca, Menorca, and Formentera. A publicly accessible interface allows easy searching. The database and, separately, the interface code will be made publicly available.

Keywords: Language resources for Romance, speech corpora, dialectal corpora

1. Introduction: Variation in Romance Clitic Syntax and Morphology

Romance languages are the group of languages that descend from Vulgar Latin, and include French, Italian, Portuguese, Romanian, Spanish, and Catalan, the official languages of France, Italy, Portugal, Romania, and Spain respectively. However, there are also many smaller, less-known varieties spoken throughout Mediterranean Europe.

This paper introduces the Romance Clitics Database (RoCDat), which focuses on clitics in these lesser-known varieties. Clitics are morphemes that require a phonological host. An interesting property of clitics in most Romance languages is that they are stressless, meaning they do not carry lexical stress or prosodic prominence. However, in various minority Romance varieties spoken in the South of France, the North and South of Italy, the Balearic Islands, Corsica, and Sardinia, stress or stress shift can occur through interaction with the host word. The RoCDat focuses on stressed clitics but also includes phonological, morphological, and syntactic information, as well as sociological metadata. We provide an example to illustrate the issue of stress. Stress is marked with an acute accent on the stressed vowel.

(1) a. Gloss: 'Give-to me-it'

b. [da-mm-íllu] <dammillu>[1] Basilicata, Italy

c. [dá-mme-li] <dammeli> Standard Italian

There are two principal reasons why a database of Romance clitics is important: language documentation, and low-resource NLP.

Many Romance varieties are endangered and understudied, and are being supplanted by the standard national languages (Cerruti and Regis, 2014; Repetti, 2018). This is particularly true of varieties with clitics that interact with stress, which is a rare phenomenon, attested in some Romance varieties such as Sardinian and Gascon, among others (Ordóñez and Repetti, 2006). In addition to the simple goal of documenting linguistic varieties, this kind of variation is empirically important for linguistic theory because it highlights the interaction between morphology, syntax, and prosody (Ordóñez and Repetti, 2006; Lai, 2017).

Concerning NLP, we note that these interactions change how the language sounds Levis (2018) and is written in informal communication (Ramponi and Casula, 2023). Prior work shows that small, targeted linguistic resources can yield substantial improvements. Zheng et al. (2024) improve low-resource machine translation for Formosan languages by incorporating bilingual lexical resources during training, demonstrating that lexicons can be valuable when large corpora are unavailable. Erdmann and Habash (2018) shows that combining a small set of out-of-context morphological rules with embedding-based methods improves low-resource morphological modeling. Thus, building databases for such low-resource languages is important for documenting their features and providing data for future applications.

Our contributions are as follows:

1. We present a freely available database of Romance clitic data (spoken, transcribed, and annotated). The data is available via an interface and is downloadable.

2. We describe the database and interface software, which is available for other such projects.

The paper proceeds as follows: Section 2 describes RoCDat's place relative to prior resources available. Section 3 details the data sources and elicitation methods, and Section 4 details data statistics and the annotation schema. We describe the database in Section 5 and its statistics in Section 6. Section 7 presents the website for RoC-

[1] This is an approximation because there is no standardized orthography for the Basilicata variety.

Proceedings of SIGUL 2026 Joint Workshop with ELE, EURALI, and DCLRL @ LREC 2026, pages 83–93
11–12 May 2026.

Dat and a minimal use case, while Section 8 discusses the codebase, which can be reused for similar projects. Section 9 concludes the paper and describes possible future work.

2. Related Work

There is no dataset directly comparable to the work presented here, but there are three types of resources to which our work can be compared to. The first type is datasets that document spoken Romance languages, the second type documents dialectal variation in the Romance languages, and the third type are textual treebanks that provide annotation of clitic features. We discuss them in turn.

The first part focuses on the prosodic properties of spoken Romance languages. For example, Cresti et al. (2004) developed a multilingual corpus of spontaneous speech for main Romance languages (Italian, French, Portuguese, and Spanish). There is another database targeting French that contains prosodic annotation in (Lacheret et al., 2014), which provides a documented and reproducible syntactic-prosodic treebank of spoken French. The data contains 57 samples, each around 5 minutes long, totaling around 3 hours and 33k words. It provides orthographic information, and it is phonetically time-aligned by (phoneme, syllable, and word) with micro and macro syntactic and prosodic layers. The Atlàs interactiu de l'intonacion de l'occitan provides a curated prosodic atlas with audio and intonation contours (Prieto et al., 2007–2014).

For the second part, there are a few corpora that cover rural and minority Romance varieties. The Corpus Oral y Sonoro Del Español Rural (COSER) (Fernández-Ordóñez, 2005), which offers a search interface for Iberian transcription aligned recordings, and there is also a universal dependency grammar annotation of the corpus. The French Atlas sonore des langues régionales (Boula de Mareüil et al., 2017) provides another interactive website that allows searching for different minority languages in Europe, including France. The project provides IPA transcription of Aesop fable readings across regional varieties. Finally, we have the digitized Atlas Lingüístico de la Península Ibérica (ALPI) (García Mouton, 2016). While it is textual data without audio, it is still considered a baseline for geographic studies of varieties in the Iberian Peninsula. The NavigAIS/NavigAIS-Web (Tisato, 2009) digitization of the Sprach- und Sachatlas Italiens und der Südschweiz (AIS) (Linguistic and Ethnographic Atlas of Italy and Southern Switzerland) (Hall Jr, 1942) includes dialect maps (1928–1940) but lacks audio.

Finally, we will go over corpora that have annotated their data to include information on clitics in Romance languages. We will start with AnCora (Taulé et al., 2008), which has 500k words for Spanish and Catalan and provides multilayer annotations that include morphology and syntax. There is the Spanish HPSG Treebank, which is built on AnCora that adds clitic-aware morphosyntax and semantic information (Chiruzzo and Wonsever, 2018). There is also the Italian Stanford Dependency Treebank (Bosco et al., 2013), which provides a strong textual baseline for studying clitics.

3. Data Collection

Romance Clitics Database (RoCDat) is a searchable database that allows users to examine one of the rare patterns involving a class of pronouns (clitics) that are normally stressless, but which, in the languages under investigation, exhibit or modify stress. This pattern is found in some minority Romance languages spoken in the South of France, the South and North of Italy, the Balearic Islands, and the islands of Corsica and Sardinia.

For participant recruitment, we depended on local guides and their suggestions for other interviewees. In each region, we started at town halls, where we were able to meet with community leaders who helped us find the first few contacts. We also contacted people working at the hotels we were staying at to introduce us to additional speakers, thus forming a chain of participants to interview.

All interviews were conducted in accordance with the ethics procedures registered with our institution, and informed consent was obtained from each participant. Participant information was later anonymized (though retaining demographic information of sociolinguistic interest). The participants were all paid, but a majority refused to accept it, as they felt they had helped preserve their language.

Over 75 people were interviewed for this project between 2007 and 2010. The informants are native speakers of the particular Romance variety being investigated. They are also fluent in another standard Romance language (Italian, Spanish, Catalan, or French), which is the language in which the interview was conducted. The complete list of interview locations can be found in Table 15 and in Figure 1.

The database includes demographic information for each informant, including age at the time of the interview, year of birth, gender, and the location where the dialect is spoken (town, region, country). We also include the year of the interview, the interview language, and the name of the interviewer.

The interview consisted of a translation task from the standard Romance language the speaker

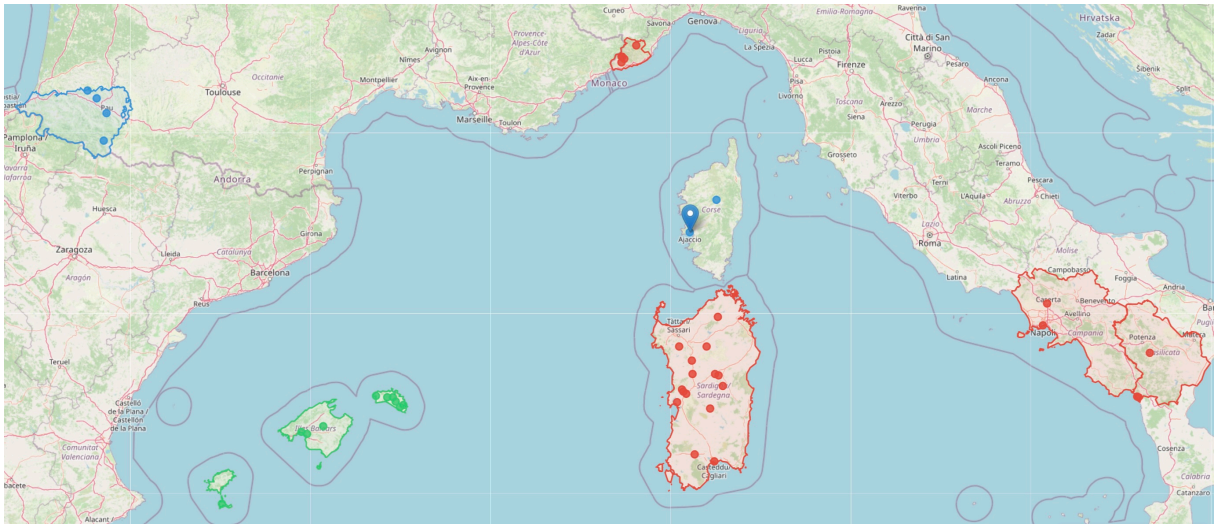

Figure 1: Geographic Spread of the Data. Italy is shaded in red, France is blue, and Spain is green.

is fluent in into their dialect. Each interview included approximately 100 utterances per speaker. The phrases the informant was asked to translate usually consisted of a verb + a clitic unit. We included the whole paradigms of verbs as well as all clitics and clitic combinations possible for each verb.

4. Data Annotation

We coded each phrase for 42 properties that are relevant to understanding the data. For example, for each verb in the phrase, we coded the person (1st, 2nd, 3rd) and number (singular, plural), whether it is tensed or not, and whether it is an imperative, infinitive, gerund, participle, auxiliary, or modal. For each pronoun in the utterance, we noted its role (direct object, indirect object, etc.), whether it is impersonal, locative, neuter, partitive, possessive, or reflexive.

We also coded relevant pronouns for gender (masculine, feminine, neuter), number (singular, plural), and person (1st, 2nd, 3rd). Each utterance is coded for phonological properties such as the position of stress in verb + clitic phrases (final, penultimate, antepenultimate, preantepenultimate), whether or not there is consonant gemination in the verb + clitic unit, etc. Many other aspects of the utterances are also coded; a complete list is shown in Table 16.

(2) Same utterance across two languages (form & order differ; meaning: 'give it.MS to me!').

a. [donə-l-má]
60_36 Manacor, Mallorca, Spain

b. [da-mm-íllu]
31_9, Massa di Maratea, Basilicata, Italy

Country	# Speakers	Minutes (Total)
Italy	35	142.5
Spain	28	59.52
France	12	64.17
TOTAL	**75**	**266.19**

Table 1: Data statistics by country.

(3) Same language (Corte, Corsica, France): preverbal vs. postverbal clitics.

a. [u-mi-wɛ́ndi]
'you sell it.MS to me'

b. [kómpʁa-mi-lu]
'buy it.MS for me!'

Example 2 shows two features that are documented in the database (clitic order, position of stress), and we can examine their interaction, namely the relationship between clitic order (ACC-DAT in Manacor, and DAT-ACC in Mass di Maratea), and stress position (final in Manacor, and penultimate in Massa di Maratea).

The pair of utterances in Example 3 illustrates in how clitic order and form can vary in the same language depending on the position of the pronouns: ACC-DAT (with ACC /u/) in preverbal position, and DAT-ACC (with ACC /lu/) in postverbal position

5. Database

The main release of the database is a CSV file with one row per utterance. Each row will

Region	# Speakers	Minutes (Total)
France, Corsica	7	31.35
France, Pyrénées-Atlantiques	5	32.82
Italy, Sardinia	24	121.81
Italy, Liguria	6	9.15
Italy, Basilicata	3	11.01
Italy, Campania	2	0.53
Spain, Balearic Islands: Mallorca	13	23.76
Spain, Balearic Islands: Menorca	9	20.11
Spain, Balearic Islands: Formentera	6	15.65

Table 2: Sample Data Statistics by Region.

Interview year	# Speakers	Minutes (Total)
2007	36	164.7
2008	34	68.67
2010	5	32.82

Table 3: Speakers by Interview Year.

	Female	Male
Number of speakers	36	39
Minutes for all speaker	117.27	142.48

Table 4: Data statistics by gender.

include utterance_id, variety_code, speaker_id, orthographic transcription, IPA transcription, translation, morphosyntax for each verbal head (person/number, tense/mood/aspect, finiteness/imperative/participle/gerund/aux/modal), clitic features (function, case/role, person/number/gender), prosody/phonology (stress position, gemination, stress shift, and displaced stress), and a url for the corresponding sound file.

There is another companion file, a CSV containing the speakers' information, such as age at interview, year of birth, gender, town, region, country, year of interview, and interview language.

6. Dataset Statistics

The dataset has 12,529 word tokens and 5,077 utterance. In Table 1 shows that the corpus spans three countries: Italy, Spain, and France. It includes 75 speakers in total. Italy has the largest number of speakers (35) and the greatest total recording time (142.5 minutes). Spain has 28 speakers, concentrated in the Balearic Islands (Mallorca, Menorca, Formentera), while France has 12 speakers from Corsica and the Pyrénées-Atlantiques. Averaged across speakers, France has the longest recordings, whereas Spain has shorter sessions per speaker.

Table 2 provides a detailed regional view of where the speech material was collected. Sardinia accounts for the largest share of total recording time, followed by the Balearic Islands (with separate coverage of Mallorca, Menorca, and Formentera) and the French sites (Corsica and the Pyrénées-Atlantiques). This distribution reflects two priorities of the project: (i) sampling areas where clitic behavior interacts robustly with prosody (stress placement and gemination), and (ii) covering multiple Romance sub-varieties within each macro-region to allow for some cross-variety comparisons.

Table 3 shows the recordings' temporal distribution, which is clustered, with the majority collected in 2007 and 2008 and the rest in a follow-up trip in 2010. Table 4 demonstrates that the speaker pool is balanced by gender (36 women, 39 men), with comparable total recording time across groups. This balance is important for both documentation and modeling: it reduces the risk that speaker-specific acoustic or prosodic tendencies will spuriously correlate with linguistic variables of interest (e.g., clitic position or stress). We also retain demographic metadata (age at interview, year of birth, town/region) to allow researchers to control for sociolinguistic factors when analyzing clitic realization and prosodic patterns.

Table 5 and Table 6 show statistics about prosodic and clause-level information. Geminate consonants are approximately 34%, while displaced stress is less common approximately 24%. Negation and interrogatives are rare, 5.9% and 2.0%, respectively, and prepositional complementizers occur in 11.6% of cases. The stress-position table shows that the penultimate stress is the most common pattern (55.60%), followed by antepenultimate (25.91%), final (14.60%), and preantepenultimate (3.88%). The verb-feature tables also demonstrate strong asymmetries: for Verb 1, imperative marking is frequent (53.81%), while auxiliary, modal, gerund, and participle values are less common. In Verb 2 and Verb 3, annotations are sparser. Tables 7–14 provide a more detailed picture of morphosyntactic structure.

Feature	Count	Yes n (%)	No n (%)
Prosody / phonology			
Geminate Consonant	4996	1696 (33.95%)	3299 (66.03%)
Displaced Stress	4472	1083 (24.22%)	3389 (75.78%)
Verb 1 binary features			
Inf	4431	327 (7.38%)	4104 (92.62%)
Imp	4430	2384 (53.81%)	2046 (46.19%)
Aux	4430	503 (11.35%)	3927 (88.65%)
Tensed	4430	1664 (37.56%)	2765 (62.42%)
Modal	4430	378 (8.53%)	4052 (91.47%)
Gerund	4429	44 (0.99%)	4385 (99.01%)
Participle	4244	18 (0.42%)	4226 (99.58%)
Verb 2 binary features			
Inf	1196	759 (63.46%)	437 (36.54%)
Aux	1181	45 (3.81%)	1136 (96.19%)
Modal	1181	53 (4.49%)	1128 (95.51%)
Gerund	1181	254 (21.51%)	927 (78.49%)
Participle	1164	161 (13.83%)	1003 (86.17%)
Verb 3 binary features			
Inf	178	163 (91.57%)	15 (8.43%)
Participle	164	6 (3.66%)	158 (96.34%)
Clause-level binary features			
Negation	4983	293 (5.88%)	4690 (94.12%)
Interrogative	4983	98 (1.97%)	4885 (98.03%)
Prepositional Comp	5011	579 (11.55%)	4432 (88.45%)

Table 5: Binary annotation features labels.

Feature	Count	Final		Penultimate		Antepenultimate		Pre-antepenultimate	
		n	%	n	%	n	%	n	%
Stress Final	4971	726	14.60	2764	55.60	1288	25.91	193	3.88

Table 6: Stress-position distribution.

7. Interface

The public web interface provides access to the corpus through two views: the search and the speakers view. The search view allows left-hand faceted filters (variety, speaker demographics, verb/clitic feature) and a text search through orthography, IPA, or both. This view is shown in more detail in Figure 2.

This results in a paginated list of utterances that includes playable audio, a map of the approximate speaker location, and the utterance's properties.

The speaker view lists all speakers and their associated utterances. Similar to the search view results, the results include a paginated list of utterances with playable audio, a map showing the approximate location of the speaker, and the utterances' properties.

Figure 2 shows the search functionality view on the website, where the user can specify the properties to filter for. In this example, the user specified category Pronoun 1, the property Pronoun 1 Person, and did not choose the value singular for 1st person. They also selected geographic options by selecting the country of France and the region of Corsica.

The Figure 4 shows the result of filtering for specific demographics. In Figure 4 we show the result of filtering for France. We see the list of speakers that fulfill this category.

The last Figure 3 shows the information contained in a single utterance. If we clicked on one of the speakers from the previous result, we would get the view in Figure 3. This view contains metadata about the speaker, such as age, gender, and location, along with a map view and information about all utterances for that speaker. In Figure 3, we expanded a single utterance to show the information contained there. The utterance can be played in audio and contains the transcription, properties, and values for these entries.

This design serves two groups at once: linguists and other academics who need to listen to and browse the data on the web, and NLP and software

Source	Count	1		2		3	
		n	%	n	%	n	%
Verb 1	4123	1041	25.25	2316	56.17	766	18.58
Pronoun 1	4180	1396	33.40	472	11.29	2312	55.31
Pronoun 2	1638	192	11.66	25	1.52	1421	86.33
Pronoun 3	16	3	18.75	–	–	13	81.25

Table 7: Person distributions.

Source	Count	Singular		Plural	
		n	%	n	%
Verb 1	4133	3336	80.62	797	19.26
Pronoun 1	3539	2764	77.40	775	21.70
Pronoun 2	1585	1238	77.76	347	21.80
Pronoun 3	16	14	73.68	2	10.53

Table 8: Number distributions.

engineers who can script and write code to query the API for the data. While users can currently download the RoCDat via the CSV, we also provide access to data via an API for future databases built on top of the RoCDat infrastructure, if needed.

8. Interface Code

We are also releasing our interface code, so that other researchers can easily create an interface to their own database. To use the interface code, the user needs to provide the data in CSV files. The input CSV files require the utterance text, a speaker ID, and a URL corresponding to the audio file hosting location. Based on this input, an interface is created. The new user can make minor edits to specified files to update the interface text to reflect the new data.

The interface code is a small, modular Django 5 application that exposes the corpus via a REST API using the Django REST framework and a native HTML/CSS/JavaScript front-end. The data model separates geographical entities by country, region, and town, speaker demographics, and utterances, with a flexible linguistics_features JSON field that allows any type of attribute, including those not described in the paper.

To add audio, users need to create an S3 URL path for utterances with audio data. Then the configuration requires adding the DATABASE URL, S3 bucket, and Allowed Hosts.

The repository is shipped with Docker/Gunicorn/Whitenoise for one-command deployment and static asset handling. We will include a GitHub URL in the final version of the paper.

9. Conclusion

We have presented the Romance Clitics Database, a database of clitics in low-resource Romance dialects. The data includes metadata, sound files, and orthographic and IPA transcriptions, and syntactic, morphological, and phonological annotations. It is accessible through an interface, and we also distribute the raw data. In addition, we are distributing the code that takes the raw data and returns the interface to encourage code reuse.

In future work, we aim to expand our database by encouraging linguists to contribute relevant data through crowdsourcing. (Prieto et al., 2007–2014)

10. Acknowledgments

This work was supported by the National Science Foundation (grant #0617471) and NSF Supplemental Funding “Research Experiences for Undergraduates” awarded to Lori Repetti and Francisco Ordonez (2006-2011), and by a Stony Brook University FAHSS grant (2011).

11. Bibliographical References

Cristina Bosco, Simonetta Montemagni, Maria Simi, et al. 2013. Converting italian treebanks: Towards an italian stanford dependency treebank. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 61–69. The Association for Computational Linguistics.

Philippe Boula de Mareüil, Frédéric Vernier, and Albert Rilliard. 2017. Enregistrements et tran-

Source	Count	Masculine		Feminine	
		n	%	n	%
Pronoun 1	1614	1116	67.35	498	30.05
Pronoun 2	1352	958	70.23	394	28.89
Pronoun 3	13	12	92.31	1	7.69

Table 9: Pronoun gender distributions.

Source	Count	DO		IO		Subject	
		n	%	n	%	n	%
Pronoun 1	3769	1820	48.19	1843	48.80	106	2.81
Pronoun 2	1598	1326	82.57	271	16.87	1	0.06
Pronoun 3	16	10	62.50	6	37.50	–	–

Table 10: Pronoun role distributions.

scriptions pour un atlas sonore des langues régionales de france. *Géolinguistique*, (17):23–48.

Massimo Cerruti and Riccardo Regis. 2014. Standardization patterns and dialect/standard convergence: A northwestern italian perspective. *Language in society*, 43(1):83–111.

Luis Chiruzzo and Dina Wonsever. 2018. Spanish hpsg treebank based on the ancora corpus. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Emanuela Cresti, Fernanda Bacelar Do Nascimento, Antonio Moreno-Sandoval, Jean Veronis, Philippe Martin, and Khalid Choukri. 2004. The c-oral-rom corpus. a multilingual resource of spontaneous speech for romance languages. In *LREC*.

Alexander Erdmann and Nizar Habash. 2018. Complementary strategies for low resourced morphological modeling. In *Proceedings of the Fifteenth Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 54–65.

Inés Fernández-Ordóñez. 2005. Corpus oral y sonoro del español rural. *Madrid, Universidad Autónoma de Madrid*.

Pilar García Mouton. 2016. Inés fernández-ordóñez, david heap. *Maria Pilar PEREA, João SARAMAGO & Xulio SOUSA*.

Robert A Hall Jr. 1942. Sprach-und sachatlas italiens und der südschweiz.

Anne Lacheret, Sylvain Kahane, Julie Beliao, Anne Dister, Kim Gerdes, Jean-Philippe Goldman, Nicolas Obin, Paola Pietrandrea, and Atanas Tchobanov. 2014. Rhapsodie: a prosodic-syntactic treebank for spoken french. In *Language Resources and Evaluation Conference*.

Rosangela Lai. 2017. Stress shift under cliticization in nuorese sardinian. *Quaderni di linguistica e studi orientali*, 3:183–199.

John M Levis. 2018. *Intelligibility, oral communication, and the teaching of pronunciation*. Cambridge University Press.

Francisco Ordóñez and Lori Repetti. 2006. Stressed enclitics? *AMSTERDAM STUDIES IN THE THEORY AND HISTORY OF LINGUISTIC SCIENCE SERIES 4*, 276:167.

Alan Ramponi and Camilla Casula. 2023. Diatopit: A corpus of social media posts for the study of diatopic language variation in italy. In *Tenth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2023)*, pages 187–199.

Lori Repetti. 2018. Fieldwork and building corpora for endangered varieties. *Ayres-Bennet, W. & J. Carruthers (edd.): Manual of Romance Sociolinguistics. De Gruyter, Berlin/Boston*, pages 114–133.

Mariona Taulé, Maria Antònia Martí, and Marta Recasens. 2008. Ancora: Multilevel annotated corpora for catalan and spanish. In *Lrec*, volume 2008, pages 96–101.

Francis Zheng, Edison Marrese-Taylor, and Yutaka Matsuo. 2024. Improving low-resource machine translation for formosan languages using bilingual lexical resources. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 11248–11259.

Source	Count	-1-		Reflexive		Possessive		Locative		Partitive		Impersonal		Other	
		n	%	n	%	n	%	n	%	n	%	n	%	n	%
Pronoun 1	2615	1430	54.68	613	23.44	206	7.88	180	6.88	167	6.39	10	0.38	9	0.34
Pronoun 2	1487	1315	88.43	48	3.23	–	–	67	4.51	54	3.63	–	–	3	0.20
Pronoun 3	12	9	75.00	–	–	–	–	1	8.33	2	16.67	–	–	–	–

Table 11: Pronoun form distributions.

Feature	Count	Enclisis		Proclisis		Other-pattern	
		n	%	n	%	n	%
Single-verb clitic position	3197	2188	68.44	1007	31.50	2	0.06

Table 12: Single-verb clitic position.

12. Language Resource References

Bosco, Cristina and Montemagni, Simonetta and Simi, Maria and others. 2014–present. *Converting italian treebanks: Towards an italian stanford dependency treebank*. Stanford deps/UD-converted Italian treebank; Evalita 2014 origin.

Boula de Mareüil, Philippe and Vernier, Frédéric and Rilliard, Albert. 2017–. *Enregistrements et transcriptions pour un atlas sonore des langues régionales de France*. Interactive recordings+transcriptions of regional languages of France.

Cresti, Emanuela and Do Nascimento, Fernanda Bacelar and Moreno-Sandoval, Antonio and Veronis, Jean and Martin, Philippe and Choukri, Khalid. 2004. *The C-ORAL-ROM CORPUS. A Multilingual Resource of Spontaneous Speech for Romance Languages.* ELRA. Multilingual spontaneous speech (IT/FR/PT/ES); LREC 2004 resource description.

Fernández-Ordóñez, Inés. 2005–present. *Corpus oral y sonoro del español rural*. Searchable audio+transcripts of rural Spanish; ongoing releases.

Lacheret, Anne and Kahane, Sylvain and Beliao, Julie and Dister, Anne and Gerdes, Kim and Goldman, Jean-Philippe and Obin, Nicolas and Pietrandrea, Paola and Tchobanov, Atanas. 2014. *Rhapsodie: a prosodic-syntactic treebank for spoken french*. 57 five-minute samples, orthographic+phonetic alignment; LREC 2014.

Pilar Prieto, Rafèu Sichel-Bazin, and Trudel (eds.) Meisenburg. 2007–2014. Atlàs interactiu de l'intonacion de l'occitan. Online resource. Accessed 24 Oct 2025.

Taulé, Mariona and Martí, Maria Antònia and Recasens, Marta. 2008–. *Ancora: Multilevel annotated corpora for Catalan and Spanish.* 500k words per language; multilayer annotations; LREC 2008.

Graziano Tisato. 2009. Ais digital atlas and navigation software. Online resource (no audio). Accessed 24 Oct 2025.

Feature	Count	Proclisis		Enclisis		Verb-cl-verb		Other-pattern	
		n	%	n	%	n	%	n	%
Two-verb clitic position	839	442	52.68	189	22.53	186	22.17	22	2.62

Table 13: Two-verb clitic position.

Feature	Count	Proclisis		V-V-cl-V		Enclisis		V-cl-V-V		Other-pattern	
		n	%	n	%	n	%	n	%	n	%
Three-verb clitic position	125	53	42.40	32	25.60	27	21.60	8	6.40	5	4.00

Table 14: Three-verb clitic position.

Country	Region / Area	Towns
Italy	Sardinia	Bonorva, Cabras, Cagliari, Fonni, Ittiri, Laconi, Luras, Macomer, Nino, Milis, Nuoro, Oristano, Ozieri, Seneghe, Siliqua, Orani.
	Campania	Napoli, San Leucio.
	Basilicata	Anzi, Maratea, Massa di Maratea.
	Liguria	Baiardo, Ormea, Perinaldo, Pigna, Viozene, Castelvittorio.
France	Corsica	Ajaccio, Boccognano, Corte.
	Pyrénées Atlantiques	Arnos, Sault, Vallée d'Ossau.
Spain	Balearic Islands Formentera	Formentera
	Balearic Islands Mallorca	Manacor, Sineu, Llucmajor
	Balearic Islands Menorca	Es Mercadal, Ciutadella, Maó, Ferreries, Alaior, Sant Climent

Table 15: Towns Where Interviews Were Conducted.

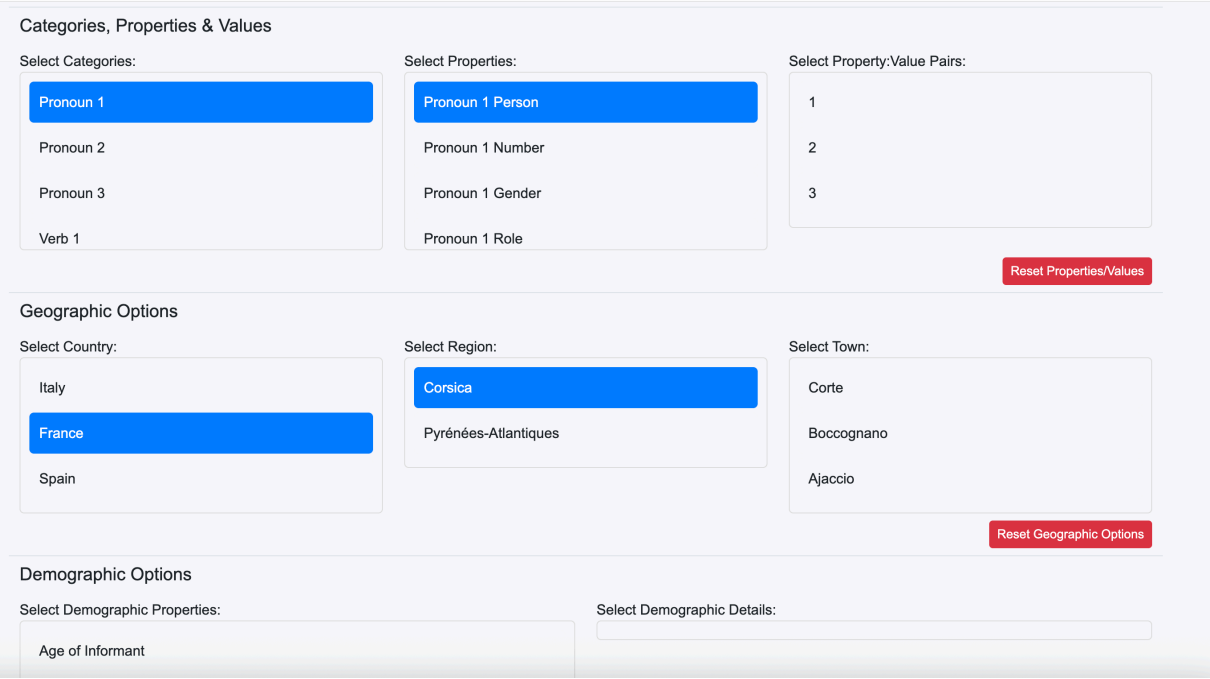

Figure 2: Search View

Feature	Values / coding
Stress	fin / pen / antepen / pre-
Displaced stress	y / n
Geminate consonant	y / n
clitic position (single verb)	pro / en / clVcl / etc.
clitic position (two verbs)	pro / en / clVcl / etc.
clitic position (three verbs)	pro / en / clVcl / etc.
verb 1 person	1 / 2 / 3
verb 1 number	s / p
verb 1 tensed	y / n
verb 1 imperative	y / n
verb 1 infinitive	y / n
verb 1 gerund	y / n
verb 1 participle	y / n
verb 1 auxiliary	y / n
verb 1 modal	y / n
verb 2 infinitive	y / n
verb 2 gerund	y / n
verb 2 participle	y / n
verb 2 auxiliary	y / n
verb 2 modal	y / n
verb 3 infinitive	y / n
verb 3 past participle	y / n
pronoun 1 role	DO / IO
pronoun 1 form	impers / loc / etc.
pronoun 1 gender	m / f
pronoun 1 number	s / p
pronoun 1 person	1 / 2 / 3
pronoun 2 role	DO / IO
pronoun 2 form	impers / loc / etc.
pronoun 2 gender	m / f
pronoun 2 number	s / p
pronoun 2 person	1 / 2 / 3
pronoun 3 role	DO / IO
pronoun 3 form	impers / loc / etc.
pronoun 3 gender	m / f
pronoun 3 number	s / p
pronoun 3 person	1 / 2 / 3
interrogative sentence	y / n
negation sentence	y / n
prepositional complementizer sentence	y / n
unknown element	y / n

Table 16: Feature inventory used in the corpus annotation. Abbreviations: y/n = yes/no, s/p = singular/plural, DO/IO = direct/indirect object, pro/en/clVcl = proclitic/enclitic/clitic–verb–clitic.

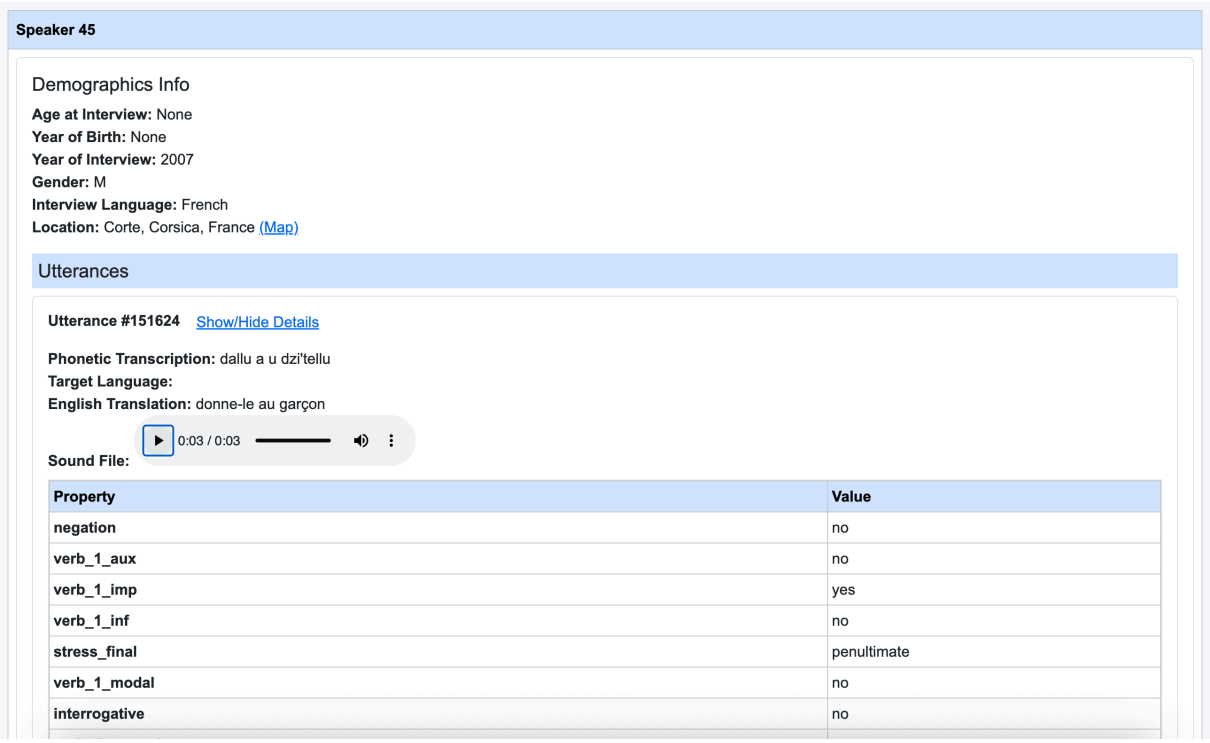

Figure 3: An Utterance Information

Figure 4: List of Utterances for France