Ciallabacialla! Modeling and Linking a Regional Lexical Resource to Include Sicilian in the Semantic Web

Rachele Sprugnoli[1,*,†], Giovanni Moretti[1], Domenico Giuseppe Muscianisi[2] and Eleonora Litta[1]

[1] *Università Cattolica del Sacro Cuore, Largo Gemelli, 1, 20123 Milano, Italy*

[2] *Università di Parma, Via D'Azeglio, 85, 43125 Parma, Italy*

Abstract

This paper describes the inclusion of Sicilian in the Semantic Web through the development of new resources aligned with Linguistic Linked Open Data principles. More specifically, we model and publish the first Sicilian Lemma Bank and a bilingual Sicilian–Italian glossary extracted from the Sicilian Wiktionary (*Wikizziunariu*). These resources are formalized using the OntoLex-Lemon and LiLa (Linking Latin) ontologies with the aim of enabling cross-lingual interoperability. The glossary is also linked to the LiITA (Linking Italian) knowledge base. In addition, two preliminary experiments are reported: the first evaluates the translation capabilities of commercial Large Language Models (LLMs) from Sicilian into Italian; the second investigates bilingual lexicon induction through cross-lingual embedding alignment, with results indicating the challenges posed by low-resource dialects. This work aims to demonstrate the feasibility and importance of integrating under-resourced languages into broader Computational Linguistics and Semantic Web infrastructures.

Keywords

Sicilian, Linguistic Linked Open Data, Semantic Web, lexical resources, dialectology

1. Introduction

The LiITA (Linking Italian) project is dedicated to developing an interoperable Knowledge Base (KB) for Italian linguistic resources. Its primary goal is to construct a network that interconnects diverse Italian language datasets (such as dictionaries, lexicons, and textual corpora) by publishing them as Linked Open Data (LOD). At the core of LiITA is the Lemma Bank (LB), a continually expanding repository of canonical citation forms (lemmas) for Italian words [1]. The LB functions as a central hub, enabling interlinking and interoperability across various linguistic datasets. By aligning lexical entries and word occurrences from distributed resources with their corresponding lemmas, LiITA supports federated search capabilities and facilitates advanced linguistic analyses.

LiITA adopts the OntoLex-Lemon [2] model as its foundational standard for the representation of lexical resources. This ensures that data is structured according to widely accepted Semantic Web principles, thereby promoting interoperability and reusability. OntoLex-Lemon provides a framework for linking lexical entries to their meanings and to related linguistic properties. LiITA utilizes this framework to establish connections between lemmas in the LB, their occurrences in texts, and their corresponding entries in lexicons and dictionaries. Although the LiITA Knowledge Base primarily focuses on resources related to the Italian language, it is important to acknowledge that Italy is home to a rich array of local languages. Many of these are endangered, predominantly oral, and often lack standardized orthographies. A recent paper [3] offers a critical examination of Italy's linguistic landscape, challenging mainstream Natural Language Processing (NLP) approaches. The study highlights the fragmented and underdeveloped state of NLP research for many Italian language varieties. Given that language inherently encodes local knowledge, cultural traditions, and historical memory, the loss of these varieties entail a significant erosion of cultural heritage. Despite this, the language varieties of Italy are increasingly represented in multilingual NLP initiatives. These include participation in shared tasks on morphological inflection and on language identification (see for example [4]). Additional contributions include cross-lingual word embeddings for low-resource settings and the inclusion of Italian varieties like Lombard, Piedmontese, and Sicilian in multilingual pretrained language models, such as mBERT [5].[1] How-

CLiC-it 2025: Eleventh Italian Conference on Computational Linguistics, September 24 — 26, 2025, Cagliari, Italy

[*] Corresponding author.

[†] This paper is the result of the collaboration between the authors. For the specific concerns of the Italian academic attribution system, Rachele Sprugnoli is responsible for sections 3, 5, 6; Eleonora Litta for section 1; Domenico Giuseppe Muscianisi for sections 2 and 4 (the latter with the aid of Rachele Sprugnoli). Section 4 was collaboratively written by Rachele Sprugnoli and Giuseppe Muscianisi. Giovanni Moretti is responsible for the technical implementation of the Sicilian Lemma Bank and of the modeling of the bilingual lexicon.

rachele.sprugnoli@unicatt.it (R. Sprugnoli); giovanni.moretti@unicatt.it (G. Moretti); domenicogiuseppe.muscianisi@unipr.it (D. G. Muscianisi); eleonoramaria.litta@unicatt.it (E. Litta)

0000-0001-6861-5595 (R. Sprugnoli); 0000-0001-7188-8172 (G. Moretti); 0000-0002-2964-856X (D. G. Muscianisi); 0000-0002-0499-997X (E. Litta)

[1] See [3] for other bibliographical details about these recent efforts.

ever, these varieties remain under-represented in terms of training data volume and quality. On the other hand, a tendency of multilingual NLP to treat language varieties "monolithically", without adequate consideration for their distinct orthographic conventions, sociolinguistic contexts, or community-specific needs, remains. In this light, the integration of bilingual dictionaries and other lexical resources for Italy's minority languages into the LiITA LOD framework would represent a concrete step toward supporting these under-resourced languages. Such inclusion would enhance their digital visibility, promote accessibility, and contribute to the broader goal of preservation and exchange of information on linguistic diversity. The first bilingual glossary to be included in the LiITA KB was the one published in the *Vocabolario della lingua parmigiana* [6]: data in RDF and CSV format together with a set of SPARQL queries are available online [7].[2] This paper, instead, concerns the modeling and linking of the Sicilian Wiktionary (*Wikizziunariu*). More specifically, this paper provides the following three contributions:

1. the modeling of the first Lemma Bank for Sicilian and of a Sicilian-Italian glossary extracted from the *Wikizziunariu* according to the Linguistic Linked Open Data principles;[3]
2. the linking of the glosssary to the KB of the LiITA project[4]
3. the results of two preliminary NLP experiments using the aforementioned bilingual glossary.

2. The Sicilian Dialect

Dialects constitute an essential component of Italy's linguistic heritage. In this study, it is important to clarify the intended meaning of the term *dialect*, which corresponds to the Italian *dialetto*, i.e., a regional or areal language that is genealogically a sister language to so-called Standard Italian, as defined in the *Vocabolario Treccani*[5] (see also [8]). The dialects of Italy are, in fact, independent Romance languages that, over the centuries, have become minoritized local varieties. This shift is primarily attributable to the prestige and diffusion of the *volgare fiorentino* following the works of Dante, Petrarch, and Boccaccio, whose literary influence from the 14th century onward played a central role in shaping the literary language of the Italian Peninsula and, eventually, the codification of present-day Standard Italian. Today, Standard Italian functions as the roof-language for the various Italo-Romance dialects spoken across the country [9]. However, the medieval Sicilian *volgare* was among the first Romance varieties to be used as a literary language, particularly at the court of Emperor Frederick II of Swabia, who established his principal seat in Palermo between 1220 and 1250. During this period, poetry and the arts flourished, giving rise to the *Scuola poetica siciliana*, which Dante, in his *De vulgari eloquentia*, regarded as the earliest manifestation of an "Italian" literary tradition.

According to the *Carta* by Giovan Battista Pellegrini, the Sicilian dialect is placed as group III (*siciliano*) among the Extreme Southern dialects of Italy (henceforth abbreviated as ESI, or *Meridionale Estremo* in Italian), with seven varieties based on the presence of umlaut (*metafonesi*), namely Western Sicilian, Central umlauted area, South-Eastern umlauted area, Original non-umlauted area, Messinese, Aeolian and Pantesco. This classification, along with others that have been proposed (see, for example, [10]), highlights the structural and sociolinguistic complexity of the Sicilian dialect. Moreover, due to its geographical location at the crossroads of the Mediterranean, Sicily has historically been (and continues to be) a site of intense cultural, communal, and linguistic contact [11]. Although the Sicilian dialect retains its core Italo-Romance structural features, it has undergone significant stratification due to successive waves of linguistic contact from Late Antiquity through the Middle Ages. Early layers include influences from (Byzantine) Greek, particularly in eastern Sicily, and from Sicilian Arabic in the west. Subsequent periods of contact include the Norman era (10th–12th centuries) and the reign of Frederick II (ending in 1266), followed by the Angevin rule and the Sicilian Vespers (1282), which introduced Gallo-Romance elements. Later, during the Aragonese and Spanish periods (14th–17th centuries), further Ibero-Romance influences were integrated into the language. Following the medieval period, Sicilian dialects began to evolve into their modern forms. In addition, various linguistic minority communities have historically settled in Sicily. The oldest still active is that of Piana degli Albanesi, the largest Arbëreshë (Italo-Albanian) settlement on the island, established at the end of the 16th century. Another notable case is the Gallo-Italic of Sicily, comprising approximately 15 isolated communities in central and eastern Sicily, whose origins trace back to the Norman period. A third group is the Sicilian Greek community in Messina, officially recognized as a linguistic minority in 2012, which descends from settlers who migrated from the Peloponnese in the mid-16th century. Today, the varieties of Sicilian spoken in these areas exhibit significant influence from these non-Italo-Romance minority languages. The long and complex sociolinguistic history of the Sicilian dialect, together with its internal variation

[2] https://github.com/LiITA-LOD/LocalVarieties/tree/main/Parmigiano

[3] The data in both Turtle RDF and TSV format are available on a dedicated GitHub together with a set of ready-to-use queries: https://github.com/LiITA-LOD/LocalVarieties/tree/main/Siciliano

[4] http://www.liita.it

[5] https://www.treccani.it/vocabolario/dialetto_res-545debd7-0018-11de-9d89-0016357eee51/

and multilingual contact layers, renders it a particularly rich and compelling subject for investigation through computational methods.

2.1. Dictionaries and Grammars of Sicilian

With such a history, the studies on the dialects of Sicily, both in language and culture, show a long-lasting tradition already from the Middle Ages. However, for a comprehensive understanding of the present-day language, the most informative period for the study of Sicilian begins with Italian Romanticism, specifically in the mid to late 19th century. Shortly after the Unification of Italy, Antonino Traina published the *Nuovo vocabolario siciliano–italiano*, a dictionary lemmatized according to Sicilian entries, which provided Italian translations as well as phraseological examples drawn from idiomatic expressions and literary sources, encompassing both cultivated and popular registers [12]. As was typical of the period, Traina's underlying objective was to promote the Tuscan-based national language, thereby contributing to the broader project of fostering social and linguistic unification among the newly formed Italian citizenry. In the same period, the most influential scholar of the Sicilian language and cultural traditions was Giuseppe Pitrè, author of the monumental *Fiabe, novelle e racconti popolari siciliani* [13] and *Grammatica Siciliana* [14]. In his linguistic work, Pitrè approached Sicilian as a Romance language in its own right, analyzing its phonology diachronically from Latin without reference to Tuscan (i.e., Italian), which he explicitly treated as a separate variety rather than a standard of comparison. Both Traina and Pitrè promoted a spelling standardization rooted in Latin orthographic principles. This approach had a dual effect: on the one hand, it contributed to the definition of a kind of Sicilian *koine* (common language), but on the other hand this introduced a bias towards the Latinization of Sicilian [15]. This process of standardization continues to play a fundamental role today. In 2024, the *Cademia Siciliana* (Sicilian Academy) published the *Documento per l'ortografia del siciliano* (Document for the spelling of Sicilian), aiming to be friendly for those who want to write in Sicilian. On the scientific and academic side, the most important linguistic and ethnographic research on Sicilian consists of the pioneering investigation by Franco Fanciullo on the Aeolian Islands [16].

Besides the *Dictionary* by Traina, two other fundamental lexicographic resources for the Sicilian dialect are the *Vocabolario storico-etimologico del siciliano* and the *Vocabolario siciliano*, both published on paper by *Centro di studi filologici e linguistici siciliani.* As far as digital dictionaries are concerned, there is the *Vocabolario del siciliano medievale*[6] of the University of Catania, which collects lemmas of the *volgare siciliano* from the mid 13th to the mid 16th century and provides a Web interface [17]. Within this context of rich historical and linguistic tradition, *Wikizziunariu* emerges as a collaborative resource that is easily accessible, machine-readable, and free from copyright restrictions.

3. Workflow

This work was carried out in two main phases. The first involved parsing a dump of the Sicilian Wiktionary (*Wikizziunariu*) to extract information relevant to our objectives. The second phase focused on modeling and creating resources in RDF format. This latter step includes the construction of a Sicilian LB, the transformation of Wiktionary data into RDF triples, and the linking of Italian translations to the LiITA LB developed within the LiITA project.

3.1. Data Extraction

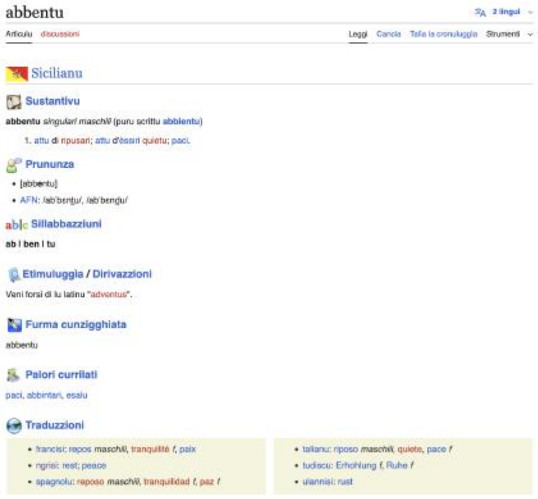
The dump of the Sicilian Wiktionary, downloaded from the Academic Computer Club archive in Umeå,[7] was parsed using a custom script designed to extract relevant data. Figure 1 illustrates the structure of an entry from which the following elements were retrieved: the page title (*abbentu*), the grammatical category (*Sustantivu*, i.e., common noun), number and gender (*singulari maschili*), alternative forms (*puru scrittu abbientu*), and the Italian translation(s) (i.e., values following the label *talianu* in the *Traduzzioni* section, such as *riposo*, *quiete*, *pace*).

The main challenge in the extraction process stemmed from the variability in how information is structured across entries. For example, number and gender may be represented using initials (e.g., *s* for *sostantivo*, noun, *m* for masculine, and *f* for feminine). Furthermore, while alternative forms are always enclosed in parentheses, they are not always preceded by the phrase *puru scrittu*, and the number of translations varies. In some cases, these translations are accompanied by information about the grammatical gender of the Italian equivalents (e.g., *maschili* and *f*, as shown in the figure).

A total of 14,464 entries were extracted through this process, distributed across 20 distinct classes. Twelve of these correspond to traditional grammatical categories: adjectives, adverbs, articles, coordinating conjunctions, interjections, common nouns, proper nouns, numerals, prepositions, pronouns, subordinating conjunctions, and verbs. In addition, the entries included acronyms, confixes, prefixes, suffixes, nominal phrases, multiword ex-

[6] http://artesia.unict.it/vocabolario

[7] https://hammurabi.ftp.acc.umu.se/mirror/wikimedia.org/dumps/backup-index.html



abbentu

Sicilianu

Sustantivu

abbentu *singulari maschili* (puru scrittu **abbientu**)

1. attu di ripusari; attu d'èssiri quietu; paci.

Prununzia

- [abbentu]
- AFN: /ab'bentu/, /ab'bendu/

Sillabbazziuni

ab | ben | tu

Etimuluggìa / Dirivazzioni

Veni forsi di lu latinu "adventus".

Furma cunzigghiata

abbentu

Palori currilati

paci, abbintari, esalu

Traduzzioni

- francisi: repos *maschili*, tranquillité *f*, paix
- ngrisi: rest; peace
- spagnolu: reposo *maschili*, tranquilidad *f*, paz *f*
- talianu: riposo *maschili*, quiete, pace *f*
- tudiscu: Erholung *f*, Ruhe *f*
- uiannisi: rust

Figure 1: Screenshot of an entry in *Wikizziunariu*.

pressions, proverbs, and conjugated verb forms. These latter entries were not included in the subsequent stages of the work, as they cannot be directly mapped to a LB. Table 1 presents the final number of entries considered for each grammatical category and provides example for each category; the original categories have been converted into UPOS (Universal Dependencies Part of Speech) tags [18]. The low number of determiners (DET) is due to the fact that, in the original classification, this category includes only articles, while other types of determiners are assigned to different classes; for example, possessive determiners are categorized as adjectives or pronouns.

Table 1
Number and examples of entries per grammatical category.

NOUN	8302	*puntaperi* (kick), *ràrica* (root)
VERB	2722	*açç̧iari* (to find), *studiari* (to study)
ADJ	1696	*nastenti* (stubborn), *sicilianu* (sicilian)
ADV	477	*nsièmmula* (together), *viatu* (soon)
ADP	340	*cu* (with), *nt'a* (in the)
PRON	152	*iddi* (them), *nui* (we)
NUM	93	*cincu* (five), *sìrici* (sixteen)
PROPN	42	*Cifaru* (Lucifer), *Aropa* (Europe)
INTJ	38	*olè*, *osara*
DET	21	*nu* (a/an), *lu* (the)
SCONJ	10	*mentri* (while), *pirchistu* (therefore)
CCONJ	7	*anchi* (also), *nì* (neither)
TOTAL	13900	

3.2. Data Modeling and Linking

The Sicilian entries were used to build the Sicilian LB. Lemmas are described with the OntoLex model in conjunction with the LiLa ontology. The latter provides a structured representation of the linguistic features of each lemma, including part-of-speech classification, via the `lila:hasPos` property, and grammatical gender, via the `lila:hasGender` property. The total number of lemmas in the Sicilian LB is 10,232. The discrepancy with respect to the number of entries in the *Wikizziunariu* (see Table 1) is primarily due to the fact that some of them are written representations, rather than distinct standalone lemmas. The following RDF triple, expressed in Turtle syntax, represents the Sicilian lemma *middeu*,[8] classified as a masculine noun. It includes multiple written representations (*amiddeu*, *amoddei*, *middeu*, *muddeu*, *muddìu*) each annotated with the language ISO tag `@scn`. These forms are considered orthographic or graphical variants of the same lemma and do not affect its morphological interpretation; all share the same grammatical gender (masculine). Additionally, the lemma is related to a lemma variant identified by an URI[9] corresponding to the lemma *muddìa*.[10] In our example, *middeu* and *muddìa* can be used alternatively but they differ in gender, being the second a feminine noun.

```
<http://liita.it/data/id/
    DialettoSiciliano/lemma/753> a
    lila:Lemma;
  lila:hasGender lila:masculine;
  lila:hasPOS lila:noun;
  lila:lemmaVariant <http://liita.it/
      data/id/DialettoSiciliano/lemma
      /1010>;
  dcterms:isPartOf <http://liita.it/
      data/id/DialettoSiciliano/lemma/
      LemmaBank>;
  rdfs:label "middeu";
  ontolex:writtenRep "amiddeu"@scn, "
      amoddei"@scn, "middeu"@scn, "
      muddeu"@scn, "mudd u"@scn .
```

Subsequently, the bilingual glossary was modeled. The Sicilian lexical entries were linked to the corresponding lemmas in the Sicilian LB via the `ontolex:canonicalForm` property. The Italian translations were connected to the Italian LB developed within the LiITA project using the same property. Furthermore, the lexical entries of the two languages were directly related through the `vartrans:translatableAs` property, which establishes a correspondence between trans-

[8] With URI:http://liita.it/data/id/DialettoSiciliano/lemma/753
[9] http://liita.it/data/id/DialettoSiciliano/lemma/1010
[10] The Property lila:lemmaVariant relates two lemmas that are semantically related to one another but differ in some linguistic feature, such as gender or number.

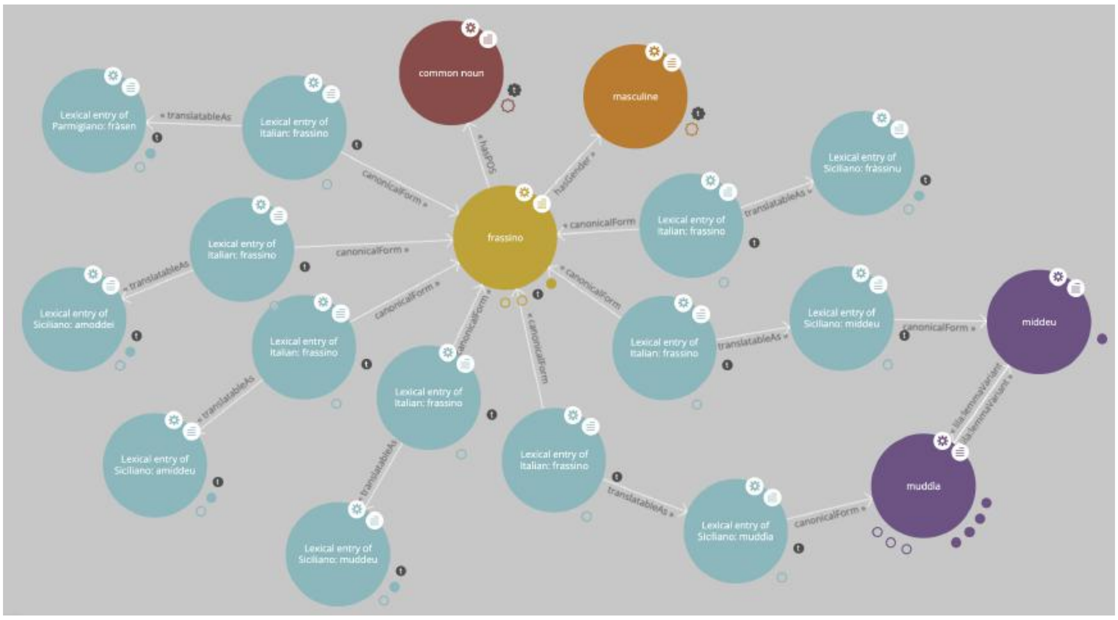

common noun
masculine
Lexical entry of Parmigiano: fràsen
« translatableAs
Lexical entry of Italian: frassino
canonicalForm »
« hasPOS
hasGender »
Lexical entry of Siciliano: fràssinu
translatableAs »
Lexical entry of Italian: frassino
« canonicalForm
frassino
Lexical entry of Italian: frassino
canonicalForm »
Lexical entry of Siciliano: amoddei
« translatableAs
Lexical entry of Italian: frassino
canonicalForm »
canonicalForm »
« canonicalForm
« canonicalForm
Lexical entry of Italian: frassino
translatableAs »
Lexical entry of Siciliano: middeu
canonicalForm »
middeu
Lexical entry of Siciliano: amiddeu
« translatableAs
Lexical entry of Italian: frassino
Lexical entry of Italian: frassino
« lila:lemmaVariant
lila:lemmaVariant »
translatableAs
translatableAs »
Lexical entry of Siciliano: muddeu
Lexical entry of Siciliano: muddia
canonicalForm »
muddia

Figure 2: Lemmas and corresponding translations: the example of *frassino* (ash).

lations. The following RDF triple defines a lexical entry in Italian for the word *frassino* (ash) associated with a canonical form which represents the corresponding lemma in the LiITA LB. Furthermore, this entry is linked to its corresponding Sicilian lexical entry (*middeu*), establishing a cross-lingual correspondence between the Italian and Sicilian lexical resources.

```
<http://liita.it/data/
    LexicalResources/DialettoSiciliano
    /id/LexicalEntry/italian/328>
  a ontolex:LexicalEntry;
  rdfs:label "Lexical entry of
      Italian: frassino";
  ontolex:canonicalForm <http://liita
      .it/data/id/lemma/993692>;
  vartrans:translatableAs <http://
      liita.it/data/LexicalResources/
      DialettoSiciliano/id/
      LexicalEntry/siciliano/753> .
```

Figure 2 displays the lemma *frassino* (ash) as it appears in the LiITA LB, together with information regarding its grammatical gender (masculine) and part of speech (common noun). The node is linked to the lexical entries in the linked lexical resources through the property `ontolex:canonicalForm`. In particular, there are six entries connected via the `vartrans:translatableAs` property related to the Sicilian dictionary and one related to the dialect of Parma. The visualization also shows the `lemmaVariant` relation between *middeu* and *muddìa*.

The linking process with the Italian LB was conducted in two distinct phases. In the initial phase, an automatic alignment was performed between the string of each translation of Sicilian glossary entry and those recorded in the Italian LB, considering the part of speech. This procedure successfully accounted for 55% of the entries. An additional 19% of entries were identified as ambiguous, i.e., a single Italian entry corresponded to multiple lemmas within the LB, thus requiring manual disambiguation. For instance, the entry *caglio*, whose Sicilian translation is *quagghialatti*, could be linked either to the lemma identified by the URI http://liita.it/data/id/lemma/972573, corresponding to the meaning "rennet", or to http://liita.it/data/id/lemma/972574, which refers to a type of herb or artichoke. To resolve such ambiguities, additional information was consulted from *Wikizziunariu* or other Sicilian-language dictionaries.

Currently, 26% of the entries lack a corresponding linking to the Italian LB. These terms include, among others, feminine or plural forms absent from the LB, as well as culturally specific terms unique to the Sicilian context, such as *spènsiri* translated as *largo mantello utilizzato dai contadini* (a wide cloak worn by peasants) or *carpita* translated as *coperta rustica tessuta con ritagli di stoffa* (a rustic blanket woven from fabric scraps).

4. Case Studies

Using SPARQL queries, it is possible to extract linguistically meaningful information from multiple perspectives.[11]

For instance, one can retrieve Sicilian lemmas having written representations beginning with a *d* and an *r*; the complementary distribution [d] ∼ [r] is especially attested in the western variant from Palermo when those sounds appear in intervocalic or initial position. Among such cases is the lemma *dicembri* (December) (< Latin *DECEMBRE(M) ∼ *DECEMBRU(M)) that witnesses several written representations, namely (a) *diciè̀mmuru*, (b) *dicèmmiru*, (c) *dicembru*, (d) *dicèmmuru*, (e) *diciè̀mmiru*, (f) *riciè̀mmiru* and (g) *riciè̀mmuru*. The lemmas (d) and (f) indeed show the aforementioned allophony [d] ∼ [r] but there are also other interesting phenomena. The lemmas (a) and (e) show the *metafonesi* (umlaut) in tonic syllables, i.e. a process of vowel assimilation; the lemmas (a), (b), (d), (e), (f) and (g) witness the lag assimilation of Latin *MB > Sicilian MM [19]. Finally, the lemmas (a), (d) and (g) attest a u-vowel, while the lemmas (b), (e) and (f) an e-vowel: these are epentheses, thus random insertions of one or more sounds to favor the pronunciation.

It is also possible to search for lemmas having written representations that include *ed* or *ied*, an alternation which graphically renders the umlaut of vowels in tonic syllables. This is a significant linguistic phenomenon in Sicilian, serving as a marker for distinguishing dialectal variants. It is generally attested in central and western regions of the island, while it is absent in the north-eastern areas. For example, in the Sicilian word (a) *aceddu* (bird) (< Latin *AU(I)CELLU(M)), the actual pronunciation of *dd* is retroflexed as *ḍḍ* [ɖː] but it is here not represented [14]. This feature is contained in all the following written representations, that is (b) *acieddu*, (c) *ancieddu* and (d) *oceddu*. The tonic syllable is the middle one and witnesses either (1) no changes in lemmas (a) and (d) deriving from Latin *-CE- or (2) umlauted vowels in lemmas (b) and (c) both bisyllabic [ˈɪ.e]. The same phenomenon occurs with, among others, *(ab)bruciareddu* ∼ *(ab)bruciarieddu* (ripe ear), *beddu* ∼ *bieddu* (beautiful), *ciuceddu* ∼ *ciucieddu* (soup, broth, delicacy), *frateddu* ∼ *fratieddu* (brother), *marzamareddu* and *mazzamareddu* ∼ *marzama(u)rieddu*, *mazzumaurieddu* and *mazzamarieddu* (whirlwind, whirlpool, demon), *munzeddu* ∼ *munzieddu* (stack, pile), *pisciteddu* ∼ *piscitieddu* (small fish).

As for morphology, we can search for nouns ending with *-ìa* (< Greek -ía), an abstract suffix which is one of the most common and attested. We can thus notice that the Sicilian suffix is variously represented in Italian translations. More specifically, Sicilian *-ìa* corresponds to the following Italian suffixes:

1. *-ia* (same Greek ía-suffix for abstractivize nominals), as in *ancarìa* ∼ *angheria* (vexation), and *magarìa* ∼ *stregoneria* (witchcraft);
2. *-ità* (< Latin *-ITÁ(TEM)), as in *avracìa* ∼ *altezzosità* (haughtiness) and *liccum(ar)ìa* ∼ *golosità* (delicacy);
3. *-ezza* (< Latin *-ITIA(M) ∼ *-ITIES), as in *laccanìa* ∼ *debolezza* (weakness);
4. various other abstractivizing suffixes, such as *-eccio*, *-io*, *-enza*, *-ita* (with the accent on the antepenultimate syllable).

5. Experiments

Beyond the specific linguistic analyses enabled by interoperability, such as those presented in Section 4, the data we provide can support a variety of experimental applications. A couple of examples are given in the following subsections.

5.1. How much Sicilian do LLMs know?

The bilingual glossary may be used to assess the ability of Large Language Models (LLMs) to translate from Sicilian into Italian. Specifically, we randomly selected 20 nouns, 20 adjectives, 20 verbs, and 20 adverbs, and prompted the main commercially available LLMs to translate each word into Italian. We chose to focus on commercial systems (namely, ChatGPT, Gemini, and Claude) because they are the most widely used by non-experts due to their user-friendly interfaces. A simple zero-shot prompt was employed uniformly across all models: *Traduci ogni parola dal siciliano all'italiano* (Translate each word from Sicilian to Italian). The responses were compared against the translations provided in the glossary and were also evaluated by one of the authors, a linguist and native speaker of Sicilian. This additional human evaluation was intended to determine whether certain translations, even if not identical to those recorded in the resource, could nonetheless be considered acceptable. For example, while the adjective *bacioccu* is officially translated only as *sempliciotto* (nitwit), the alternatives *sciocco* (foolish) (proposed by GPT-4o) and *tonto* (dumb) (provided by Claude Sonnet 4) were considered equally valid. Table 2 presents the results of this evaluation in terms of (synonym-aware) accuracy.

Table 2 reveals not very high accuracies even with synonym tolerance. Gemini 2.5 Flash tops the list at 67% accuracy, about 6 points ahead of GPT o3 and roughly 15 points above Claude 4 Sonnet (51%) and GPT-4o (52%). Even the best-performing model thus mistranslates

[11] Queries can be found in the GitHub repository: https://github.com/LiITA-LOD/LocalVarieties/tree/main/Siciliano. They can be tested on the following endpoint: https://liita.it/sparql.

Table 2
Synonym-aware accuracy on 80 randomly chosen Sicilian words translated into Italian.

	Accuracy
Gemini 2.5 Flash	67%
GPT o3	61%
GPT-4o	52%
Claude 4 Sonnet	51%

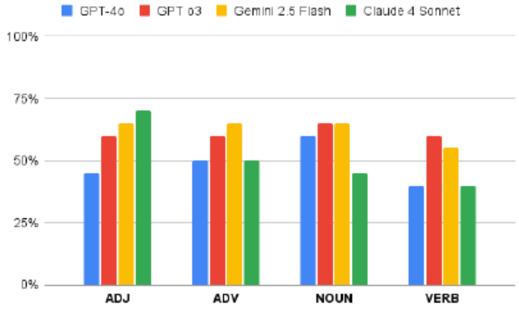


Figure 3: Accuracy per part-of-speech tag.

roughly one word out of three, underscoring how low-resource dialects remain challenging for general-purpose systems. An interesting case is that of GPT o3, which, during the reasoning process, retrieves information from the Web. For certain translations, it explicitly cites its sources, including the *Wikizziunariu*, the vocabulary published on the *TerraLab* blog,[12] and the lexicon curated by the group *Salviamo il siciliano*.[13] This approach leads to better accuracy than the GPT-4o model but still lower than that of Gemini 2.5 Flash.

Two noteworthy observations can be drawn from Figure 3, which shows with a bar chart the accuracy calculated for each part of speech. First, verbs consistently emerge as the most challenging grammatical category to translate across all four models. Second, GPT o3 and Gemini 2.5 Flash exhibit relatively stable performance across categories, whereas Claude 4 Sonnet and GPT-4o show greater variability. However, given the limited sample size of only 80 items, the results are subject to a high sampling error, and the observed differences are not statistically significant. Future work should expand the benchmark and incorporate a broader range of dialectal variants to enable more robust evaluation.

Error analysis shows that 18 words were incorrectly translated by all the systems. More generally, all models exhibit a systematic tendency to infer translations on the basis of superficial orthographic similarity between the Sicilian lemma and a resembling Italian word, which is then selected as the output. For example, *mbròcculi* is rendered as *broccoli*, although its actual meaning is *moina* (flattery), and *pisuliddu* is rendered as *pisellino* (little pea), whereas the intended sense is *permaloso* (touchy).

5.2. Evaluating Bilingual Lexicon Induction

A second experiment used the bilingual glossary to build cross-lingual word embeddings and to evaluate the resulting mapped vectors on the Bilingual Lexicon Induction (BLI) task. Irvine and Callison-Burch [20] define BLI as “the task of inducing word translations from monolingual corpora in two languages.” Although recent work has introduced solutions based on LLMs [21] [22], one of the most widely adopted methods is still to align embeddings trained separately on monolingual corpora into a shared vector space. We therefore applied vecmap[14] [23] in its supervised mode to map Sicilian and Italian fastText embeddings.[15] The glossary was partitioned into training and test sets using a 90:10 ratio after removing homographs and Sicilian lemmas whose Italian equivalents were multi-token expressions, yielding 9,698 Sicilian–Italian pairs for training and 1,079 pairs for testing. Evaluation employed the nearest-neighbor retrieval method (with k=10) and resulted in an accuracy of 19.8% (coverage=50.6%). By using the Cross-domain Similarity Local Scaling (CSLS) retrieval, a cosine-similarity variant that attenuates the hubness problem, namely the tendency of a small subset of vectors to appear disproportionately often as nearest neighbors of other points [24], the result is even lower, i.e., 14.68%. These low scores suggest that, although more than 9.6 K seed pairs are non-trivial for a low-resource variety such as Sicilian, there are many out-of-vocabulary words.

6. Conclusions

This work represents a step toward the integration of the Sicilian dialect into the ecosystem of Linguistic Linked Open Data [25]. By modeling and publishing a bilingual Sicilian–Italian glossary extracted from Wikizziunariu, and by aligning it with the LiITA LB through established ontologies such as OntoLex-Lemon and LiLa, we provide a reusable, interoperable lexical resource that promotes the visibility and accessibility of Sicilian in digital environments. The two preliminary NLP experiments, evaluating LLMs’ translation capabilities and testing BLI, highlight both the potential and the current limitations of applying computational methods to under-resourced varieties.

[12] https://www.terralab.it
[13] http://www.salviamoilsiciliano.com
[14] https://github.com/artetxem/vecmap
[15] https://fasttext.cc/docs/en/crawl-vectors.html

Future work will proceed along multiple directions. First, we plan to model and integrate additional Sicilian resources, with particular attention to Antonino Traina's *Nuovo vocabolario siciliano–italiano*, which is already available in digital format. Second, we aim to broaden the scope of the LiITA KB by incorporating resources from other dialects. An expanded multilingual dataset will enhance interoperability and enable richer cross-lingual analyses. Third, we intend to link textual resources to the LB. However, this will require reliable lemmatization procedures, a non-trivial task for dialects with non-standardized orthographies and scarce annotated corpora. Finally, we plan to extend the range and depth of NLP experiments to evaluate downstream tasks with the goal of advancing computational support for Italy's linguistic diversity.

Acknowledgments

This contribution is partly funded by the European Union - Next Generation EU, Mission 4 Component 1 CUP J53D23017270OO1. The PRIN 2022 PNRR project **"LiITA: Interlinking Linguistic Resources for Italian via Linked Data"** is carried out jointly by the Università Cattolica del Sacro Cuore, Milano and the Università di Torino.

References

[1] E. Litta, M. Passarotti, P. Brasolin, G. Moretti, V. Basile, A. Di Fabio, C. Bosco, The lemma bank of the LiITA knowledge base of interoperable resources for Italian, in: F. Dell'Orletta, A. Lenci, S. Montemagni, R. Sprugnoli (Eds.), Proceedings of the 10th Italian Conference on Computational Linguistics (CLiC-it 2024), CEUR Workshop Proceedings, Pisa, Italy, 2024, pp. 517–522. URL: https://aclanthology.org/2024.clicit-1.61/.

[2] J. P. McCrae, J. Gil, J. Gràcia, P. Bitelaar, P. Cimiano, The OntoLex-Lemon Model: Development and Applications, 2017. URL: https://www.semanticscholar.org/paper/The-OntoLex-Lemon-Model%3A-Development-and-McCrae-Gil/3ab2877e3cf9d8f7bad3a4fb9a03602010e00691.

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

[4] N. Aepli, A. Anastasopoulos, A.-G. Chifu, W. Domingues, F. Faisal, M. Gaman, R. T. Ionescu, Y. Scherrer, Findings of the vardial evaluation campaign 2022, in: Proceedings of the Ninth Workshop on NLP for Similar Languages, Varieties and Dialects, 2022, pp. 1–13.

[5] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers), 2019, pp. 4171–4186.

[6] M. Mori, U. Pavarini, Vocabolario della lingua parmigiana. Tutte le voci e i modi di dire autentici del dialetto parmigiano, Valentino, Parma, 2017.

[7] R. Sprugnoli, D. G. Muscianisi, Linked data and italian dialectology: A new case study on the dialect of parma, To appear in: L'Analisi Linguistica e Letteraria 35 (2025).

[8] J. Van Keymeulen, The dialect dictionary, The handbook of dialectology (2017) 39–56.

[9] M. Loporcaro, Profilo linguistico dei dialetti italiani, volume 275, Laterza, 2013.

[10] G. Ruffino, Sicilia, Laterza, 2001.

[11] Y. Matras, Language contact, Cambridge University Press, 2020.

[12] A. Traina, Nuovo vocabolario siciliano-italiano, volume 1, Lauriel, 1868.

[13] G. Pitrè, Fiabe, novelle e racconti popolari siciliani, Donzelli, 2016.

[14] G. Pitrè, Grammatica siciliana, Pedone Lauriel, 1875.

[15] F. Fanciullo, Il siciliano e i dialetti meridionali, in: Tre millenni di storia linguistica della Sicilia (Atti del Convegno della Società Italiana di Glottologia, Palermo, 25-27 marzo 1983), Giardini Editori, 1984, pp. 139–159.

[16] F. Fanciullo, Dialetto e cultura materiale delle isole Eolie, Palermo: Centro di Studi Filologici e Linguistici Siciliani, 1983.

[17] S. Arcidiacono, Da lexicad a lexichub: note sull'interope-rabilità tra risorse lessicografiche, Quaderni Veneti 13 (2024) 165–174.

[18] M.-C. De Marneffe, C. D. Manning, J. Nivre, D. Zeman, Universal dependencies, Computational linguistics 47 (2021) 255–308.

[19] A. Vàrvaro, Capitoli per la storia linguistica dell'italia meridionale e della sicilia: I. gli esiti di -nd-, -mb-, Medioevo Romanzo 6 (1979) 189–206.

[20] A. Irvine, C. Callison-Burch, A comprehensive analysis of bilingual lexicon induction, Computational Linguistics 43 (2017) 273–310.

[21] Y. Li, A. Korhonen, I. Vulić, On bilingual lexicon induction with large language models, in: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, 2023, pp. 9577–9599.

[22] Y. Li, A. Korhonen, I. Vulić, Self-augmented in-

context learning for unsupervised word translation, in: L.-W. Ku, A. Martins, V. Srikumar (Eds.), Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), Association for Computational Linguistics, Bangkok, Thailand, 2024, pp. 743–753. URL: https://aclanthology.org/2024.acl-short.67/. doi:10.18653/v1/2024.acl-short.67.

[23] M. Artetxe, G. Labaka, E. Agirre, Learning bilingual word embeddings with (almost) no bilingual data, in: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2017, pp. 451–462.

[24] A. Conneau, G. Lample, M. Ranzato, L. Denoyer, H. Jégou, Word translation without parallel data, arXiv preprint arXiv:1710.04087 (2017).

[25] P. Cimiano, C. Chiarcos, J. P. McCrae, J. Gracia, Linguistic Linked Data. Representation, Generation and Applications, Heidelberg, Berlin: Springer, 2020.