Verso la creazione di una Treebank per il sardo

NICOLETTA PUDDU[1], MANUELA SANGUINETTI[1], LUIGI TALAMO[2]
[1]UNIVERSITÀ DEGLI STUDI DI CAGLIARI
[2]UNIVERSITY OF SAARBRÜCKEN

ABSTRACT (ITALIANO)

In questo contributo intendiamo presentare una prima proposta per la creazione di una treebank per il sardo secondo il framework delle Universal Dependencies (UD). Verranno illustrate le considerazioni teoriche introduttive, i principi alla base della selezione dei testi, le fasi di tokenizzazione e annotazione e i primi risultati di training di un annotatore sintattico. Delineeremo quindi i passi futuri per l'estensione dell'annotazione.
Parole chiave: Universal Dependencies, sardo, lingue di minoranza, annotazione morfosintattica

ABSTRACT (ENGLISH)

This paper presents an initial proposal for the development of a Sardinian treebank within the Universal Dependencies (UD) framework. We outline the theoretical motivations informing the design of the resource, the criteria adopted for corpus construction and text selection, and the methodological workflow implemented for tokenisation and morpho-syntactic annotation. The contribution also reports preliminary evaluation results from the first training cycle of a UD-compliant syntactic parser, providing insights into the feasibility and future directions of large-scale syntactic annotation for Sardinian.
Keywords: Universal Dependencies; Sardinian; Less resourced languages; POS-tagging

1. INTRODUZIONE

In questo contributo intendiamo presentare una prima proposta per la creazione di una *treebank* per il sardo secondo il sistema di annotazione delle Universal Dependencies (UD). Nonostante esistano infatti diverse treebank per le lingue romanze, non ne esistono per il sardo e, più in generale, sono ancora poche quelle per lingue romanze minoritarie, in particolare se prive di standardizzazione. Descriveremo quindi le prime fasi di una creazione di una treebank per il sardo, evidenziando le principali criticità incontrate e delineando soluzioni possibilmente estensibili anche ad altre varietà di minoranza.

2. LA LINGUA SARDA E LE SUE VARIETÀ

Il sardo è una lingua romanza parlata nell'isola di Sardegna tradizionalmente suddivisa in quattro varietà: campidanese, logudorese, nuorese e arborense (Virdis, 1988). Tali varietà sono però riconducibili a due macrovarietà divise da un fascio di isoglosse in molti casi sovrapponibili. Le due macrovarietà vengono classificate come varietà distinte in ISO 639-3 con le denominazioni di campidanese (src), varietà del centro-sud, e logudorese (sro), varietà del centro-nord. Altre due varietà che ISO 639-3 identifica come sarde (gallurese e sassarese) non sono in realtà ritenute in letteratura come appartenenti al sardo, ma piuttosto affini a varietà corse e toscane e non sono state pertanto considerate ai fini della costituzione della nostra Treebank.
Dal 2006 esiste inoltre una varietà scritta per i documenti in uscita dall'Amministrazione regionale, denominata Limba Sarda Comuna (LSC) elaborata dalla Regione autonoma della Sardegna. La stessa Regione, nel 2022 ha promulgato, nell'ambito della Certificazione sperimentale della lingua sarda, dei criteri ortografici orientativi per la lingua sarda.

3. STATO DELL'ARTE

La scarsità di risorse annotate è un problema comune alle varietà di minoranza in Italia. Segnaliamo comunque che per il ligure[1], il napoletano[2] e il siciliano[3] sono state realizzate e rese disponibili nel repository di UD delle treebank.
Non esistono al momento corpora di riferimento in lingua sarda, nemmeno non annotati. Risulta attualmente in preparazione un corpus di parlato di emigranti sardi (Mura et al., 2023) sul quale è stato creato un *parts-of-speech* (POS) *tagger* automatico per il sardo che utilizza un *language model* neurale basato su BERT (Carta et al., 2025). Esistono inoltre alcune risorse per le varietà storiche: segnaliamo l'archivio testuale ATLISOR (Fortunato & Ravani, 2015) un corpus non annotato che raccoglie testi di età medioevale, e il corpus di testi sardi di epoca moderna EModSar (Puddu & Talamo, 2020), attualmente annotato per POS. Primi tentativi di annotazione sia per POS che sintattica erano stati fatti per il sardo moderno da Puddu & Stein (2018) tramite il software Arborator (Gerdes, 2013).

4. LA CREAZIONE DEL CORPUS

La prima scelta che abbiamo dovuto operare nella ideazione della treebank è stata quella della varietà di sardo da utilizzare per il corpus di partenza. Ciò che qui interessa notare è che le differenze tra le varietà sono perlopiù di tipo fonetico e lessicale, mentre gli studiosi concordano sulla maggiore unità del sardo a livello morfosintattico. A ciò si aggiunga che la discussione sulla effettiva esistenza di due macrovarietà è piuttosto articolata, dato che, al confine tra le due varietà, è possibile individuare un'area mediana che presenta tratti di entrambe le macrovarietà (Hajek & Goebl, 2021; Loporcaro & Putzu, 2024).
La nostra ipotesi di lavoro è pertanto che sia possibile creare un annotatore a livello morfosintattico e un'unica treebank per tutte le varietà di sardo.
Per quanto riguarda la scelta dei testi, nella creazione di corpora di lingue minoritarie si fa spesso largo uso del web, anche quando, come nel caso del sardo, siano disponibili testi a stampa. Il sardo è documentato tra le lingue utilizzate sul web da W3CTechs, anche se, secondo una ricerca condotta da Soria et al. (2018), i parlanti non sono solitamente a conoscenza di media e servizi in sardo, anche quando questi siano disponibili.
Uno dei problemi nel caso di lingue non standardizzate, è quello di identificare la subvarietà alla quale i testi appartengono in modo da creare un corpus per quanto possibile bilanciato.
A questo proposito, le pagine di Wikipedia in sardo risultano particolarmente informative, dato che indicano la varietà nella quale sono redatte: limba sarda comuna (LSC), logudorese, campidanese e nuorese.
Per la creazione del corpus sono stati scelti tre testi: due pagine di Wikipedia (una in LSC e una in campidanese) e una fiaba in dorgalese, subvarietà del nuorese orientale e meridionale per un totale di 3862 token.

5. ANNOTAZIONE

Si è partiti da un'annotazione automatica dei tre testi effettuata tramite il parser Stanford Stanza (Qi et al. 2020) versione 1.10.1, usando il modello per l'italiano, addestrato su quattro treebank UD: ISDT, VIT, PoSTWITA, e TWITTIRO. Il modello–ha segmentato in frasi, tokenizzato, lemmatizzato e annotato per *parts of speech* (POS), caratteristiche morfologiche e dipendenze sintattiche. In assenza di modelli pre-addestrati per il sardo, si è scelto di utilizzare un modello disponibile per l'italiano in virtù delle similarità morfosintattiche tra le due lingue, al fine di partire da una versione pre-annotata dei testi che, pur non essendo accurata dal punto di vista dell'annotazione linguistica, rispecchiasse le specifiche del formato CoNLL-U.
Due parlanti native e linguiste hanno poi corretto in forma collaborativa l'output dell'annotazione, utilizzando la piattaforma Arborator-Grew (Guibon, 2020). La correzione è stata condotta in maniera indipendente e poi controllata per verificarne l'uniformità. Per l'attuale proposta, sono state corrette 65 frasi per un totale di 1280 token. Tokenizzazione e segmentazione in frasi non hanno avuto necessità di particolari interventi.

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

La lemmatizzazione ha invece richiesto una fase di normalizzazione. La variabilità fonetica tra le diverse varietà ha imposto il ricorso ad una normalizzazione prevalentemente ortografica, sulla base dei criteri proposti per la LSC e per la Certificazione provvisoria sperimentale della conoscenza del sardo. A livello morfologico sono state operate delle normalizzazioni, quali la riduzione delle forme di infinito in *-are* e l'uscita dei nomi maschili in *-e*. Sottolineiamo però che la normalizzazione ha riguardato unicamente il lemma e non la forma ed è fondamentale sia per la creazione di un lemmario unico, sia per future ricerche di tipo sintattico estese a tutto il sardo.
Per quanto riguarda il *POS-tagging*, il *tagset* utilizzato (cfr. tabella 1) è Universal Parts of Speech (UPOS) di Universal Dependencies (de Marneffe et al., 2021), standard ampiamente consolidato e utilizzato anche per lingue romanze di minoranza come l'occitano. Al momento, non è stata corretta l'annotazione per caratteristiche morfologiche.

Open class	Closed class	Other
ADJ	ADP	PUNCT
ADV	AUX	SYM
INTJ	CCONJ	
NOUN	DET	
PROPN	NUM	
VERB	PART	
	PRON	
	SCONJ	

Tabella 1 Il tagset per POS utilizzato nella Treebank

Anche per quanto riguarda le relazioni sintattiche, sono stati utilizzati i tipi definiti nella versione 2 di UD, indicati nella tabella 2.

	Nominals	**Clauses**	**Modifier words**	**Function words**
Core arguments	nsubj ob iobj	csubj ccomp xcomp		
Non-core dependents	obl vocative expl dislocated	Advcl	Advmod discourse	aux cop mark
Nominal dependents	nmod appos nummod	acl	Amod	det clf case
Coordination	**Headless**	**Loose**	**Special**	**Other**
conj cc	fixed flat compound	list parataxis	Orphan goeswith reparandum	punt root dep

Tabella 2. I tipi di relazione sintattica marcati nella Treebank

6. ADDESTRAMENTO DI UNA NLP PIPELINE COMPLETA PER IL SARDO

Le 65 frasi sono state utilizzate per addestrare una NLP pipeline completa di Stanford Stanza per il sardo, in maniera da poter annotare in maniera automatica ulteriori frasi e poi correggerne l'output. La pipeline comprende modelli per i seguenti *task* di *natural language processing*: tokenizzazione e segmentazione in frasi, *multi-word token expansion* (mwt), lemmatizzazione, annotazione per *part-of-speech (POS)* e caratteristiche morfologiche e annotazione sintattica. Stanford Stanza richiede l'utilizzo di *word embeddings* per gli ultimi due task. A tale scopo, abbiamo utilizzato nell'addestramento i *word*

vectors per il sardo forniti dal progetto Fasttext[4] (Bojanowski, Grave, Joulin e Mikolov 2017) ed estratti da Wikipedia.
Il set per l'addestramento è ripartito in *train*, *dev* e *test* come descritto nella seguente Tabella numero 3 (Set per l'addestramento).

Set	Token	Frasi
Train	911	52
Dev	110	6
Test	259	7

Tabella 3 – Set per l'addestramento

Come di consueto, il parser addestra il modello sulle frasi del set di *train*, aggiustando poi automaticamente gli iperparametri sul set di *dev*. Infine, una valutazione della bontà del modello è condotta sul set di *test*, utilizzando come riferimento le misure utilizzate nel task di valutazione CoNLL-U 2018 (Zeman et al., 2018).
Di seguito, discutiamo i risultati di questa valutazione per ciascun modello della pipeline.
Modello per la tokenizzazione e la segmentazione in frasi. Questo modello combina i task di divisione in token e in frasi. Il modello dimostra ottime performance nella tokenizzazione (F1 = 96.79), ma purtroppo le sue capacità di determinare l'inizio e la fine di una frase (*sentence-splitting*) risultano insufficienti (F1 = 58.82). La scarsa efficacia del sentence splitting è probabilmente dovuta anche alle caratteristiche di uno dei testi utilizzati per il training che presenta molte istanze di discorso diretto e numerosi incisi.
Modello per la *multi-word token expansion*. Il modello 'espande' (*expansion*) i token multi-parola, sia legati che intervallati da spazi bianchi, nei token costituenti, ad es. *Art Nouveau* o *Intamis chi*. Nonostante gli ottimi risultati raggiunti (F1 = 99.09), il modello non è probabilmente valutabile a causa dell'esiguo numero di token multi-parola presenti nel set di addestramento, soltanto quattro.
Modello per la lemmatizzazione. Il modello presenta buone performance (F1 = 87.27) nel ricondurre la forma al lemma.
Modello per l'annotazione per parts of speech e caratteristiche morfologiche. Il modello annota i token per UPOS e per caratteristiche morfologiche, descrivendo le categorie nominali, verbali e aggettivali del token. Nella parte relativa alle UPOS, il modello raggiunge performance discrete (78.18). I risultati sono anche discreti per l'annotazione morfologica (77.27): dobbiamo però notare che tale annotazione al momento non è stata corretta manualmente, ma riflette ancora quella del modello italiano (vedi sopra).
Modello per l'annotazione sintattica. Il modello annota i token per dipendenze sintattiche, individuando le relazioni di testa e dipendente e descrivendo la relazione sintattica. Dato l'esiguo set di annotazione, i risultati non sono purtroppo soddisfacenti, configurando questo modello come il tallone d'Achille della nostra *NLP* pipeline. Lo UAS (*Unlabeled Attachment Score*) è del 63.64%, ovvero il modello ha difficoltà nell'identificare la corretta testa sintattica, a prescindere dalla relazione sintattica, in almeno un token su tre; se prendiamo in considerazione anche la relazione sintattica (LAS: *Labeled Attachment Score*), il risultato scende a 50.91%, ovvero una volta su due il modello sbaglia ad annotare la relazione sintattica. Escludendo le parole funzionali e considerando solo aggettivi, nomi e verbi, i risultati sono ancora più deludenti: il CLAS (*Content-word Labeled Attachment Score*) è infatti del 34.43%. Infine, il MLAS (*Morpho-Syntactic Labeled Attachment Score*), una misura che combina parte del discorso, dipendenza sintattica e caratteristiche morfologiche, è del 29.51%.

7. CONCLUSIONI E PROSPETTIVE FUTURE

La creazione di risorse elettroniche per lingue di minoranza non standardizzate rappresenta una sfida in quanto, innanzi tutto, impone delle scelte nelle varietà da utilizzare sia per il corpus di partenza che per l'addestramento. Nel caso del sardo, abbiamo deciso di utilizzare diverse subvarietà per il corpus di partenza, normalizzando le forme a livello di lemma. La scelta appare per ora corretta, in quanto il modello presenta buone performance nel ricondurre forme appartenenti a subvarietà distinte ad un unico lemma. Nel prossimo futuro ci proponiamo di estendere l'addestramento del modello al maggior numero possibile di subvarietà.

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

Anche a livello di POS-tagging, il modello dimostra livelli di correttezza discreti, che potranno essere sicuramente migliorati tramite l'aggiunta delle feature morfologiche all'annotazione.
Il modello risulta decisamente meno performante nell'analisi sintattica. Uno dei motivi, come detto, potrebbe essere l'esiguo numero di frasi sinora annotate, per cui ci riserviamo di rivalutare gli esiti una volta incrementato tale numero. Condurremo inoltre un'analisi di tipo qualitativo sulla tipologia degli errori, per verificare se una percentuale non possa essere ricondotta a particolari strutture sintattiche del sardo non riconosciute dall'annotatore. L'obiettivo è quello di pubblicare la treebank in una release ufficiale di UD, presumibilmente a maggio 2026.

BIBLIOGRAFIA

Bojanowski, P., Grave, E., Joulin, A., Mikolov, T. (2017) Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics*, 5 (pp. 135-146). https://doi.org/10.48550/arXiv.1607.04606.

Carta, S. M., Concas, F., Fenu, G., Giuliani, A., Manca, M. M., Mura, P., & Pisano, S. (2025). A BERT-based Approach for Part-of-Speech Tagging in the Low-Resource Context of Sardinian. In C. Bosco, E. Jezek, M. Polignano, & M. Sanguinetti (A c. Di), *CLiC-it 2025 Italian Conference on Computational Linguistic*. https://aclanthology.org/2025.clicit-1.18/.

de Marneffe, M.-C., Manning, C., Nivre, J., & Zeman, D. (2021). Universal Dependencies. *Computational Linguistics*, *47*(2), 255–308. https://doi.org/10.1162/coli_a_00402.

Fortunato, M., & Ravani, S. (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.

Gerdes, K. (2013). Collaborative dependency annotation. *Proceedings of the Second International Conference on Dependency Linguistics (DepLing 2013)*, *88–97*. https://aclanthology.org/W13-3711/.

Guibon, G. (2020). When Collaborative Treebank Curation Meets Graph Grammars. Arborator With a Grew Back-End. *Proceedings of the 12th Conference on Language Resources and Evaluation*, 5291–5300. https://aclanthology.org/2020.lrec-1.651/.

Hajek, S., & Goebl, H. (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)* (Vol. 2, pp. 979–991). Société de linguistique romane.

Loporcaro, M., & Putzu, I. (2024). Sardinian. In *Oxford Encyclopedia of Romance Linguistics* (pp. 1–42). Oxford University Press. https://dx.doi.org/10.1093/acrefore/9780199384655.013.725.

Mura, P., Carta, S., Giuliani, A., & Manca, M. (2023). *The corpus of Sardinian emigrants:a tool for a quantitative approach to contact phenomena,*. MiLES: Minority Languages in European Societies - International Conference-Turin, Torino/Bard.

Peng, Q., Qi, P., Zhang, Yuhao, Zhang, Yuhui, Bolton, J., & Manning, C. (s.d.). Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. *ACL2020 System Demonstration*. https://doi.org/10.18653/v1/2020.acl-demos.14.

Puddu, N., & Stein, A. (2018). Word-level and higher level annotation of the Sardinian Medieval Corpus. *Proceedings of the Second Workshop on Corpus-Based Research in the Humanities*.

Puddu, N., & Talamo, L. (2020). EModSar: A Corpus of Early Modern Sardinian Texts. In *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* (pp. 210–215). Associazione per l'Informatica Umanistica e la Cultura Digitale. https://doi.org/10.6092/unibo/amsacta/6316. https://doi.org/10.6092/unibo%2Famsacta%2F6316.

Regione autonoma della Sardegna. (2006). *Limba sarda comuna. Norme linguistiche di riferimento a carattere sperimentaleper la lingua scritta dell'Amministrazione regionale*. https://www.regione.sardegna.it/documenti/1_72_20060418160308.pdf.

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*. https://www.regione.sardegna.it/documenti/1_38_20220721131659.pdf.

Soria, C., Quochi, V., & Russo, I. (2018). The DLDP Survey on Digital Use and Usability of EU Regional and Minority Languages. *Proceedings of the Eleventh International Conference on Language Resources and Evaluation*. LREC 2018. https://aclanthology.org/L18-1656/.

Virdis, M. (1988). Sardisch: Areallinguistik. In G. Holtus, M. Metzeltin, & C. Schmitt (A c. Di), *Lexikon der Romanistischen Linguistik: IV* (pp. 897–913). Niemeyer.

Zeman, D., Hajič, J., Straka, M., Nguyen, D. Q., Potthast, M., Stepanov, E., Di Fabbrizio, G., Promrit, N., & Popel, M. (2018). CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies. *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, 1–21. https://doi.org/10.18653/v1/K18-2001.