Introducing Universal Dependencies for Sardinian: the UD ContSar Treebank

Nicoletta Puddu[1], Manuela Sanguinetti[2], Luigi Talamo[3]
[1]Dept. of Literature, Languages and Cultural Heritage, University of Cagliari (Italy)
[2]Dept. of Mathematics and Computer Science, University of Cagliari (Italy)
[3]Dept. of Language Science and Technology, Saarland University (Germany)
{nicoletta.puddu, manuela.sanguinetti}@unica.it
luigi.talamo@uni-saarland.de

Abstract

This paper introduces the first steps towards the creation of a novel resource for contemporary Sardinian within the Universal Dependencies framework. Sardinian is a Romance language spoken in Sardinia, an island belonging to the Italian Republic and located in the center of the western Mediterranean. It is a minority and endangered language, traditionally transmitted mainly orally, and characterized by a multiplicity of varieties (usually grouped into two macro-varieties Logudorese and Campidanese), all recognized as part of the Sardinian linguistic continuum. These varieties share basic morphosyntactic features, while presenting differences at the lexical level and in the realization of specific constructions. This internal variation can be particularly challenging with regard to the normalization of lemmas and the linguistic characterization of certain phenomena. The development of the treebank therefore aims to provide an annotated resource for contemporary Sardinian that takes into account the specificities of the different varieties, using Universal Dependencies to represent them within a unified theoretical framework, in order to facilitate both linguistic analysis and automatic processing. The present paper thus describes some linguistic characteristics of Sardinian and the attempts to encode them within the UD framework. Finally, we present the results of our evaluation of an NLP pipeline for Sardinian, trained on our corpus, for the Stanford Stanza parser.

Keywords: Sardinian, UD, Treebank

1. Introduction

The possibility of using a unified representation scheme such as Universal Dependencies, which has consolidated its role over the years as the standard for treebank annotation, has led to the proliferation of new resources of this type for both widely spoken languages and minority or under-resourced languages. Narrowing it down to Italy alone, it is well known that the country is rich in dialects and linguistic varieties in addition to Italian (see, for example, the overview proposed in Ramponi 2024). Some of these have also found their way into the UD project: this is particularly the case with Ligurian[1], Neapolitan[2], and Sicilian (Cappello et al., 2025)[3]. Along the lines of such efforts, in this paper we introduce an ongoing work on the development of a UD-based treebank for contemporary Sardinian.

Sardinian (ISO 639-3 srd) is a Romance language, spoken on the island of Sardinia, belonging to the Republic of Italy. It shows many conservative features, especially at the phonetic and lexical level, but it has also been considerably influenced by Iberian languages (Castilian and Catalan), given that it was under the Catalan-Aragonese domination for more than four centuries. It is considered an endangered language, and is protected by National Law n.482/1999 and Regional Laws n.26/1997 and n.22/2018. There are no recent surveys on the actual number of speakers, but in a study conducted by the Autonomous Region of Sardinia on 2.437 inhabitants, 68.4% of the respondents claimed to be able to understand and speak the language and a further 29% only to understand but not speak it (Oppo, 2007).

Sardinian is traditionally subdivided into four varieties: Campidanese, Logudorese, Nuorese and Arborense (Virdis, 1988). Two further varieties, Sassarese and Gallurese, will not be addressed here since they are classified by scholars as non-Sardinian varieties, being closer to Corsican and Tuscan varieties. The sub-varieties of Sardinian are usually grouped into two macro-varieties: a center-southern variety, named Campidanese (ISO 639-3 sro), and a center-northern variety, named Logudorese (ISO 639-3 src). It is, however, important to note that this traditional division into varieties has been challenged, due to the presence of a transitional area in the central region. (Hajek and Goebl, 2021; Loporcaro and Putzu, 2024).

Moreover, in 2006, a standard variety, called Limba Sarda Comuna (LSC), meaning "Common Sardinian Language", was developed by the Autonomous Region of Sardinia for documents is-

[1] https://github.com/UniversalDependencies/UD_Ligurian-GLT
[2] https://github.com/UniversalDependencies/UD_Neapolitan-RB
[3] https://github.com/UniversalDependencies/UD_Sicilian-STB

Proceedings of the Ninth Workshop on Universal Dependencies (UDW 2026), pages 242–251
May 16, 2026.

sued by the regional administration (Regione Autonoma della Sardegna, 2006). LSC is intended to be a "mediation variety", taking into account the most widespread features of Sardinian, and also picking several features of the transition varieties spoken in the centre of the island. Finally, the regional administration, in cooperation with the University of Cagliari, also issued provisional orthographic guidelines for Sardinian as part of the experimental certification of language competence C1 (Regione Autonoma della Sardegna, 2022; Grosso et al., 2024).

Scholars generally agree that the differences across varieties are mostly phonetic and lexical, while at the morphosyntactic level Sardinian shows a general uniformity (Virdis, 1988). Given this substantial unity, our working hypothesis is that it is possible to develop a single morphosyntactic annotation and a unified treebank covering all main Sardinian varieties. The availability of a resource with such characteristics can be particularly useful both for consulting and analyzing specific linguistic constructs of Sardinian across its varieties, but also for the development of NLP tools.

In this respect, although the development of the resource is currently in its infancy, the resource has already made it possible to (i) formalize annotation strategies for a number of Sardinian-specific constructions and (ii) carry out a preliminary training of the Stanza dependency parser. While the results are necessarily exploratory given the limited size of the dataset, they provide a first indication of the feasibility of building UD-based parsing models for Sardinian. This paper will therefore present some linguistic features of Sardinian that have been represented according to UD principles, together with the preliminary results obtained with the parser.

2. Related Work

Concerning Sardinian, early attempts to build text corpora can be found in Fortunato and Ravani (2015) with the ATLISOR textual archive[4], an unannotated corpus of Medieval texts, and later on in Puddu and Talamo (2020) with the EModSar[5], a corpus of Early Modern Sardinian texts that were also POS-tagged. Preliminary experiments on both POS and syntactic annotation, instead, were carried out in Puddu and Stein (2018) using the Arborator web interface (Gerdes, 2013) to syntactically annotate a corpus of Medieval Sardinian. However, to the best of our knowledge, no reference treebank exists for contemporary Sardinian. A spoken corpus of Sardinian emigrants has been introduced in Pisano et al. (2019) and Mura et al. (2023), with POS annotation only. The same resource was later used to train a BERT-based POS tagger (Carta et al., 2025).

Further collections of spoken data were also created in Chizzoni and Vietti (2024) and De Cristofaro et al. (2025), where the recorded data served the purpose of developing and inspecting an ASR model for Campidanese Sardinian. Finally, some tools have been developed for rule-based Machine Translation from Italian to Sardinian (Tyers et al., 2017) and from Catalan to Sardinian (Fronteddu et al., 2017)

The present paper thus represents the first attempt to build an openly available and syntactically annotated resource of texts in contemporary Sardinian.

3. Data

The construction of corpora for minority languages often relies heavily on web sources, even when printed texts exist, essentially for copyright reasons. Due to its open license, Wikipedia often represents a valuable source of data for treebank development. In the case of Sardinian, this is particularly relevant, as Wikipedia pages explicitly indicate the variety in which each entry is written: Limba Sarda Comuna (LSC), Logudorese, Campidanese or Nuorese. This is particularly useful in the context of non-standardized languages, where identifying the corresponding sub-variety may not be trivial and may thus compromise a balanced diatopic representation. At the same time, relying on a single platform would risk limiting the range of registers and constructions represented in the corpus. Textual diversity is thus often required, in order to ensure a broader coverage of linguistic phenomena.

To balance the data scarcity problem with principles of varietal and genre variation, we preliminarily selected a sample to use for the pilot annotation, which features two Wikipedia pages (one in LSC[6] and one in Campidanese[7]), and a fairy tale in Dorgalese, a sub-variety of Eastern and Southern Nuorese, titled *Mannoi Corrias e sa pudda bianca*, by Gonario Carta-Brocca[8].The basic statistics on this preliminary sample are also reported in Table 1.

[4] http://atlisorweb.ovi.cnr.it/

[5] https://linguistica.dh.unica.it/emodsar

[6] Source: https://sc.wikipedia.org/wiki/Arte

[7] Source: https://sc.wikipedia.org/wiki/Casteddu/campidanesu

[8] Source: Public URL no longer available at the time of writing.

Section	Variety	sent.	tok.
Mannoi	src	138	2312
Wiki-Casteddu	sro	39	1218
Wiki-Arte	LSC	17	488
	tot.	194	4.018

Table 1: Basic statistics of the pilot annotated sample. `src` and `sro` refer to the ISO code of the Logudorese and Campidanese Sardinian varieties, respectively, while LSC stands for *Limba Sarda Comuna*.

4. Annotation Principles

This section provides an overview of the main linguistic characteristics of Sardinian focusing on their representation according to the core principles of Universal Dependencies.

4.1. Tokenization

Tokenization was carried out following orthographic criteria that are consistent with standard UD practices. In general, words are delimited by whitespaces or apostrophes. Punctuation marks are treated as independent tokens; the only exception concerns apostrophes, which are always attached to the adjacent word. More specifically, in cases of elision, the apostrophe is attached to the end of the word containing the elided vowel (*unu ecantu*$_{src}$[9] → *un'ecantu*, 'a piece'). In cases of apheresis, the apostrophe is attached to the beginning of the word undergoing the sound loss (*de* → *'e*, 'of').

Multi-word tokens are introduced only in cases of contractions between verbs and clitic pronouns, where the orthographic word corresponds to multiple syntactic words (*daemilu*$_{src}$ → *dae mi lu*, 'give to-me it').

4.2. Orthography and Lemmatization

The spelling and lemmatization choices adopted in this work are mainly based on a set of general rules, that we have followed to ensure consistency and uniformity across the different Sardinian varieties. Given the mediating nature of the LSC, these general rules primarily refer to the guidelines provided by the Autonomous Region of Sardinia, that consider the following underlying principles:

- Accent marks in Sardinian serve exclusively to indicate stress placement and do not distinguish vowel quality. Specifically, the use of accents is compulsory in oxytone and proparoxytone words (e.g., *fàghere*$_{src}$ 'to do', *èssere*$_{src}$ 'to be', *merì*$_{sro}$ 'evening'). Consider, however, that oxytones are not very common in Sardinian, since they often develop a paragogic vowel as in *tue*$_{src}$ 'you' or *caffei*$_{sro}$ 'coffee' (Virdis, 1978; Dettori, 2002).
- Given the diffused lenition of intervocalic singleton stops and fricatives, Sardinian does not display, for many consonants, a clear opposition between singleton and geminate consonants. Consequently, except for the consonants [b], [d], [l], [m], [n], [r], and [s], which can be either simple or geminate, all the other consonants in intervocalic position are written as simple.

We thus resorted to these rules to properly handle a wide range of spelling and accent inconsistencies at the lemma level, probably due, in many cases, to the scholarization in Italian of the writers. In particular, we found frequent omission of accents especially in proparoxytone words (accordingly, the noun *omine* 'man' has been lemmatized as *òmine*), where Italian ortographic rules do not prescribe it. We also found variation in consonant gemination, with double consonants occurring in contexts where Sardinian orthography does not allow them. Consequently, we have instances like *fèmminas* 'women', which has been lemmatized as *fèmina*). This tendency likely reflects not only an attempt to encode a more intense pronunciation, but, again, a possible interference of Italian orthographic conventions.

Another recurring case of normalization, always at the lemma level, concerned the so-called *euphonic d*, that refers to the insertion of a final [d] in certain words (the most frequent being the prepositions *in* 'in' and *cun*,'with'), when the following word begins with a vowel, to facilitate pronunciation. In line with LSC guidelines, these occurrences were thus reduced to their basic form (*totu ind'unu* → *totu in unu*, 'all in one').

The lemmatization process was supported by lexicographical tools and reference resources. In particular, the Dictionary of Sardinian Language and Culture by Mario Puddu (Puddu, 2015) (specifically, its online version)[10] was used as a guide for selecting lemma forms, and the CROS spell checker - Corpus de Referèntzia de su Sardu,[11] also accessible online, was used to verify the consistency of the choices made.

It is worth pointing out that, whenever possible, a unified lemma form was adopted across different Sardinian varieties. This choice was motivated not only by the need for internal coherence, but also by methodological considerations

[9] As mere convention, we will use subscripts to specify, whenever required, the variety - expressed with its ISO code - to which the reported form belongs to. Word forms without subscripts are thus common to all the varieties considered.

[10] https://ditzionariu.nor-web.eu/en

[11] https://cros.nor-web.eu/

related to corpus querying. Adopting a unified lemma makes it possible to retrieve and compare the same morphosyntactic phenomenon across varieties, avoiding the fragmentation of the data into variety-specific lemma entries.

4.3. Morphology and POS Tagging

Similarly to other Romance languages, Sardinian has a rich inflectional morphology, which also compounds with internal variations across its main varieties. While a full account of the inflectional morphology of contemporary Sardinian is well beyond the scope of this paper, we discuss here some relevant phenomena and their implications in terms of UD representation (for a brief sketch of Sardinian from a typological perspective see Putzu 2017; Loporcaro and Putzu 2024).

Sardinian marks gender and number in nominals. Adjectives generally agree with the noun in gender and number, including possessive adjectives (with the exception of the invariant third person plural *issoro*, 'their/theirs'). Degree morphology is limited: comparative and superlative meanings are typically expressed analytically, with lexicalized exceptions such as *mèzus*$_{src}$/*mèllus*$_{sro}$ (meaning both 'better' and 'best') and *peus*$_{src}$/*pejus*$_{sro}$ ('worse' or 'worst'). The suffix *-ìssimu* exists but is restricted to marked or poetic registers. As a result, the feature `Degree` is only marginally relevant when annotating texts in contemporary Sardinian.

Sardinian has both definite and indefinite articles, while it lacks partitive articles (Virdis, 2007). Definite articles mark gender and number in the singular (*su, sa*), but in the plural, while Logudorese has a masculine *sos* opposed to a feminine *sas*, Campidanese only has the unmarked form for gender *is*.

Quantifiers such as *totu* ('all'), *dogni*$_{src}$/*cada*$_{sro}$ ('every'), and *carchi* ('some') are invariable and typically precede the noun, whereas *meda* ('much, many') is generally invariable but tends to occur postnominally (Mensching, 2017). Their morphosyntactic behavior informs their POS assignment and feature specification, but does not require language-specific morphological categories.

Concerning the pronominal system, Sardinian distinguishes between tonic (stressed) and atonic (clitic) pronouns (Loporcaro and Putzu, 2024). In particular, third-person clitic pronouns are morphologically differentiated for accusative and dative functions. This distinction is encoded in our annotation using the feature `Case=Acc` or `Case=Dat`.

Tonic pronouns have an additional property in the first and second person singular: special forms occur when the pronoun is preceded by the preposition *a* (e.g., *a mie*$_{src}$/*mimi*$_{sro}$, 'to me'). This behavior, which is exceptional within Romance, is treated in UD by marking the pronoun with the appropriate `Case` value (typically `Dat` or `Acc`, depending on syntactic function, see Section 4.4), while the preposition is annotated independently as `ADP`. A further distinctive feature concerns comitative constructions with *cun* ('with'). Sardinian, in the first and second singular form, allows forms such as *cun megus/tegus* ('with me/you'), where the pronominal form differs from the canonical nominative. This phenomenon is shared with languages such as Spanish and Portuguese. In UD terms, this is represented by assigning the comitative feature `Case=Com` to the pronoun, thus capturing the morphologically marked status of these forms while preserving the standard syntactic relation (`obl` with case marking).

The system of demonstratives is tripartite, distinguishing proximal, medial and distal pronouns and adverbs (Putzu, 2015)

Sardinian, as the other Romance languages, also has a dedicated reflexive pronoun for third person (tonic *se*, clitic *si*), which is used in reflexive, impersonal and passive constructions (Puddu, 2005).

As for the verb, Sardinian has three conjugation classes: *-àre*$_{sro}$ *-ài*$_{src}$; *-ere -i(ri)*$_{src}$; *-ìre*$_{sro}$*-ìri*$_{src}$. One of the most notable features of Sardinian is the preservation of final *-s* and *-t* in verb endings. However, in spoken Sardinian, final consonants are often followed by a paragogic vowel, which some authors may mark also in the written languages.

Sardinian has lost the original Latin simple perfect and, as northern Romance varieties uses the Romance compound Latin perfect (Virdis, 2007).

As in other Romance varieties, the synthetic forms for future and conditional derive from perifrastic constructions. The future comes from Latin HABĔO+AD+VERB.INF which results in the forms *apo a cantare*$_{sro}$/*apu a cantai*$_{sro}$/,'I will sing', lit. 'I have to sing'.

The conditional shows a more complex distribution: in Campidanese it derives from HABĒBAM (AD) + VERB.INF, so it is formed by the imperfect indicative of the verb *ai*$_{src}$ 'to have' + *a* + VERB.INF resulting in *ia/emu a cantai* 'I would sing'. lit. 'I had to sing'. In Logudorese, the conditional *dia cantare* is formed by a now opaque past form of the verb *dèpere*, derived from Latin DEBĒBAM, directly followed by the infinitive (Pisano, 2009).

It must be noticed that in Sardinian, as in Romanian but differently from other Romance languages, in both future and conditional, the auxiliary precedes the verb and it is clearly detached from it. In other words, while in Italian or Spanish the forms *canterò* and *cantarè* have inglobated the original auxiliary in the verb ending, in Sardinian and Rumenian forms, the auxiliary is still clearly recognizable. From the annotation perspective, this means that these are treated as analytic periphrastic construc-

tions; in these cases, tense-mood inflections are not directly encoded in the morphological features of the lexical verb. Instead, the auxiliary (e.g. *ai* 'have' or the past form of *dèpere* 'have to') bears the relevant features (e.g. `Mood=Ind`, `Tense=Pres` or `Imp`, `Person`, `Number`), while the infinitival verb simply includes the `VerbForm=Inf` feature.

The gerund is highly productive in Sardinan, and can be used both after a verb of perception and as a modifier of a noun phrase (Cuzzolin, 2005).

For all other morphological features, including the language-specific feature `Clitic=Yes`, we aligned our annotation choices with the conventions adopted in the UD Italian treebank, due to the structural proximity between Sardinian and Italian.

With respect to part-of-speech tagging, Sardinian aligns straightforwardly with the UPOS inventory and makes use of all 17 universal tags. The mapping between traditional grammatical categories and UD labels is largely unproblematic and reflects the profile of the language as a Romance system.

The tag `AUX` is assigned to functional verbs that contribute tense, aspect, voice, or modality without introducing an independent lexical predicate. This includes the copular verb *èssere* 'to be'; the auxiliaries *èssere* and *àere* 'to have' in compound tenses and passive constructions; *dèpere* 'must' in the analytic formation of the conditional; èssere in progressive constructions; and modal verbs such as *pòdere* ('can'), *dèpere* ('must'), and *chèrrere*$_{src}$/*bòlli(ri)*$_{sro}$ ('want') when they function as modal auxiliaries rather than as full lexical verbs. When these verbs introduce their own argument structure, they are tagged as `VERB`, following a distributional criterion.

The tag `DET` is used for elements occupying the determiner position within the noun phrase, including articles, demonstratives, indefinites, and exclamatives when they modify a nominal head. Two noteworthy cases concern the pronominal use of the form *su*, that is analogous to Spanish *lo*, and possessive adjectives. Regarding the former, when *su* functions as a nominal head, it is annotated as `PRON` with `PronType=Dem`, due to its deictic nature and syntactic autonomy. Possessive adjectives, instead, although semantically related to determiners, are obligatorily postnominal and have full adjectival agreement. Their distribution therefore aligns with the adjectival class rather than with determiners and are annotated as `ADJ`. To encode their possessive semantics, the feature `Poss=Yes` is added.

Finally, the tag `PART` is marginal in the current dataset. The only attested instance is *nanca*, a discourse particle coming from *narant+ca* ('they say that') conveying reported, not directly witnessed information (Cruschina and Remberger, 2008; Brunelli, 2013; Loporcaro and Pisano, 2021). In our treebank *nanca* occurs once in the example *su mere non mi cheret prus ca nanca no apo fattu bona guardia* 'the lord does not want me anymore since he says that I did not stand guard well'. Its annotation as `PART` reflects its clause-level pragmatic function and lack of argument structure, distinguishing it from adverbs or complementizers.

Overall, the POS tagging and morphological annotation of Sardinian demonstrate a high degree of compatibility with the UD framework, even in case of language-specific properties that require a suitable encoding.

4.4. Syntax

From a typological point of view, Sardinian, as most Romance varieties, is an SVO, pro-drop language, showing nominative-accusative alignment (Putzu, 2017). Determiners precede the noun, while all the modifiers, including possessives, follow the noun. Moreover, Sardinian, like other Romance languages of the southern area, shows differential object marking, which is driven by syntactic-semantic parameters of the object, as its animacy, specificity, and definiteness (see the example in Figure 1).

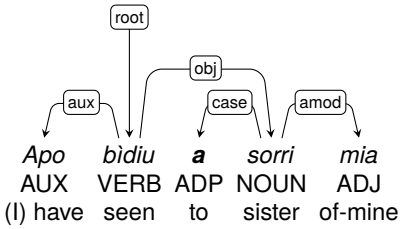

Figure 1: Example in Campidanese of direct object introduced by preposition. English translation: 'I've seen my sister'.

It also features some specific properties that are particularly salient from a structural perspective. Among these are the case, already described in Section 4.3 of *su* as a pronominal head, as well as reduplication strategies used as intensifiers. These cases required explicit annotation choices, partly drawing inspiration from other UD language-specific guidelines.

Concerning the pronominal use of the form *su*, we already introduced in Section 4.3 the case where it does not modify a noun, but rather stands independently. Accordingly, it functions as syntactic head, either governing a subordinate clause, typically a relative, or a nominal modifier. The example in Figure 2 shows the former instance, with the finite clause - introduced by the relative pronoun *chi* - attached to *su* with the relation `acl:relcl`. Figure 3 instead shows the latter case, where the phrase *de Aragoni* 'of Aragon' is attached as nominal modifier of the pronoun according to standard UD practice.

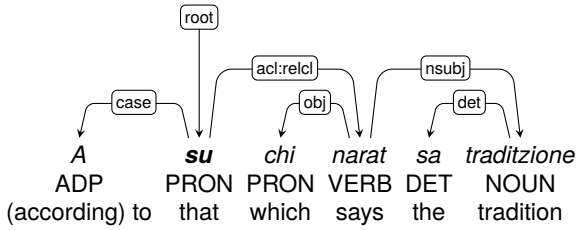

Figure 2: Example in LSC of pronominal *su* governing a relative clause. English translation: 'according to the tradition'.

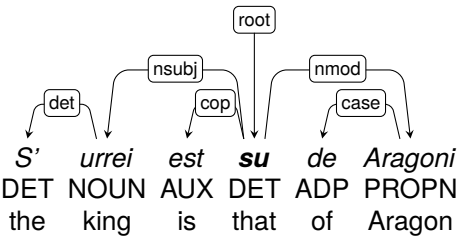

Figure 3: Example in Campidanese of pronominal *su* governing a noun. English translation: 'the king is that of Aragon'

As in other Mediterranean languages, reduplication can be used in Sardinian for intensive purposes and is a productive and structurally regular strategy (Stolz, 2003). It generally involves adjectives (e.g. *mannu mannu* (lit. 'great great' → 'very great'), adverbials (*abbellu abbellu*, 'very slowly'), but also nouns, particularly in spatial constructions (*caminai muru muru*$_{sro}$ 'to walk along a wall', referring to the pathway along which the movement is performed), and finally with verbal forms (*canta canta*, 'singing continuously') (Floričič, 2012). Putzu (2017) considers these cases as "reduplicative compounds". Accordingly, the general annotation criterion we adopted in UD is to treat this phenomenon as a case of compounding rather than modification or coordination, also following the example of other UD treebanks where reduplication serves a similar function, such as in Greek[12] or Turkish[13]. However, contrary to the latter two, and consistently with the head-initial principle adopted for determining the syntactic governor in Sardinian, the head is always the first element of the compound.

5. Annotation Methodology

Following the preliminary selection of the texts, the annotation process began with an initial automatic pass over the raw data. Since no parser specifically trained on Sardinian was available, it was not possible to rely on a language-specific model. Instead, we employed an existing model trained on the Italian ISDT treebank (Simi et al., 2014), motivated by the typological and genealogical proximity between Italian and Sardinian, as well as by partial overlap in morphosyntactic structures. In fact, this choice was intended as a bootstrapping strategy rather than as a fully reliable annotation solution. As expected, the quality of the parser's output was extremely low, and errors affected all linguistic levels. Nonetheless, this preliminary step proved useful in providing a first segmentation of the texts and a basic tokenization that would align with UD principles, thus facilitating the subsequent revisions.

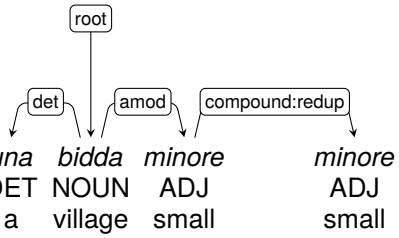

Figure 4: Example in Nuorese of reduplication. English translation: 'a very small village'.

The core of the annotation work was therefore carried out manually, using Arborator-Grew as graphical interface (Guibon et al., 2020). Two native-speaker annotators were involved, one speaker of Campidanese and the other of Nuorese. In light of their respective linguistic competence and variety-specific expertise, the annotation tasks were distributed accordingly: the Campidanese annotator was primarily responsible for texts belonging to the Campidanese variety, while the Nuorese annotator worked on Logudorese texts, due to the greater structural and lexical affinity between Logudorese and Nuorese compared to Campidanese. The corpus section comprising texts in LSC was instead reviewed jointly by both annotators. However, due to the very limited number of LSC sentences included in the corpus so far, it was considered premature to compute inter-annotator agreement at this stage. Such evaluation is planned for the next phase of the treebank development, once the LSC section will be sufficiently expanded to provide more meaningful results.

For validation purposes, a third annotator was involved, in order to fix issues concerning in particular inconsistencies in feature or dependency annotations. Finally, after configuring the language-specific settings, the official validation script was used to make sure the annotated data was fully compliant with UD requirements.

[12] https://universaldependencies.org/el/dep/compound-redup.html

[13] https://universaldependencies.org/tr/dep/compound-redup.html

Although the manual annotation was conducted independently by the annotators, the process was accompanied by frequent coordination meetings. These sessions served to discuss problematic cases, refine annotation decisions, and progressively define and revise the project's annotation guidelines.

6. NLP Pipeline Training and Evaluation

We used our treebank to train an NLP pipeline for Stanford Stanza (Peng et al., 2020). Stanza's NLP pipeline is made of the following models: sentence splitter and tokenizer, multi-word token extender, lemmatizer, parts-of-speech (POS) and morphological tagger (UFeats) and dependency parser. Training the last two models in Stanza requires word embeddings. We obtained word embeddings from the FastText project[14] in the form of word vectors extracted from the Sardinian version of Wikipedia.

To train the parser, we applied a stratified k-fold method for data sampling. This approach ensured both a balanced distribution of the three varieties across the training and test sets and allowed us to verify - despite the limited data size - the robustness of our results. The treebank was partitioned to obtain a training set of 155 sentences and a test set of 39 sentences, with the varieties in each split distributed approximately as follows: 71% Logudorese, 20% Campidanese, and 9% LSC. We used 5-fold cross-validation as the training method. Stanza was trained from scratch for each step of the pipeline, from tokenization to parsing, using the hyperparameter configurations suggested on the related GitHub page. Table 2 reports the results obtained, using the metrics developed for the CoNLL 2018 Shared Task (Zeman et al., 2018), macro-averaged across all folds. Table 3 shows instead a detailed account of the averaged F1 score obtained on each dependency relation.

The results show performance that is generally consistent with expectations for a parser trained on a limited-size dataset, demonstrating good quality in the more superficial aspects of linguistic analysis and progressively poorer performance as structural complexity increases.

Overall, the tokenizer performed very well (F1 = 99.46), while we observe lower - though still acceptable -results with sentence splitter and the lemmatizer. The lower results with the former are mainly attributed to the nature of the texts included in the Logudorese section, where the presence of several parentheticals and direct speech make sentence splitting more challenging. As regards lemmatization, despite the higher sparsity due to the presence different varieties, performance results are reasonable, but also more susceptible to change depending on the fold used for training and evaluation.

As for the POS and morphological tagger, we did not develop a language-specific POS tagset (XPOS) for our corpus. The trained UPOS model performs well (F1 = 87.68), an so does the morphological tagger (F1 = 81.89), although the latter - similarly to lemmatization - shows a greater sensitivity to the fold used. When considering the correctness of both UPOS and UFeats, the performance of the model decreases significantly (F1 = 75.23), suggesting errors distributed across the various components of the annotation.

The dependency parser has a fair understanding of the basic Sardinian syntactic structure, identifying the head-modifier relation in about 70% of tokens (Unlabeled Attachment Score = 70.36). However, the model is able to assign the correct syntactic relation only for approximately the 59% of tokens (Labeled Attachment Score = 59.58). If we exclude functional syntactic relations such as auxiliaries, copulas, adpositional markers and determiners (see also Table 3), performances are even worse (Content Label Score = 47.42). The model is particularly poor at identifying relations between clauses (relative and adverbial clauses) and performs sub-optimally in identifying the clausal structure, thus in properly finding subject, object, and obliques relations.

The lowest performance is achieved when lemmas and morphological features are included in the evaluation of the syntactic parser. The Bi-Lexical dependency score is identical to the Content Label Score (F1 = 47.45), meaning that the syntactic parser has a poor understanding of non-functional, meaningful syntactic relations. Finally, although the morphological tagger is fairly robust, the Morphology-Aware Labeled Attachment Score is the lowest of all scores (F1 = 43.37). This confirms that the model struggles to capture structure, labels, and morphological information simultaneously, as would be expected under conditions of data scarcity. However, the relatively low standard deviations across all these metrics suggest that the model is quite robust with respect to the cross-validation splits.

The analysis by dependency label provides further insights. The most frequent and locally determinable relationships, such as `det` (90.45%), `aux` (80.38%), and `case` (80.12%), achieve high scores, confirming that the model effectively learns simple and regular syntactic patterns. Labels such as `cc` and `root` also maintain decent performance. In contrast, more complex or less frequent relations, such as `ccomp`, `obl`, `conj`, and `advcl`, show a

[14] https://fasttext.cc/docs/en/crawl-vectors.html

marked decline in performance (F1 often below 40%), showing the difficulties in modeling long-distance , coordination and subordinate structures. The extremely low or zero performance for the remaining labels is clearly due to the significant imbalance in the dataset and insufficient coverage during training, resulting from the low frequency of these relations.

	Avg. (± st.dev)
Tokens	99.46 (± 0.20)
Sentences	78.68 (± 1.56)
Lemma	81.33 (± 1.76)
UPOS	87.68 (± 0.74)
Ufeats	81.89 (± 1.77)
All tags	75.23 (± 1.83)
UAS	70.36 (± 1.64)
LAS	59.58 (± 2.51)
CLAS	47.42 (± 2.57)
MLAS	43.37 (± 1.74)
BLEX	47.42 (± 2.57)

Table 2: Average scores (with standard deviation) over the five folds.

7. Conclusions

In this paper, we introduced the preliminary stages of the development of ContSar, the first treebank for contemporary Sardinian. In particular, we outlined the main linguistic feature of Sardinian, highlighting some of its peculiarities, and describing how such features were mapped into the UD formalism. The annotation process has shown that UD is fully compatible with the structural properties of Sardinian, including its analytic constructions and pronominal system. This confirms the suitability of UD for representing the language and enables direct comparison with other treebanks built under the same framework.

As next step, we plan to expand the corpus with additional texts, following, as done in the previous stages, an iterative process of automatic bootstrapping with the latest trained model, manual revision and further parser improvement.

We intend to release the first batch of annotated data for the treebank to be openly available with the upcoming official UD release on May, 2026.

8. Limitations

The definition of the annotation guidelines followed a mixed approach. On the one hand, it was developed top down, taking into account the existing normative descriptions of Sardinian. On the other hand, it was refined bottom up, on the basis of the linguistic phenomena observed in the pilot sample and the need to find a compatible UD-based counterpart. However, the pilot dataset is very small and does not ensure adequate coverage of the range of possible constructions. As a result, some annotation choices may need to be revised as new data are included.

deprel	**F1 (± st.dev)**
det	90.45 (± 0.03)
aux	80.38 (± 0.08)
case	80.12 (± 0.05)
cc	74.43 (± 0.06)
root	71.69 (± 0.06)
flat	68.88 (± 0.12)
expl	59.42 (± 0.16)
amod	59.20 (± 0.09)
nmod	56.42 (± 0.04)
nummod	55.14 (± 0.16)
nsubj	54.55 (± 0.04)
punct	53.61 (± 0.05)
cop	50.44 (± 0.18)
advmod	48.17 (± 0.10)
obj	47.30 (± 0.07)
iobj	39.28 (± 0.15)
mark	39.13 (± 0.12)
ccomp	34.67 (± 0.33)
obl	33.37 (± 0.19)
conj	32.94 (± 0.03)
discourse	31.25 (± 0.24)
parataxis	29.59 (± 0.23)
flat:name	26.15 (± 0.28)
xcomp	23.22 (± 0.19)
acl:relcl	16.48 (± 0.11)
fixed	16.29 (± 0.11)
acl	12.45 (± 0.15)
advcl	9.37 (± 0.07)
appos	8 (± 0.11)
compund:redup	8 (± 0.18)
compound	0
csubj	0
dislocated	0
nsubj:pass	0
obl:agent	0
orphan	0
vocative	0

Table 3: Average F1 score (with standard deviation) for each dependency label.

The limited size of the dataset, although largely due to the scarcity of freely accessible sources, also affects the overall coverage and balance of the three main varieties included in the treebank. The current distribution does not yet guarantee a fully representative diatopic and structural coverage.

For these reasons, the extension of the dataset is not only a quantitative objective but also a methodological necessity. Increasing the size and diversity of the corpus will allow us to test and refine the

guidelines, improve balance across varieties, and strengthen the overall consistency of the resource.

9. Bibliographical References

Michele Brunelli. 2013. An evidential marker in Sardinian: nanca in Santa Maria Navarrese. In *Atti della XVIII Giornata di dialettologia. Quaderni di lavoro delASIt*, volume 16.

S.M. Carta, F. Concas, G. Fenu, A. Giuliani, M.M. Manca, P. Mura, and S. Pisano. 2025. A BERT-based Approach for Part-of-Speech Tagging in the Low-Resource Context of Sardinian. In *Proceedings of the Eleventh Italian Conference on Computational Linguistics (CLiC-it 2025)*, Cagliari, Italy.

Ilaria Chizzoni and Alessandro Vietti. 2024. Towards an ASR system for documenting endangered languages: A preliminary study on Sardinian. In *Proceedings of the Tenth Italian Conference on Computational Linguistics (CLiC-it 2024)*, pages 214–220, Pisa, Italy. CEUR Workshop Proceedings.

Silvio Cruschina and Eva-Maria Remberger. 2008. Hearsay and reported speech. Evidentiality in Romance. *Rivista di Grammatica Generativa*, 33:95–116.

Pierluigi Cuzzolin. 2005. Some remarks on the gerund in Sardinian. *Sprachtypologie und Universalienforschung*, 58(2/3):176–187.

Domenico De Cristofaro, Alessandro Vietti, Marianne Pouplier, and Aleese Block. 2025. When less is more? diagnosing ASR predictions in sardinian via layer-wise decoding. In *Proceedings of the Eleventh Italian Conference on Computational Linguistics (CLiC-it 2025)*, pages 363–370, Cagliari, Italy.

Antonietta Dettori. 2002. La Sardegna. In Manlio Cortelazzo, editor, *I dialetti italiani. Storia struttura uso*, pages 897–958. UTET, Torino.

Franck Floričič. 2012. On reduplicated imperatives in Sardinian. *Lingue e linguaggio*, 11(1):71–96.

Gianfranco Fronteddu, Hèctor Alòs i Font, and Francis M. Tyers. 2017. Machine translation from Catalan to Sardinian: a translation tool for a language in the process of standardisation. *Linguamatica*, 9(2):3–20.

Kim Gerdes. 2013. Collaborative dependency annotation. In *Proceedings of the Second International Conference on Dependency Linguistics (DepLing 2013)*, volume 88-97.

Giulia Grosso, Giulia Murgia, and Antonietta Marra. 2024. Certifying a minority language: a first description of the Sardinian case. *MOSAIC*, 15:105–121.

Gaël Guibon, Marine Courtin, Kim Gerdes, and Bruno Guillaume. 2020. When collaborative treebank curation meets graph grammars. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 5293–5302, Marseille, France. European Language Resources Association.

Sandra Hajek and Hans Goebl. 2021. Le strutture profonde del dominio linguistico sardo: Un'analisi dialettometrica. In *Actes du XXIXe Congrès international de linguistique et de philologie romanes, Copenhague 1er–6 juillet 2019. Section 7. Dialettologia e geolinguistica medievale e moderna (Europa e fuori dall'Europa)*, volume 2, pages 979–991, Strasbourg. Société de linguistique romane.

Michele Loporcaro and Simone Pisano. 2021. I complementatori chi e ca in Sardegna centrale:estensione areale e dinamiche di variazione. In *Actes du XXIXe Congrès international de linguistique et de philologie romanes, Copenhague 1er–6 juillet 2019. Section 7. Dialettologia e geolinguistica medievale e moderna (Europa e fuori dall'Europa)*, volume 2, pages 993–1005, Strasbourg. Société de linguistique romane.

Michele Loporcaro and Ignazio Putzu. 2024. Sardinian. In *Oxford Encyclopedia of Romance Linguistics*, pages 1–42. Oxford University Press, Oxford.

Guido Mensching. 2017. Morfosintassi: sincronia. In *Manuale di linguistica sarda*, pages 376–396. Mouton de Gruyter, Berlin.

P. Mura, S. Carta, A. Giuliani, and M. Manca. 2023. The corpus of Sardinian emigrants:a tool for a quantitative approach to contact phenomena. In *Minority Languages in European Societies International Conference. Book of Abstracts*.

Anna Oppo. 2007. *Le lingue dei sardi: una ricerca sociolinguistica*. Rapporto finale, Regione Autonoma della Sardegna.

Qi Peng, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Cristopher Manning. 2020. Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. In *ACL2020 System Demonstration*.

Simone Pisano. 2009. I futuro e il condizionale analitici in alcune varietà sarde moderne: genesi di marche grammaticali da forme verbali lessicalmente piene. *Bollettino di studi sardi*, 2:147–166.

Mario Puddu. 2015. *Ditzionàriu de sa limba e de sa cultura sarda*, 2 edition. Condaghes, Cagliari.

Nicoletta Puddu. 2005. Reflexives and intensifiers in Sardinian. *Sprachtypologie und Universalienforschung*, 3:246–261.

Ignazio Putzu. 2015. 2. Sardinian. In Konstanze Jungbluth and Federica Da Milano, editors, *Manual of Deixis in Romance Languages*, pages 45–58. De Gruyter, Berlin, München, Boston.

Ignazio Putzu. 2017. Tipologia del sardo. In Eduardo Blasco Ferrer, Peter Koch, and Daniela Marzo, editors, *Manuale di linguistica sarda*, pages 303–319. de Gruyter, Berlin.

Alan Ramponi. 2024. Language varieties of Italy: Technology challenges and opportunities. *Transactions of the Association for Computational Linguistics*, 12:19–38.

RAS Regione Autonoma della Sardegna. 2006. Limba sarda comuna. Norme linguistiche di riferimento a carattere sperimentale per la lingua scritta dell'Amministrazione regionale.

RAS Regione Autonoma della Sardegna. 2022. Certificazione provvisoria sperimentale della conoscenza delle lingue di minoranza storica parlate in Sardegna. Criteri ortografici orientativi per la lingua sarda.

Thomas Stolz. 2003. A New Mediterraneanism: Word Iteration in an Areal Perspective. *Mediterranean Language Review*, 15(4):1–47.

Francis M. Tyers, Hèctor Alòs i Font, Gianfranco Fronteddu, and Adrià Martìn Mor. 2017. Rule-based machine translation for the Italian-Sardinian language pair. *The Prague Bulletin of Mathematical Linguistics,*, 108:221–232.

Maurizio Virdis. 1978. *Fonetica del sardo campidanese*. Edizioni della Torre, Sassari.

Maurizio Virdis. 1988. Sardisch: Areallinguistik. In Günter Holtus, Michael Metzeltin, and Christian Schmitt, editors, *Lexikon der Romanistischen Linguistik*, volume IV, pages 897–913. Niemeyer, Tübingen.

Maurizio Virdis. 2007. Tipologia e collocazione del sardo tra le lingue romanze. In *Atti del convegno di Santa Cristina di Paulilatino, 6– 8 dicembre 2001*, pages 141–152, Cagliari. Condaghes.

Daniel Zeman, Jan Hajič, M. Straka, D.Q. Nguyen, M. Potthast, E. Stepanov, G. Di Fabbrizio, N. Promrit, and M. Popel. 2018. CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 1–21.

10. Language Resource References

Caterina Maria Cappello, Sabrina D'Alì, Mario Guglielmetti, Elisa Di Nuovo, and Cristina Bosco. 2025. Arbuli sunnu: A Sicilian-Italian parallel treebank. In *Proceedings of the Eleventh Italian Conference on Computational Linguistics (CLiC-it 2025)*, pages 155–168, Cagliari, Italy. CEUR Workshop Proceedings.

Maria Fortunato and Sara Ravani. 2015. L'informatica al servizio della filologia e della linguistica sarda: il corpus ATLiSOr (archivio testuale della linguasarda delle origini). *Bollettino di studi sardi*, 8:53–90.

Simone Pisano, Valentina Piunno, and Vittorio Ganfi. 2019. Appunti per un corpus di sardo multimediale. In Daniela Marzo, Valentina Piunno, and Simone Pisano, editors, *Per una pianificazione del plurilinguismo in Sardegna*, pages 147–164. Condaghes, Cagliari.

Nicoletta Puddu and Achim Stein. 2018. Word-level and higher level annotation of the Sardinian Medieval Corpus. In *Proceedings of the Second Workshop on Corpus-Based Research in the Humanities*, Vienna. Gerastree Proceedings.

Nicoletta Puddu and Luigi Talamo. 2020. EModSar: A Corpus of Early Modern Sardinian Texts. ISBN: 978-88-942535-4-2 Pages: 210–215 Publication Title: Atti del IX Convegno Annuale dell'Associazione per l'Informatica Umanistica e la Cultura Digitale (AIUCD). La svolta inevitabile: sfide e prospettive per l'Informatica Umanistica.

Maria Simi, Cristina Bosco, and Simonetta Montemagni. 2014. Less is more? Towards a reduced inventory of categories for training a parser for the Italian Stanford Dependencies. In *Language Resources and Evaluation 2014*, pages 83–90. European Language Resources Association (ELRA).