Arbuli sunnu: a Sicilian-Italian Parallel Treebank

Caterina Maria Cappello[1,†], Sabrina D'Alì[1,†], Mario Guglielmetti[1,†], Elisa Di Nuovo[2,*] and Cristina Bosco[1]

[1]Università di Torino, Dipartimento di Informatica, Corso Svizzera 185, Torino, 10149, Italia
[2]European Commission, Joint Research Centre (JRC), Via Enrico Fermi, 2749, Ispra (VA), 21027, Italia

Abstract

The Natural Language Processing (NLP) community has recently begun to engage with endangered languages and dialects which encode culturally different perspectives and local knowledge. Regardless of the usefulness and applicability of NLP tools for such languages, creating resources for dialects increases our knowledge of them, encourages the community to study them further, and supports the preservation of an important heritage. As part of this endeavour, we are focussing on Sicilian, a dialect spoken in Sicily, with a rich cultural history. Sicilian preservation is crucial to maintaining Southern Italy's linguistic diversity. In this paper, we present the first release of a novel treebank called Sicilian3bank. On the one hand, to improve the usability of this resource and provide access to non-Sicilian speakers, all sentences are linked to their translation into Italian, resulting in a 1:1 parallel resource. On the other hand, by applying the Universal Dependencies format, a widely used standard for the annotation of treebanks, we pave the way for data-driven cross-linguistic research. We hope that this work can serve as a basis for further linguistic research and computational applications for the Sicilian dialect.

Keywords

Sicilian, treebank, parallel texts, Universal Dependencies, translation

1. Introduction

Recent developments in generative Artificial Intelligence (genAI) have increasingly highlighted the importance of taking more into account a larger variety of the languages spoken in the world. Developing tools and resources to deal with a language has meaningful effects, among which the most important is an improvement of the awareness of the underlying cultural heritage, an aspect that can be crucial for the achievement of better performances by Large Language Models (LLMs) in several tasks.

According to [1], the world's living languages can be categorised into 500 institutional languages and a further 6,500 local vernaculars, or oral languages. While institutional languages feature standardised orthographies and widespread literacy, the local languages include ancestral languages, with an unbroken history of oral transmission, and languages in danger of disappearing. Most Natural Language Processing (NLP) tools and resources developed until now are almost only for institutional languages, since only in the very last years the NLP community has begun to engage with local and endangered languages. Therefore the challenges to address are still many.

In this paper, we focus on the first steps of developing a resource for one of the most spoken Italian dialects, which is featured in a long tradition of studies in linguistics, but not considered enough in NLP until now.[1] The aim of this study goes beyond introducing a specific novel resource and consists of starting a discussion on the challenges that can be encountered when NLP meets a dialect or a language without a standardised orthography and reference grammar.[2] Starting this discussion may be especially relevant in the context of the CLiC-it conference, since Italy is characterised by a one-of-a-kind linguistic diversity in the European landscape, where diatopic variation implicitly encodes local knowledge, cultural traditions, artistic expressions, and the history of its speakers [3]. With respect to high-resource lan-

CLiC-it 2025: Eleventh Italian Conference on Computational Linguistics, September 24—26, 2025, Cagliari, Italy.
*Corresponding author.
†These authors contributed equally.
caterina.cappello@edu.unito.it (C. M. Cappello); sabrina.dali@edu.unito.it (S. D'Alì); mario.guglielmetti@edu.unito.it (M. Guglielmetti); elisa.di-nuovo@ec.europa.eu (E. Di Nuovo); cristina.bosco@unito.it (C. Bosco)
https://www.unito.it/persone/crbosco (C. Bosco)
https://orcid.org/0000-0002-4814-982X (E. Di Nuovo); https://orcid.org/0000-0002-8857-4484 (C. Bosco)

[1]This paper has been revised for English using the LLaMa 3.3 70B model through the GPT@JRC platform, an internal JRC testbed for LLMs [2].
All cited links were last accessed on the 12th of June 2025.
Some of the reported examples have been shortened due to space constraints.

[2]In linguistics, the distinction between a language and a dialect is not always clear-cut and is often influenced by political and sociocultural factors rather than purely linguistic ones. A dialect is typically considered a regional or social variety of a language, but varieties such as Sicilian, which may lack official status or standardisation, are often labelled as dialects despite possessing many characteristics of a distinct language. For this reason, we use the terms language and dialect interchangeably when referring to Sicilian, to reflect its complex sociopolitical status.

guages, which have extensive amounts of digital data and resources available, Italian dialects are under-resourced, lacking sufficient digital representation and support.

The language observed in our study is Sicilian, as the title *arbuli sunnu* suggests, literally 'trees (they) are', showing a common predicate-initial structure. Sicilian is a vernacular language with local functions that include intergenerational knowledge transmission. The resource presented comprises diachronic and diatopic variants, enabling the analysis of linguistic changes in certain phenomena along these axes. In addition, it features orthographic variability due to the non-standardised transfer from oral to written form.

In order to make the resource accessible to a bigger audience, we provide the Italian translation in a 1:1 alignment setting. We decided to translate into Italian rather than English to underline the importance of mitigating the over-reliance towards English [4].

It is beyond the scope of this article to cover all the challenges associated with developing a treebank[3] for Sicilian; we focused mainly on the phenomena that have a major impact at (morpho-)syntactic level. By showing some of the major challenges in the treebank annotation, we hope to pave the way for the future development of an expanded resource and the discussion about the involved phenomena.

The paper is organised as follows: the next section (Sec. 2) presents an overview of related work, followed by Sec. 3, which describes the data collection and annotation process for the first release of the SICILIAN3BANK, including the translation of Sicilian sentences into Italian to create a parallel corpus. In Sec. 4 we show the parallel architecture of the treebank and the annotation methodology we followed. This section also highlights the challenges we faced developing a treebank for Sicilian. Finally, the last section (Sec. 5) is about conclusions and future work.

2. Related Work

This section provides a brief introduction to the Universal Dependencies (UD) formalism and existing parallel treebanks in UD, followed by a discussion on language variation in NLP, with a focus on dialects and, eventually, on Sicilian.

2.1. Universal Dependencies and Parallel Treebanks

UD [5] is a framework for annotating morphology and syntax consistently across languages. In recent decades, UD has become the *de facto* standard for treebanks. As its name suggests, UD represents syntax using dependency trees, instead of constituency trees. This is because dependency trees are perceived as better suited to represent free or flexible word order languages [6]. Furthermore, models using dependency representations have achieved promising results in many NLP tasks (e.g. in machine translation and information extraction) [6, p. 3].

UD comprises treebanks in more than 100 languages, including low-resource languages (see sec. 2.2 for a definition), e.g. Irish, Faroese, Uyghur. Among the UD treebanks, there are also parallel treebanks, i.e. treebanks that have been translated into other languages and subsequently annotated. The biggest effort in this respect has been done for the PUD treebank [7], which consists of 1,000 sentences in 18 languages (the majority originally in English). Translators were asked to opt for the translation which is fluent but also sharing the most grammatical features of the original. Another example of parallel treebanks in UD is ParTUT [8], which contains sentences from different domains in English, Italian and French. In ParTUT, the alignment is not 1:1 for all the sentences [9], though the texts coming from a more formal register, i.e. those from the JRC-Acquis corpus [10], are almost all aligned 1:1.

The 1:1 alignment has been considered as especially helpful in learning contexts, and has been therefore applied in the case of the English Second Language (ESL) [11] or VALICO-UD [12] treebanks, resources which include learner texts in English and Italian, respectively. We decided to follow their example for SICILIAN3BANK, as it might be used for language learning.

2.2. Language Variation in NLP

It is possible to distinguish two main groups of languages based on the availability of resources: high-resource languages and low-resource languages [4]. The former are languages (excluding sign languages) that have a large collection of machine-readable texts or, at the very least, a solid foundation upon which to build corpora, treebanks, and similar linguistic resources [4]. These include English, Mandarin Chinese, Arabic, and French, as well as Portuguese, Italian, Dutch, Standard Arabic, and Czech to a somewhat lesser but still significant extent [4]. Many languages, particularly local varieties and dialects, are at risk of disappearing in a relatively short time due to the lack of attention and resources they receive.

In the European context, standard languages exhibit notable diatopic variation [3]. Failing to prioritise research on language variations in the field of NLP would mean losing not only the languages as systems of communication, but also the identities, social values, and heritage of the societies they represent. It is not only a matter of increasing efforts towards these languages, but of doing so with an appropriate approach [3]. A shared

[3] A treebank is a corpus enriched with (morpho-)syntactic annotations.

goal should be established, knowledge must be made accessible to all, and subsequently disseminated beyond the community itself through engagement initiatives and the promotion of active participation. In addressing low-resource and endangered languages a novel approach would be applied based on respect, cultural awareness, and sensitivity to the wishes of their speakers.

2.3. Dialects in NLP

Focussing now specifically on dialects, it is important to note that their marginalisation is not a phenomenon exclusive to the field of NLP. A negative connotation of dialects is often rooted in complex historical, social, and political dynamics. For example in Italy, regional varieties, dialects, and other non-standard linguistic forms often coexist with the standard language in a situation known as *dilalìa* [13], where there is not a rigid compartmentalisation of the languages, as it happens in *diglossìa*, but still Italian is preferred in formal and high-prestige domains, and dialects in informal, everyday, or familial interactions. The significant linguistic loss experienced by Sicilian and other Italian dialects can also be attributed to the Fascist dictatorship, which aimed to achieve linguistic unification by suppressing regional language varieties and all that was perceived as foreign. Furthermore, the Italian language was instrumental in constructing national unity, serving as a symbol of collective identity at the expense of non-standard varieties, which were increasingly marginalised in both institutional and public domains [3].

One notable effort to address dialects and local languages is the MaiBaam project, a multi-dialectal Bavarian UD treebank [14]. It represents the first UD treebank for the Bavarian language, a West German dialect spoken in southern Germany, Austria, and northern Italy (South Tyrol). The major challenges encountered by the MaiBaam project authors, which are common issues within this field, are the difficulty to collect texts and find native-speaking annotators. While we are facing the former challenge, we did not encounter the latter, as the majority of our team members are native speakers of Sicilian. Nevertheless, there remains the necessity for a strong linguistic knowledge of the dialect being worked on—a requirement that is uncommon, given that dialects are rarely studied actively but are instead acquired through everyday use. The solution adopted by the MaiBaam group to adress this issue is making their work publicly available, which enables them to engage with the population and collect contributions from the community.[4] The Bavarian dialect is also represented for tasks such as Named Entity Recognition (NER) and Dialect Identification (DID), thanks to BarNER, a medium-sized corpus collecting Wikipedia and tweets data [15]. The authors in [16] show how such resources can be effectively utilised in NLP.

A similar initiative is the COSER-UD treebank [17], the first syntactically annotated corpus of spoken peninsular rural Spanish distributed within the UD framework [18]. The treebank addresses features such as word-order flexibility, ellipses, disfluencies, and colloquial expressions, critical for accurately representing morphosyntactic variation in oral communication.[5] By focussing on rural dialects beyond urban linguistic norms, COSER-UD enhances the diversity of linguistic data available to NLP and supports sociolinguistic preservation of under-represented varieties. The COSER-UD resource has supported the development of tasks such as Part-of-Speech (PoS) tagging, where models adapted to rural speech have been evaluated against a gold-standard dataset of over 13,000 sentences. Furthermore, the dataset has been used to test automatic speech recognition tools on dialectal Spanish audio [19].

Another noteworthy project is the East Cretan Treebank [20]. It was built from audio material of folkloric narratives collected from radio broadcasts, which were transcribed and annotated according to the UD framework. The treebank annotates dialect-specific features, such as euphonics and voicing phenomena, which are represented using dedicated tags and treated as distinct tokens in the annotated data. The East Cretan Treebank has been used for two main NLP tasks: PoS tagging and dependency parsing. Both tasks were addressed via fine-tuning of the Greek BERT model, using either exclusively the Eastern Cretan corpus data or in combination with data from the GUD, a treebank for Standard Modern Greek [21].

Focussing on Italian dialects, a treebank for Ligurian [22] is available in the UD repository which is the first-ever digital corpus of that language, comprising 316 sentences and 6,928 tokens. Like Sicilian, Ligurian is a minority variety within the Italian linguistic landscape and faces many challenges due to its low-resourced status. The project shares similar goals with ours, aiming to promote research and NLP development for endangered dialects, with a focus on supporting language preservation. The study also addresses orthographic aspects of the Genoese variety of the Ligurian dialect. The treebank was used for parsing experiments, and despite the performance of the parser is lower than those trained on high-resourced languages, the results obtained are in line with or superior to other small-scale corpora, confirming annotation consistency.

[4] Apart from sharing our resource, we mitigated this also during the annotation process by making it the most objective as possible by using shared resources.

[5] Additional information can be found at https://github.com/UniversalDependencies/UD_Spanish-COSER.

The UD repository also includes a small Neapolitan treebank that contains only 20 sentences, corresponding to 197 tokens and 199 syntactic words[6].

As far as Sicilian is concerned, a particularly interesting project is the one carried out by Arba Sicula[7] [23], which presents the first neural machine translator for the Sicilian dialect based on a deep-learning transformer model fed with Sicilian sentences augmented using back-translation [24] to cope with the lack of resources. The results were evaluated using the BLEU score metric and yielded scores of 35.0 for English>Sicilian and 36.8 for Sicilian>English. The project was later expanded into a multilingual translation system by incorporating Italian, using techniques such as transfer learning.

2.4. Studying Sicilian

When approaching the creation of a treebank for a dialect, one must come to terms with the absence of an orthographic standard and norms to regulate its development. Sicilian, as well as other dialects, exhibits great variability, especially at the diachronic and diatopic levels. To deal with these critical issues, we adopted a combined approach, drawing on different grammars and dictionaries of Sicilian and comparing them. In general, the grammars proved to be very useful to explain several phenomena and guide their representation in Sicilian3bank. However, for a few especially challenging issues, those for which we found a discordance of opinions reported in the grammars, we provided solutions based on our intuition of native speakers and consulting a linguist expert on Sicilian. We carefully discussed them and kept track of our motivations in the annotation guidelines.

For the purpose of lexical consultation and to handle different word forms, some online tools were used, such as *Wikizziunariu*[8], *Glosbe*[9], *Napizia-Chiù dâ Palora* [10], *Salviamo il siciliano*[11], plus social posts and blogs, demonstrating the importance of leveraging every available resource for dialectal language research and preservation. In addition, we consulted *Nuovo vocabolario siciliano-italiano* by Antonio Traina [25], selected for its breadth and accuracy, and various other dictionaries [26, 27, 28].

Several grammars from various time periods were also consulted [29, 30, 31, 32, 33, 34, 35, 36, 37], in order to gain a comprehensive understanding of the language also on diachronic aspect. Consulting these works revealed significant variability in the treatment of linguistic phenomena. On the one hand, some grammars document some phenomena in detail, while in others they are completely absent. On the other hand, some phenomena are mentioned in all grammars but treated differently. It was therefore necessary to make a choice based on a critical comparison of the sources and data available to us. It should be noted that, as it is common in the development of resources from scratch, some decisions were taken based on the limited set of examples currently included in the treebank. In future extensions of the resource, new comparisons with additional instances of the same or similar phenomena may prompt a revision of certain annotation choices.

3. Data Collection and Translation

In the development of a treebank, the first step to be addressed is the collection of texts to be later annotated. When the objective is a parallel treebank, texts must be made available in at least two languages. For the development of the first release of the Sicilian3bank[12] we collected a group of open source texts available on the web (sec. 3.1), and we applied to these texts a semi-automatic procedure to obtain their Italian version (sec. 3.2).

3.1. Data Collection

The first of the challenges we encountered was finding suitable texts and sources for building the treebank. We constrained our search to literature, but we do not exclude to include other genres in future enlargement of the resource, e.g. including the Sicilian pages of Wikipedia.[13] We started our search based on the criterion of contemporaneity, that is, we sought texts modern and reflecting language use consistent with present-day Sicilian. A useful source has been Panzaredda website.[14] From this source, we retrieved two of the three texts of our corpus: *U cuntu di Purpu*[15] and *Amara Sapi - Capìtulu Unu, U Zuccu*[16]. These texts do not indicate the geographic origin of the authors or the dialectal variety, which prevents us from declaring with certainty the provenance of these texts. However, based on a lexical analysis of the terms used, it is likely that the first text comes from the Agrigento area and the second from the Catania area. The third text is a collection of 18 diatopic variants[17] of

[6] The few information about this resource can be found at https://github.com/UniversalDependencies/UD_Neapolitan-RB.

[7] Arba Sicula is a non-profit international organisation that promotes the language and culture of Sicily https://arbasicula.org/.

[8] Available here: https://scn.wiktionary.org/.

[9] Available here: https://it.glosbe.com/.

[10] Available here: https://www.napizia.com/cgi-bin/cchiu-da-palora.pl.

[11] Available here: http://www.salviamoilsiciliano.com/come-si-dice/dizionario/.

[12] We plan to release it in the next official UD treebank release.

[13] Main page: https://scn.wikipedia.org/wiki/PĂăggina_principali.

[14] Available here: https://www.panzaredda.com/.

[15] Available here: https://www.panzaredda.com/post/u-cuntu-di-purpu, written by Alesci Mistretta.

[16] Available here: https://www.panzaredda.com/post/amara-sapi-capĂńtulu-unu-u-zuccu, by Goetia.

[17] This paper focuses on 17 tales from the collection, excluding the 18th tale as it is entirely written in Italian. Some parts of the 17

the legend of *Colapisci*, a very well-known folktale in Sicily, narrating the story of a merman.

In the CoNLL-U file of SICILIAN3BANK, a comment line has been added at the beginning of each text, containing information regarding the text's diatopic variant and publication year. In the specific case of *Colapisci*, this information is provided at the beginning of each story.

3.2. Creating the Parallel SICILIAN3BANK

In this section, we present the challenges of LLMs in translating the selected texts from Sicilian into Italian, and the translation principles we applied for manually correcting the automatic translations.

3.2.1. GenAI for Automatic Sicilian>Italian Translation

To translate the Sicilian texts into Italian, we exploited LLMs to obtain a first version, which was then manually revised by Sicilian native speakers.[18] We decided not to use machine translation-specific systems, because they usually do not cover dialects, and when they do, e.g. Google translate, their performance is low, as verified at a first qualitative check on our texts. We preferred to use general-purpose LLMs, as this might be the start of a more systematic study on LLMs abilities with translation of low-resource languages. The machine translated versions were produced in three different settings, giving the whole text in the prompt and asking for the translation, giving a sentence at a time with the whole text as context and giving each sentence in isolation.[19] These three versions have been produced for each of the three LLMs tested, i.e. Mistral 3 Small, LLaMA 3.3 70B and GPT-4o models. These models were accessed using GPT@JRC, a tool that enables the use of genAI models in a safe and AI-Act compliant environment [2], and using standard settings (e.g. temperature 0.7). Despite the three texts having different lengths (from less than 2k to more than 5k tokens), this did not influence the translation quality, though only qualitatively evaluated, especially in the setting asking for the translation of the whole text together, which is the one producing the best translations. This means that the degradation of performance reported in the literature about LLMs [38] (using automatic metrics such as BLEU) is not visible with our qualitative evaluation. In particular, reviewing the translations, it was observed that the best translations were generated by Mistral for the texts *Amara Sapi* and *U Cuntu di Purpu*, whereas for *Colapisci*, the most satisfactory version was the one produced by GPT-4o.[20]

We considered subjective qualitative evaluations of the overall quality of the translation, focussing on the relationship between fidelity to the original text and fluency of the translated text. Notably, despite not being specifically trained on dialect data, the LLMs demonstrated a remarkable ability to generate meaningful translations, producing a fluent and largely accurate output in both cases.[21] However, some inaccuracies regarded: (i) Untranslated or roughly translated terms—nouns in particular are the most difficult to translate and required manual corrections and lexical consultations; (ii) Cultural and linguistic nuances not correctly identified and translated; (iii) Inconsistencies in subject-verb agreement, especially in translations produced by Mistral, and the use of verb tense, which impaired temporal coherence; (iv) Omitted content—a few cases were observed where the models failed to translate parts of the text, producing incomplete results and requiring manual intervention.

3.2.2. Translation Choices

We created fluent translations into Italian, opting for the variant that has the most grammatical features of the original, when possible, as in the PUD treebank [7]. Nevertheless, fully rendering the meaning of certain expressions in the translation has been challenging. We have indeed encountered words that did not have an equivalent in Italian, or had one or more meanings. For example, in *U cuntu di Purpu*, the nickname of the main character, 'Purpu'[22], literally means 'octopus', but it is commonly used also to offensively indicate homosexual people. Nowadays, in the translation literature, it is commonly agreed that proper names are not translated, unless they carry a meaning or the target audience requires it. A thoroughly studied case is the translation of names in Harry Potter [39, 40], where localisation seems to be the most adopted technique. Since our primary aim is not translation, we decided to opt for a one-size-fits-all strategy instead of localisation, which involves an ad hoc solution for each different case: proper names were not translated, even when they carried meaning. However, in the document with the whole translated text, provided in the resource repository[23], we added footnotes provid-

collected tales contained Italian sentences, particularly in explanations of details or cross-references to similar versions. These sections were not included in the corpus.

[18] The first authors of this paper.

[19] We are aware that giving the whole text as a context per sentence is not efficient considering computation costs, but we tried this setting as we had only three texts.

[20] It must be noted that safety filters were triggered in some cases, especially in the short story *U Cuntu di Purpu*, as it is mentioned a dead body. This hindered the possibility for a full comparisons of the models and settings.

[21] Qualitatively better than translations obtained using Arba Sicula translator or Google Translate (Sicilian>English).

[22] See https://it.wiktionary.org/wiki/purpu for the translation of the term and this Quora thread https://it.quora.com/Perché-in-Sicilia-gli-omosessuali-vengono-chiamati-purpi for a discussion of its common use.

[23] Available here: https://github.com/ElisaDiNuovo/Sicilian3bank.

ing translation and further explanation where necessary. Other examples of proper names we met in the texts included in the Sicilian3bank—which are known in the translation literature as challenging since rich in social, geographical, or cultural references—are 'Liotru' (from *U Cuntu di Purpu*), literally translatable as 'elephant', but also bearing a reference to the city of Catania, that any Sicilian reader would also recognise; 'Zuccarata' (from *Amara Sapi*), which is not only an affectionate epithet used to describe a person, but also the name of a traditional dessert typical of the region.

A different approach was taken with the toponyms that had a direct equivalent in Italian, which were indeed translated, e.g. *Missina*, *Turri di Faru*, and *Napuli* (from the text *Colapisci*), rendered respectively as Messina, Torre Faro, and Napoli. Finally, fictional toponyms, such as *Cirasitu*, found in the text *Amara Sapi*, was Italianised as *Cirasito*, however the Italian reader would lose the reference to cherries.

4. Sicilian3bank in UD

In this section, we describe the annotation process and the challenges we faced in applying the UD format to our collection of texts described in Sec. 3. All the annotation choices are documented in the annotation guidelines, provided in the resource repository.

4.1. Parsing Sicilian in UD

There is no annotated resource or treebank in UD format for the Sicilian dialect. Based on the supposed similarity of Sicilian with Italian and the availability of UD treebanks for this latter, we decided to create a first draft of the Sicilian annotated data using the models for Italian, expecting to find a significant amount of errors in the output to be manually corrected. We selected the models trained on ISDT [41] and POSTWITA [42] treebanks, which are the biggest resources for Italian available in the UD repository, and we have a performance evaluation of these models in non standard Italian texts (i.e. [12]). A preliminary comparison of the outputs generated by UDPipe[24] trained on them showed that the model based on ISDT outperforms that based on POSTWITA in dealing with Sicilian data. We started therefore the manual check and correction of the output of UDPipe trained on ISDT, feeding it with gold sentence segmentation.[25]

The three first authors, all native Sicilian speakers skilled in linguistics and computational linguistics, carried out this manual revision of the automatic annotation leading to the first version of the Sicilian3bank. The tool used for the correction was Arborator [43].[26] Each of the three texts was annotated by one annotator. The annotation was reviewed by a second annotator. Problematic phenomena were discussed by the three annotators together, and specific cases also with the rest of the authors.[27] In Table 2 in Appendix A we report an example of the CoNLL-U file for a Sicilian sentence of the treebank, featuring a comment line with the Sicilian text, and the aligned Italian translation.

When it comes to this parallel dataset composed of the translations into Italian of the Sicilian sentences (described in Sec. 3), the same parsing approach has been applied, thus creating the Sicilian-Italian parallel treebank. Nevertheless, considering that our main focus is on the Sicilian dialect, we decided to concentrate our current efforts on the creation of the parallel data (translation into Italian) and the manual correction of the annotation of the Sicilian data, carefully checking them both, and planning instead the manual check of the annotation of the Italian parallel data of the Sicilian3bank as a future work. This is further justified as automatic parsers for Italian are considered good enough, although some marginal phenomena still are consistently wrongly annotated [44, 12]. The next section is therefore focused on the analysis based on the Sicilian data only.

4.2. A Quantitative Analysis of the Sicilian Data

After the manual check and correction, the Sicilian resource annotated in CoNLL-U format consists of a total of 505 sentences and 11,709 tokens (Table 1). Each annotated sentence of each of the three texts presented in Sec. 3.1 includes a comment text line that reports the sentence in Sicilian dialect followed by a comment text line containing the translation into Italian. Following this, the UD annotation of the sentence is provided organised in the ten columns typical of this format (Table 2).

Text	Number of sentences	Number of tokens
Amara Sapi	246	4723
Colapisci	179	5092
U cuntu di Purpu	80	1894
Total	505	11709

Table 1
The distribution of sentences and tokens in the Sicilian data of the Sicilian3bank.

[24] Available here: https://lindat.mff.cuni.cz/services/udpipe/.

[25] For sentence segmentation we followed the VALICO-UD project, which does not split sentences on colons and treats direct speech as single segment.

[26] We noticed that Arborator (https://arborator.ilpga.fr) allowed to split tokens only into two, so in case of verb + double clitic we had to further tokenise manually.

[27] To further ensure annotation quality, an inter-annotator agreement score (Krippendorff's kappa) will be computed for future releases of the treebank.

A comparison of the annotation provided by UDPipe with the manually corrected data enables us to evaluate the transfer domain abilities of the parsing models when applied on the Sicilian data. In Table 3 in Appendix A, we report the scores (precision, recall and F1 for UPOS, LAS and UAS) obtained by UDPipe models trained on ISDT and on PoSTWITA. These results confirm that the model based on ISDT outperforms the other one, but it must be observed that it may depend at least in part on the fact that the output of UDPipe trained on ISDT was the base for the manual correction. The table shows that the best performance based on ISDT can be referred to *Colapisci* (LAS F1 72.87) while the worst to *Amara Sapi* (LAS F1 59.80). An in-depth investigation of these results is beyond the scope of this paper, but will be addressed in our future work. However, we can qualitatively observe that the performance of the two models differs for some phenomena. For example, the model trained on PoSTWITA was more robust in annotating verbs containing double clitic pronouns.

4.3. Challenges in Dealing with the Sicilian Dialect

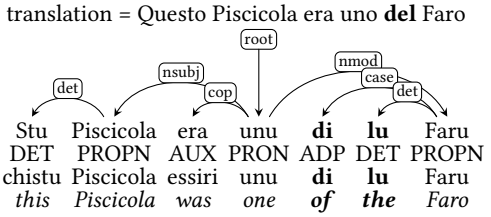
The approach used for the generation of the annotated data, based on models available for Italian, has clearly brought out some characteristics and phenomena that differentiate Sicilian from Italian. It is in dealing with these phenomena that the parser has produced more annotation errors, and it is on them that the work of manual correction was mostly concentrated.

This section presents some choices we had to make to deal with some features of the Sicilian texts considered. In particular, we focus on tokenisation (articulated prepositions), lemmatisation (orthographic variations of some pronouns reflecting suprasegmental traits), and syntactic (focussing here on the reduplication phenomenon) choices.

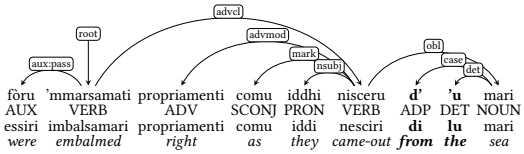
4.3.1. Tokenisation Issues

A particularly relevant phenomenon that emerged during the annotation is that represented by articulated prepositions, for which there has been, over time, a process of grammaticalisation that has determined their evolution. Generally, many prepositions that in Italian occur in a unified form have undergone a transformation in Sicilian, first passing through a disjunct form (Example 1)[28], until arriving at forms with elision (Example 2)[29] [34, 31] and, in more recent times, with contraction (Example 3)[30], although the disjunct form is still present, at least in some areas [33].

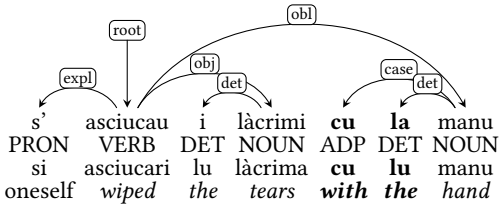
(1) # text = Stu Piscicola era unu **di lu** Faru
translation = Questo Piscicola era uno **del** Faro



(2) # text = fòru 'mmarsamati propriamenti comu iddhi nisceru **d' 'u** mari
translation = furono imbalsamate proprio quando uscirono **dal** mare



(3) # text = S'asciucau i làcrimi **câ** manu
translation = Si asciugò le lacrime **colla** mano



Contracted articulated prepositions—graphically marked by the circumflex accent [29, 30, 32]—were split into two different tokens, as shown in Example 3. In this way we show, for each articulated preposition, the morphology attached to it, even in those cases in which it is not apparently visible, as it is nevertheless part of its evolution and can be described by formal rules. A different choice, such as not splitting it into two tokens, would have highlighted the grammaticalisation of this particular phenomenon by not splitting it into two tokens. However, this choice might necessitate the creation of a specific UPOS, which would hinder cross-language comparisons.

Similarly the forms *nta* and *ntâ* differ as the former is a simple preposition, equivalent to *in* of Italian, while the latter is the articulated preposition. Depending on the gender and number of the article, it can be rendered as *ntô* (masculine singular), *ntê* (plural, both masculine and feminine).

It is worth noting in this regard that the Italian preposition *in* can be rendered in Sicilian in various ways, such as *in*, *ni*, *nni*, *nta* [29]. The same is true for the Italian simple preposition *da*, which in Sicilian occurs in the forms *di*, *ni* and *nni* [29]. These different forms are reflected also in the corresponding articulated prepositions

[28] English translation: *This Piscicola was one from Faro.*
[29] English translation: *[...] were embalmed just as they emerged from the sea.*
[30] English translation: *He wiped away his tears with his hand.*

(e.g. the Italian preposition *nello*, such as *ntô*, *nô* and *nnô*). Please see Sec. 4.3.2, for our lemmatisation choices for these variants.

The complete scheme of the articulated prepositions system in Sicilian is presented in Table 4 in Appendix A.

4.3.2. Lemmatisation Issues

Concerning lemmatisation, as Sicilian does not have a unified orthography—although recent efforts try to standardise this [32]—in the texts considered there are different variants for the same forms, which try to render different pronunciations. For example, in the considered texts there is no consistency in the transcription of the Sicilian word meaning 'no one', *nuddu*, which is pronounced reproducing a long voiced retroflex stop, but it is transcribed sometimes as *nuddu*, other times as *nuḍḍu*, stressing the retroflex pronunciation. Other variants of the same word are *nuddru*, *nuddhu*. Since our aim is not focused on phonetics, we lemmatised these occurrences without any pronunciation marks, i.e. *nuddu*, and decided not to uniform the orthographic rendering (i.e. the form) of this word and similar cases, e.g. *ci*/*cci* and *ni*/*nni*, as shown in Examples 4a-4b and 5a-5b, respectively.

(4a) # text = **ci** succidìu accussì LEMMA **ci**
translation = **gli** successe questo (*this happened **to him***)

(4b) # text = chi **cci** jemu a fari? LEMMA **ci**
translation = che **ci** andiamo a fare? (*what are we going to do **there**?*)

(5a) # text = **ni** chiamavanu "l'Armali" LEMMA **ni**
translation = **ci** chiamavano "gli Animali" (*they called **us** "the animals"*)

(5b) # text = Chi **nni** putìa sapiri iu? LEMMA **ni**
translation = Che **ne** potevo sapere io? (*How could I know **about that**?*)

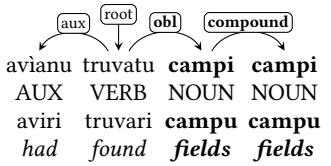
We applied the same principle to shortened oral variants of words, e.g. *diri* ('to say') or *riri*, both of which are abbreviated forms of *diciri*. All such variants have been lemmatised using the extended lemma, such as *diciri* in Example 8).

To summarise, the main aim of lemmatisation is to reduce the sparseness of forms and their variants by reducing them to a common lemma, regardless of the causes of this sparseness. Therefore, we have applied the same strategy used in other resources where sparsity is determined, for example, by the writing style of the users (or by errors due to the writing device they use), as in PoSTWITA[42], to the lemmatisation of Sicilian3bank.

4.3.3. Syntax Issues

One of the cases in which we had to take a decision about a syntactic phenomenon is reduplication, a typical and widespread phenomenon in the Sicilian dialect [45], which consists in the repetition of a word, resulting in a shift or extension of meaning within the sentence. It is a phenomenon still highly productive in contemporary Sicilian, as shown by Amenta through the analysis of a corpus from the *Atlante Linguistico della Sicilia* [46], where these forms exhibit neither diachronic nor diastratic variation, thereby confirming the ongoing vitality of this linguistic process. This phenomenon can involve the reduplication of a verb to form an adjective or a noun; a noun to form an adjective or an adverb; and other PoS [47]. This last pattern, the most frequent in our texts, reveals several semantic implications, but frequently is used as a locational nominal modifier. In order to highlight the compound nature of this phenomenon (in [45, p. 350], it is clearly stated that it is not possible to interpose any words between the two elements of the reduplicated construct), we use the relation *compound* and the relation *obl*, in line with UD guidelines, as shown in Example 6[31]. In addition we added *LOC=adv* in the last column of the CoNLL-U file, as it is done in VALICO-UD, to indicate that there is an adverbial locution.

(6) # text = avìanu truvato **campi campi**
translation = avevano trovato **tra i campi**



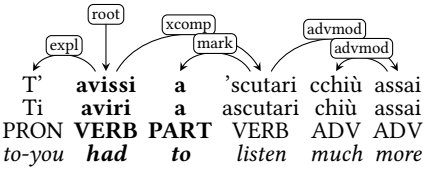
4.4. A Cross-Linguistic Analysis Example

In Sicilian, modal verbs—like the auxiliaries *essiri* ('to be') and *aviri* ('to have')—can serve two main functions: they may appear independently with their own lexical meaning, or they may function as support verbs, combining with an infinitive (without a preposition) to convey specific modal values, such as: (i) ability/possibility → *putiri* ('can'); (ii) will/desire → *vuliri* ('want'); (iii) obligation/necessity → *duviri* ('must') or *aviri a* ('have to').
In modern Sicilian, particularly in spoken usage, the periphrastic construction *aviri a* + infinitive is commonly employed to express modal meanings, especially obligation, replacing the older verb *duvìri* found in Old Sicilian [30] (see Example 7)[32]. Within this construction, the tense of *aviri* plays a central role in conveying modal values, whether epistemic or deontic. When *aviri* appears in the past remote, its perfective aspect confers an epistemic meaning, indicating certainty about the event's occurrence in the past. In contrast, when *aviri* is used in the present or imperfect—both imperfective tenses—the construction can express either an epistemic sense of probability or a deontic sense of obligation or necessity.

[31] English translation: *[...] they had found among the fields.*
[32] English translation: *I should listen to you much more often.*

In some cases, especially with the present indicative or imperfect subjunctive, an exhortative function may also emerge [48].

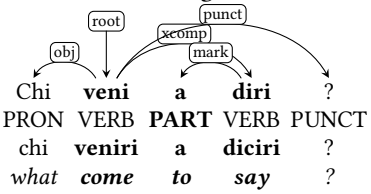
(7) # text = T'**avissi a** 'scutari cchiù assai
translation = ti **dovrei** ascoltare molto di più



The annotation in UD of such resource allows for drawing a parallel with other languages. For example, with the English *have to* construction, which is similarly used to express obligation and certainty [49, p. 210]. In the English UD treebanks *to* is consistently annotated as a particle when used in this way (see Example 9a in Appendix A). We therefore decided to treat the element *a*, which is usually tagged as a preposition in our corpus, as a particle in this specific construction. However, in Italian *avere da* can be used with the same meaning (see Example 9b in Appendix A), but *da* is not annotated as particle. This might be due to historical reasons, a different function of *da* in Italian than of *to* in English, or to highlight a less grammaticalised relation.

Another periphrastic construction found in the treebank texts is *veniri + a + diciri* (literal translation into Italian *venire a dire*), which can have the meaning of the Italian verb *significare* ('to mean'). In such cases, we treated it in the same way as the previous one, as shown in Example 8[33].

(8) # text = Chi **veni a diri**?
translation = Che **significa**?



5. Conclusion and Future Work

We can create a world that sustains its languages [50]. Among the concrete actions we can perform to achieve this goal, there is the possibility of speaking and studying the original languages of our places.

This paper describes and discusses the issues involved in the development of the first release of the Sicilian3bank. Many are the challenges we have encountered in dealing with a language which has never been treated before and which is in addition a dialect, which carries with it an uninterrupted history of oral transmission but does not have a standardised form of transcription or unified treatment of phenomena in grammars.

The project we present here is intended therefore solely as a preliminary foundation and proposal, which nonetheless requires substantial further work and numerous improvements. First, the inclusion of more texts and perform inter-annotator agreement, to verify guidelines soundness. Second, the corpus enrichment introducing Italian glosses in the MISC column of the CoNLL-U file. In the current version, each sentence is accompanied by a fluent Italian translation in a comment line, we propose the inclusion of a literal word-for-word translation from Sicilian into Italian. Although this form of translation may result in grammatically incorrect or unnatural Italian, it would provide an almost word-by-word parallel aligned resource that mirrors the syntactic structure of the original Sicilian sentences and would facilitate syntactic calque studies. Third, a future objective would be to manually validate the automatic annotation generated with UDPipe for the aligned Italian resource as well. This step is needed to give to the Italian parallel dataset the same quality we are currently providing for the Sicilian annotated data. Fourth, another interesting enhancement might be to systematically include graphic accents on all verb lemmas, to help reading them, and including in MISC column of the CoNLL-U file the International Phonetic Alphabet transcription. This idea is motivated by the desire to turn the resource not only into a syntactic dataset but also into a tool to support language learning, scientific studies and preservation of Sicilian. Finally, an aspect we would like to improve in the future concerns the translation of proper nouns. As already discussed, we encountered several challenges in translating these elements, which ultimately led us to the decision not to translate the proper nouns found in the texts at this stage. The focus of this work is the development of a Sicilian treebank, and although a deeper engagement with translation would certainly have added valuable insights, it would have diverted attention from the project's primary objective. We therefore plan to revisit this aspect in a later phase of the project.

[33] English translation: *What does it mean?*

Acknowledgment

We would like to express our gratitude to Giuseppe Domenico Muscianisi, PhD, from the University of Parma, for very kindly sharing with us his expertise, which was instrumental in resolving several of our questions and improving our knowledge about the literature on the Sicilian dialect.
A special thanks goes to the JRC internal reviewers and to the CLiC-it 2025 anonymous reviewers for their precious comments.

References

[1] S. Bird, D. Yibarbuk, Centering the Speech Community, in: Y. Graham, M. Purver (Eds.), Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics - Volume 1: Long Papers, ACL, St. Julian's, Malta, 2024, p. 826–839. URL: https://aclanthology.org/2024.eacl-long.50/. doi:10.18653/v1/2024.eacl-long.50.

[2] B. De Longueville, I. Sanchez, S. Kazakova, S. Luoni, F. Zaro, K. Daskalaki, M. Inchingolo, The Proof is in the Eating: Lessons Learnt from One Year of Generative AI Adoption in a Science-for-Policy Organisation, AI 6 (2025) 128.

[3] A. Ramponi, Language Varieties of Italy: Technology Challenges and Opportunities, Transactions of the Association for Computational Linguistics 12 (2024) 19–38. doi:https://doi.org/10.1162/tacl_a_00631.

[4] E. M. Bender, The #BenderRule: On Naming the Languages We Study and Why It Matters, The Gradient (2019).

[5] M.-C. de Marneffe, C. D. Manning, J. Nivre, D. Zeman, Universal Dependencies, Computational Linguistics 47 (2021) 255–308. URL: https://aclanthology.org/2021.cl-2.11/. doi:10.1162/coli_a_00402.

[6] H. Bunt, P. Merlo, J. Nivre (Eds.), Trends in Parsing Technology: Dependency Parsing, Domain Adaptation, and Deep Parsing, volume 43, Springer Science & Business Media, 2010.

[7] D. Zeman, M. Popel, M. Straka, J. Hajič, J. Nivre, F. Ginter, J. Luotolahti, S. Pyysalo, S. Petrov, M. Potthast, F. Tyers, E. Badmaeva, M. Gokirmak, A. Nedoluzhko, S. Cinkova, J. Hajic jr., J. Hlaváčová, V. Kettnerová, Z. Urešová, J. Kanerva, S. Ojala, A. Missilä, C. D. Manning, S. Schuster, S. Reddy, D. Taji, N. Habash, H. Leung, M.-C. de Marneffe, M. Sanguinetti, M. Simi, H. Kanayama, V. dePaiva, K. Droganova, H. Martínez Alonso, C. Çöltekin, U. Sulubacak, H. Uszkoreit, V. Macketanz, A. Burchardt, K. Harris, K. Marheinecke, G. Rehm, T. Kayadelen, M. Attia, A. Elkahky, Z. Yu, E. Pitler, S. Lertpradit, M. Mandl, J. Kirchner, H. F. Alcalde, J. Strnadová, E. Banerjee, R. Manurung, A. Stella, A. Shimada, S. Kwak, G. Mendonça, T. Lando, R. Nitisaroj, J. Li, CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies, in: J. Hajič, D. Zeman (Eds.), Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies, Association for Computational Linguistics, Vancouver, Canada, 2017, pp. 1–19.

[8] M. Sanguinetti, C. Bosco, Building the multilingual TUT parallel treebank, in: Proceedings of The Second Workshop on Annotation and Exploitation of Parallel Corpora, 2011, pp. 19–28.

[9] M. Sanguinetti, C. Bosco, PartTUT: The Turin University Parallel Treebank, in: R. Basili, C. Bosco, R. Delmonte, A. Moschitti, M. Simi (Eds.), Harmonization and Development of Resources and Tools for Italian Natural Language Processing within the PARLI Project, Springer, 2015, pp. 51–69.

[10] R. Steinberger, M. Ebrahim, A. Poulis, M. Carrasco-Benitez, P. Schlüter, M. Przybyszewski, S. Gilbro, An overview of the European Union's highly multilingual parallel corpora, Language Resources and Evaluation 48 (2014) 679–707.

[11] Y. Berzak, J. Kenney, C. Spadine, J. X. Wang, L. Lam, K. S. Mori, S. Garza, B. Katz, Universal Dependencies for Learner English, in: E. Katrin, A. S. Noah (Eds.), Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, 2016.

[12] E. Di Nuovo, Introducing Valico-UD: A Parallel, Learner Italian Treebank for Language Learning Research, Pàtron, 2023.

[13] G. Berruto, Lingua, dialetto, diglossia, dilalia, in: G. Holtus, J. Kramer (Eds.), Romania et Slavia Adriatica. Festschrift für Zarko Muljačić, Buske, Hamburg, 1987, pp. 57–81.

[14] V. Blaschke, B. Kovačić, S. Peng, H. Schütze, B. Plank, MaiBaam: A Multi-Dialectal Bavarian Universal Dependency Treebank, in: N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, N. Xue (Eds.), Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), ELRA and ICCL, Torino, Italia, 2024, pp. 10921–10938. URL: https://aclanthology.org/2024.lrec-main.953/.

[15] S. Peng, Z. Sun, H. Shan, M. Kolm, V. Blaschke, E. Artemova, B. Plank, Sebastian, Basti, Wastl?! Recognizing Named Entities in Bavarian Dialectal Data, in: N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, N. Xue (Eds.), Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), ELRA and ICCL, Torino, Italia, 2024, pp. 14478–14493. URL: https://aclanthology.org/2024.lrec-main.1262/.

[16] X. M. Krückl, V. Blaschke, B. Plank, Improving Dialectal Slot and Intent Detection with Auxiliary Tasks: A Multi-Dialectal Bavarian Case Study, in: Y. Scherrer, T. Jauhiainen, N. Ljubešić, P. Nakov, J. Tiedemann, M. Zampieri (Eds.), Proceedings of the 12th Workshop on NLP for Similar Languages, Varieties and Dialects, Association for Computational Linguistics, Abu Dhabi, UAE, 2025, pp. 128–

146. URL: https://aclanthology.org/2025.vardial-1.10/.
[17] J. E. Bonilla, Spoken Spanish PoS tagging: gold standard dataset, Language Resources and Evaluation 59 (2025) 983–1012. doi:10.1007/s10579-024-09751-x.
[18] J. E. Bonilla, Development of the first spoken spanish treebank within the universal dependencies framework: A multi-regional approach, submitted.
[19] C. Adsuar Ávila, Automatic Speech Recognition in Dialectal Data (COSER), 2024. URL: https://audias.ii.uam.es/2024/10/30/automatic-speech-recognition-in-dialectal-data-coser/, Presentation at the AUDIAS-UAM Seminar, October 30, 2024.
[20] S. Vakirtzian, V. Stamou, Y. Kazos, S. Markantonatou, Dialectal treebanks and their relation with the standard variety: The case of East Cretan and Standard Modern Greek, in: R. Johansson, S. Stymne (Eds.), Proceedings of the Joint 25th Nordic Conference on Computational Linguistics and 11th Baltic Conference on Human Language Technologies (NoDaLiDa/Baltic-HLT 2025), University of Tartu Library, Tallinn, Estonia, 2025, pp. 776–784. URL: https://aclanthology.org/2025.nodalida-1.77/.
[21] P. Prokopidis, H. Papageorgiou, Experiments for Dependency Parsing of Greek, in: Y. Goldberg, Y. Marton, I. Rehbein, Y. Versley, Ö. Çetinoğlu, J. Tetreault (Eds.), Proceedings of the First Joint Workshop on Statistical Parsing of Morphologically Rich Languages and Syntactic Analysis of Non-Canonical Languages, Dublin City University, Dublin, Ireland, 2014, pp. 90–96. URL: https://aclanthology.org/W14-6109/.
[22] S. Lusito, J. Maillard, A Universal Dependencies corpus for Ligurian, in: M. de Lhoneux, R. Tsarfaty (Eds.), Proceedings of the Fifth Workshop on Universal Dependencies (UDW, SyntaxFest 2021), Association for Computational Linguistics, Sofia, Bulgaria, 2021, pp. 121–128. URL: https://aclanthology.org/2021.udw-1.10/.
[23] E. Wdowiak, Sicilian Translator: A Recipe for Low-Resource NMT, 2021. URL: https://arxiv.org/abs/2110.01938. arXiv:2110.01938.
[24] R. Sennrich, B. Haddow, A. Birch, Improving Neural Machine Translation Models with Monolingual Data, in: K. Erk, N. A. Smith (Eds.), Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Berlin, Germany, 2016, pp. 86–96. URL: https://aclanthology.org/P16-1009/. doi:10.18653/v1/P16-1009.
[25] A. Traina, Nuovo vocabolario siciliano-italiano, Palermo, Lauriel, 1868.
[26] G. Biundi, Vocabolario manuale completo siciliano-italiano seguito da un'appendice e da un elenco di nomi proprj siciliani: coll'aggiunta di un dizionario geografico in cui sono particolarmente descritti i nomi di città, fiumi, villaggi ed altri luoghi rimarchevoli della Sicilia: e corredato di una breve grammatica per gl'Italiani, Palermo, Carini, 1851.
[27] V. Mortillaro, Nuovo dizionario siciliano-italiano. Volume unico, Palermo, Stabilimento tipografico Lao, 1876.
[28] R. Rocca, Dizionario Siciliano-Italiano compilato su quello del Pasqualino con aggiunte e correzioni. Volume unico, Catania, Pietro Giunti Editore, 1839.
[29] A. Fortuna, Grammatica siciliana: Principali regole grammaticali, fonetiche e grafiche (comparate tra i vari dialetti siciliani), Caltanissetta, Terzo Millennio Editore, 2002.
[30] F. Giacalone, Prammatica siciliana. Storia della nostra lingua, proverbi, curiosità, modi di dire, consigli pratici per una corretta scrittura, Trapani, Edizioni Colorgrafica, 2009.
[31] A. Messina, Grammatica sistematica della lingua siciliana. Dall'ortoepia all'ortografia. Dall'analisi grammaticale all'analisi logica e del periodo. Con antologia esemplificativa dei poeti. Seconda edizione riveduta e ampliata con 30 chine sui mestieri d'una volta eseguite da Francesco Nania e poesie, Assessorato alle politiche scolastiche di Siracusa, 2007.
[32] S. Baiamonte, Documento per l'ortografia del siciliano. Documentu pi l'ortugrafia dû sicilianu. II edizione, Cademia Siciliana, 2024.
[33] Lingua siciliana. Come scrivere in siciliano, n.d. URL: https://linguasiciliana.com/come-scrivere-in-siciliano/.
[34] M. Gorini, Ortografia Siculo-Calabra, 2017. URL: https://michelegorini.blogspot.com/2017/08/ortografia-siculo-calabra.html.
[35] G. Gerbino, N. Barone, Cenni di ortografia siciliana, Trapani, Jò A.L.A.S.D., 2011.
[36] V. Lumia, La Nostra Grammatica Siciliana, Trapani, Jò A.L.A.S.D., 2010.
[37] N. Russo, Corso di grammatica siciliana, Forum Lingua siciliana 2003.
[38] L. Wang, Z. Du, W. Jiao, C. Lyu, J. Pang, L. Cui, K. Song, D. Wong, S. Shi, Z. Tu, Benchmarking and Improving Long-Text Translation with Large Language Models, in: L.-W. Ku, A. Martins, V. Srikumar (Eds.), Findings of the Association for Computational Linguistics: ACL 2024, Association for Computational Linguistics, Bangkok, Thailand, 2024, pp. 7175–7187. URL: https://aclanthology.org/2024.findings-acl.428/. doi:10.18653/v1/2024.findings-acl.428.
[39] K. Brøndsted, C. Dollerup, The names in Harry Pot-

ter, Perspectives: Studies in Translatology 12 (2004) 56–72. doi:10.1080/0907676X.2004.9961490.
[40] C. Mastrangelo, Harry Potter in Translation: Comparison of Nine Romance Languages in the Translation of Proper Names in Harry Potter and the Philosopher's Stone, Transletters. International Journal of Translation and Interpreting (2024) 1–28.
[41] C. Bosco, S. Montemagni, M. Simi, Converting Italian Treebanks: Towards an Italian Stanford Dependency Treebank, in: A. Pareja-Lora, M. Liakata, S. Dipper (Eds.), Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse, Association for Computational Linguistics, Sofia, Bulgaria, 2013, pp. 61–69. URL: https://aclanthology.org/W13-2308/.
[42] M. Sanguinetti, C. Bosco, A. Lavelli, A. Mazzei, O. Antonelli, F. Tamburini, PoSTWITA-UD: an Italian Twitter Treebank in Universal Dependencies, in: N. Calzolari, K. Choukri, C. Cieri, T. Declerck, S. Goggi, K. Hasida, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis, T. Tokunaga (Eds.), Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), European Language Resources Association (ELRA), Miyazaki, Japan, 2018, pp. 1768–1775. URL: https://aclanthology.org/L18-1279/.
[43] G. Guibon, M. Courtin, K. Gerdes, B. Guillaume, When Collaborative Treebank Curation Meets Graph Grammars, in: N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis (Eds.), Proceedings of the Twelfth Language Resources and Evaluation Conference, European Language Resources Association, Marseille, France, 2020, pp. 5291–5300.
[44] E. Di Nuovo, M. Sanguinetti, A. Mazzei, E. Corino, C. Bosco, VALICO-UD: Treebanking an Italian Learner Corpus in Universal Dependencies, IJCoL. Italian Journal of Computational Linguistics 8 (2022).
[45] L. Amenta, La reduplicazione sintattica in siciliano, Bollettino del Centro di studi filologici e linguistici siciliani 22 (2010) 345–358.
[46] G. Ruffino, Linee di discussione a ipotesi di lavoro per l'Atlante Linguistico della Sicilia, in: Actas do XIX Congreso Internacional de Lingüística e Filoloxia Románicas (1989), volume VIII, A Coruña, 1996, pp. 649–682.
[47] G. Todaro, F. Villoing, P. Gréa, INTERNAL LOCALISATION NN ADV REDUPLICATION IN SICILIAN, in: Colloque International de Morphology, volume 22, Bordeaux, France, 2012.
[48] L. Amenta, Perifrasi verbali in siciliano, in: J. Garzonio (Ed.), Studi sui dialetti della Sicilia, Unipress, Padova, 2010, pp. 1–20.
[49] M. Swan, Practical English Usage 3rd edition, Oxford University Press, 2005.
[50] S. Bird, Beyond Technological Solutions: How we Create a World that Sustains its Languages, Linguapax Review 9 (2022) 167–173.

A. Appendix

```
# sent_id = 35
# text = Nuḍḍu di nuiautri sapìa soccu fari.
# translation = Nessuno di noi sapeva cosa fare.
1   Nuḍḍu     nuddu     PRON    PI    Gender=Masc|Number=Sing|PronType=Ind                          4   nsubj   _   _
2   di        di        ADP     E     _                                                            3   case    _   _
3   nuiautri  nuiautri  PRON    PE    Number=Plur|Person=1|PronType=Prs                            1   nmod    _   _
4   sapìa     sapiri    VERB    V     Mood=Ind|Number=Sing|Person=3|Tense=Imp|VerbForm=Fin         0   root    _   _
5   soccu     soccu     PRON    PQ    Number=Sing|PronType=Int                                     6   obj     _   _
6   fari      fari      VERB    V     VerbForm=Inf                                                 4   ccomp   _   SpaceAfter=No
7   .         .         PUNCT   FS    _                                                            4   punct   _   SpacesAfter=\r\n
```

Table 2
Exemplification of line comments and fields in the treebank CoNLL-U file. The first column contains the token IDs, the second the token form, the third the lemmas, the fourth the UPOS (i.e. the Universal Part of Speech, which is in common to all the languages covered in UD), the fifth, the XPOS (language specific PoS), the sixth the morphological features, the seventh the dependency head, the eighth the syntactic relation, the ninth is left blank as it is used for enhanced dependencies, not annotated in this treebank, and the last and tenth column for miscellaneous information.

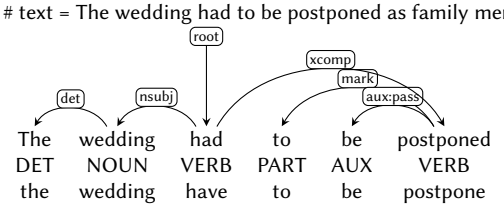
Text	Model	Metrics	Precision	Recall	F1 Score
Amara Sapi	ISDT	Tokens	97.65	97.54	97.59
		UPOS	79.59	76.48	78.00
		UAS	71.52	68.73	70.10
		LAS	61.02	58.63	59.80
	PoSTWITA	Tokens	93.45	88.41	90.86
		UPOS	69.84	63.57	66.56
		LAS	62.13	56.56	59.21
		UAS	51.66	47.02	49.23
Colapisci	ISDT	Tokens	93.56	96.59	95.05
		UPOS	82.61	84.49	83.54
		UAS	78.43	80.22	79.31
		LAS	72.06	73.60	72.87
	PoSTWITA	Tokens	91.23	92.04	91.63
		UPOS	77.64	77.59	77.61
		UAS	72.24	72.20	72.22
		LAS	65.38	65.34	65.36
U cuntu di Purpu	ISDT	Tokens	99.89	99.77	99.83
		UPOS	86.78	84.34	85.55
		UAS	76.46	74.31	75.37
		LAS	68.35	66.43	67.37
	PoSTWITA	Tokens	97.26	94.56	95.89
		UPOS	79.84	75.52	77.62
		UAS	69.23	65.49	67.31
		LAS	61.36	58.05	59.66

Table 3
Evaluation of the two models trained on ISDT and PoSTWITA output against the manually corrected CoNLL-U files, considering precision, recall, F1 of tokenisation, UPOS, UAS (i.e. unlabelled attachment score) and LAS (i.e. labelled attachment score).

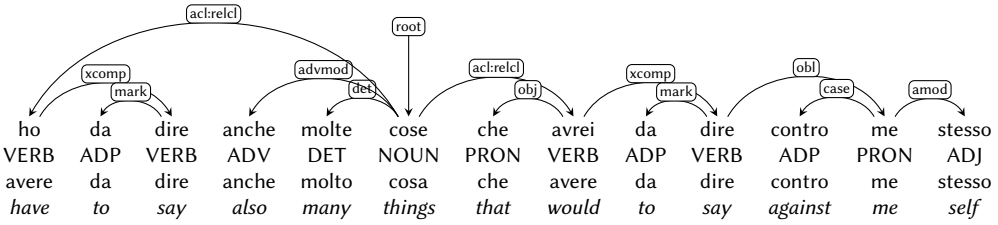
Articulated prepositions	Composition	Lemmas	Feats
dû	di+lu	di+lu	Definite=Def|Gender=Masc|Number=Sing|PronType=Art
dâ	di+la	di+lu	Definite=Def|Gender=Fem|Number=Sing|PronType=Art
dî	di+li	di+lu	Definite=Def|Gender=Masc/Fem|Number=Plur|PronType=Art
ô	a+lu	a+lu	Definite=Def|Gender=Masc|Number=Sing|PronType=Art
â	a+la	a+lu	Definite=Def|Gender=Fem|Number=Sing|PronType=Art
ê	a+li	a+lu	Definite=Def|Gender=Masc/Fem|Number=Plur|PronType=Art
nô/nnô/ntô	ni+lu/nta+lu	ni+lu/nta+lu	Definite=Def|Gender=Masc|Number=Sing|PronType=Art
nâ/nnâ/ntâ	ni+la/nta+la	ni+lu/nta+lu	Definite=Def|Gender=Fem|Number=Sing|PronType=Art
nê/nnê/ntê	ni+li/nta+li	ni+lu/nta+lu	Definite=Def|Gender=Masc/Fem|Number=Plur|PronType=Art
kû/cû	cu+lu	cu+lu	Definite=Def|Gender=Masc|Number=Sing|PronType=Art
kâ/câ	cu+la	cu+lu	Definite=Def|Gender=Fem|Number=Sing|PronType=Art
kî/chî	cu+li	cu+lu	Definite=Def|Gender=Masc/Fem|Number=Plur|PronType=Art
pû	pi+lu	pi+lu	Definite=Def|Gender=Masc|Number=Sing|PronType=Art
pâ	pi+la	pi+lu	Definite=Def|Gender=Fem|Number=Sing|PronType=Art
pî	pi+li	pi+lu	Definite=Def|Gender=Masc/Fem|Number=Plur|PronType=Art

Table 4
Scheme of the Sicilian articulated preposition system. Gender=Masc/Fem for example in the third row indicated that the same articulated preposition is used referring to masculine and feminine nouns, and its gender can only be distributionally understood.

(9a) [From EWT treebank]
sent_id = weblog-blogspot.com_alaindewitt_20060827093500_ENG_20060827_093500-0017
text = The wedding had to be postponed as family members fled the outbreak of the war, she said.



(9b) [From ISDT treebank]
sent_id = isst_tanl-1497
text = ho da dire anche molte cose che avrei da dire contro me stesso



Declaration on Generative AI

During the preparation of this work, the author(s) used ChatGPT (OpenAI), Grammarly, Other, and GPT@JRC (an internal JRC testbed for LLMs. The model used there is an on-premises installation of LLaMa 3.3 70B) in order to: Paraphrase and reword, Improve writing style, Grammar and spelling check, and Citation management. After using these tool(s)/service(s), the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.