An Etymological Dataset for Nouns and Verbs in the Gallo-Italic Variety Spoken in Nicosia and Sperlinga

Domenico Cantone[1], Vincenzo Nicolò Di Caro[2], Cristiano Longo[2], Salvatore Menza[2], Marianna Nicolosi Asmundo[1] and Daniele Francesco Santamaria[1]

[1]*University of Catania, Department of Mathematics and Computer Science, 6, Viale Andrea Doria, Catania, Italy*
[2]*University of Catania, Department of Human Sciences, 32, Piazza Dante, Catania, Italy*

Abstract

This paper presents a FAIR-compliant etymological dataset of noun and verb lemmas from the Gallo-Italic variety spoken in the Sicilian towns of Nicosia and Sperlinga. Based on Trovato and Menza (2020) and various lexicographic materials, the dataset captures borrowing relations from Sicilian, modeled using the OntoLex-lemon framework and its etymological extension. Linguistic phenomena involved in these borrowings—so-called Gallo-Sicilian features—are formalized in the Linguistic Phenomena Ontology (LiPh) as regular relations, enabling the automated generation of candidate derivations. Verified derivations, reviewed by lexicographers, are encoded using LiPh to reflect how individual features contribute to lexical transformations. The dataset is enriched with detailed metadata, geographic and linguistic contextualization, and is aligned with current standards in lexical resource publication, supporting future reuse and integration in the Linguistic Linked Open Data cloud.

Keywords

Semantic Web, OWL, FAIR, Historical Linguistics, Language Contact Theory, Gallo-Italic languages, Sicilian

1. Introduction

In general, a lexical expression in any language may be either *inherited* from its parent language or *borrowed* from a foreign one. Following [1], we define *linguistic phenomena* as processes that modify the morphology or the pronunciation of lexical expressions. These phenomena may occur during both the inheritance and borrowing processes mentioned above. A linguistic phenomenon is considered a *feature* of a given language if it has affected a significant portion of its lexical expressions. For example, a feature of Italian is the change of "t" to "d"—a form of *lenition* (see [2])—which can be observed in the inheritance of Latin *patrem*, which became *padre* in Italian.

Language features are essential tools for investigating *contact-induced changes* (see [3, 4]), that is, changes that occur in a *recipient* language spoken by a population in contact with another population speaking a different language, called the *source* language. To analyze such changes, it is necessary to identify the features retained in both the recipient and the source languages.

As reported in [5, 6], the *Gallo-Italic* varieties spoken in Sicily provide valuable case studies for contact-induced language changes. These languages were introduced in Sicily during the Middle Ages through mass migration from the northwestern Italy and have since shaped by long-term contact with Sicilian. For brevity, we will refer to these varieties as "Gallo-Sicilian" throughout the remainder of this paper.

Here we document an effort to create a dataset of nouns and verbs from the Gallo-Sicilian variety spoken in Nicosia and Sperlinga, based on the lexical material in [7]. The dataset includes etymological
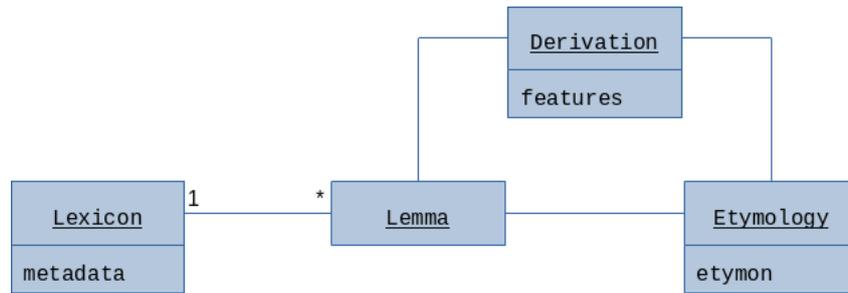
Figure 1: Knowledge domain items

information, expressed in terms of language features, for lexical expressions borrowed from Sicilian, thereby enabling automated analysis of borrowing processes.

From this perspective, we established the following representational goals for our dataset:

[G1] To maximize *findability*, *accessibility*, *interoperability* and *reusability*, the dataset must comply with the **FAIR** principles, as outlined in [8].

[G2] To ensure findability, the entire **lexicon** must be accompanied by detailed and comprehensive **metadata**, facilitating indexing in authoritative linguistic resource catalogs such as the *Linguistic Linked Open Data Cloud* (*LLOD*) [9].

[G3] The dataset must include all **lemmas** for nouns and verbs in the Gallo-Sicilian variety spoken in Nicosia and Sperlinga, as these constitute the primary objects of analysis.

[G4] Since our language-feature detection strategy is grounded in Sicilian, the dataset must include **etymologies** from Sicilian.

[G5] Each etymology must be described through a **derivation** path listing the language **features** involved in transforming the Sicilian **etymon** into the Gallo-Sicilian lemma.

The knowledge domain items referenced in the goals above are summarized in the UML diagram (see [10]) in Figure 1. In the rest of the paper, all the examples are provided using the UML formalism.

The remainder of the paper describes the structure of the dataset and discusses the design choices made. Section 2 outlines the technologies and the knowledge representation formalisms adopted to ensure compliance with the FAIR principles. Section 3 details the metadata used to enhance the findability of the lexicon. Section 4 explains how the dictionary entries from [7] were incorporated into our dataset. Sections 5 and 6 describe how Sicilian etymons and the corresponding language features involved in borrowings are reported in the dataset, respectively. Finally, Section 7 presents our conclusions and outlines directions for future research.

2. Technologies and Overall Dataset Structure

Linked Data technologies (see [11, 12]) provide all the essential features for constructing and publishing a FAIR dataset. Entities in the domain of discourse are uniquely identified by *resolvable* IRIs (see [13]), which allow one to retrieve relevant information using the standard HTTP protocol (see [14]). Data are represented using the formalisms in *Semantic Web stack* such as RDF [15], RDFS [16], and OWL [17]. These technologies support the association of rich metadata with resources and the interlinking of datasets.

Although Linked Data technologies facilitate the creation of FAIR datasets, several design aspects require careful planning (see best practices reported in [18]). In particular, one must determine how the dataset will be published—specifying its IRI and defining a consistent IRI structure for internal resources.

The dataset is provided in the *Turtle* format [19], though the choice of serialization is not critical due to the availability of standard tools for converting between formats. Throughout this paper, Turtle syntax is used for namespaces and prefixes.

As part of the Gallo-Sicilian project, our dataset is provided under the project namespace.[1] In addition, to accommodate future datasets for other Gallo-Sicilian languages, we organized our dataset IRI under the `lexica` sub-namespace, allowing additional vocabularies to be hosted within a coherent and extensible namespace structure.

We adopted *hashed* IRIs, meaning that each internal resource is identified by a fragment IRI within the following namespace:

```
@prefix: <https://gallosiciliani.unict.it/ns/lexica/nicosiaesperlinga#>
```

In the remainder of this paper, this namespace will be considered the default, and omitted for brevity when referring to dataset entities.

A core principle of both FAIR and Linked Data approaches is that data about an item should be retrievable by dereferencing its IRI. In our case, since hashed IRIs are used, a client accessing the IRI of an item will retrieve the entire dataset, and then locate the desired resource via its fragment identifier (i.e., the IRI part after the hash).

Although Linked Data and Semantic Web technologies are *general-purposes*, they support the specification of domain-specific vocabularies and thesauri.

FAIR guidelines also emphasize the importance of *domain-relevant community standards*. In the domain of computational linguistics, [20] provides a comprehensive overview of best practices for representing lexical resources as Linked Data, along with a review of key initiatives and projects in the field.

A widely adopted standard for lexical resource modeling is *OntoLex-lemon*, an OWL vocabulary introduced in [21]. It has become the *de facto* standard for lexical information encoding (see, e.g., [22, 23, 24, 25, 26]). OntoLex-lemon is modular, with each module devoted to a specific aspect of lexical information representation and defined in its own namespace. Several extensions have been proposed to address specialized representational needs. In our case, we use:

- the *Etymological Extension*, introduced in [27] and further discussed in Section 5, which we use to represent etymological relations between Sicilian etymons and Gallo-Sicilian borrowings; and
- the *Liguistic Phenomena Ontology* (LiPh) [1], used to formally capture the linguistic phenomena that occur during the borrowing process.

Our dataset is built upon OntoLex-lemon and the aforementioned extensions. The following sections illustrate how these vocabularies are used to fulfill the representational goals mentioned in Section 1.

While our resource leverages OWL and includes formal vocabularies (such as OntoLex-lemon and LiPh), it is more appropriately described as a *dataset* rather than using the general term *ontology*. This is because it primarily consists of structured lexical and etymological data—specific entries, forms, derivations, and annotations—rather than a general, reusable conceptual schema. Ontologies are indeed used within the dataset to ensure semantic precision and interoperability, but the core content is a collection of domain-specific instances derived from empirical linguistic sources.

3. Lexicon and Lexicon metadata

As stated in Goal **[G2]**, the first component to be described is the lexicon, along with its associated metadata. The OntoLex-lemon module responsible for representing lexica and their metadata is called *lime*. Conceptually, a lexicon is a collection of *lexical entries* for a particular language. Lexical entries will be discussed in detail in Section 4. For the purposes of the present section, it suffices to note that lexical entries correspond to the individual entries of a vocabulary.

The namespace of the lime module is as follows:

---

[1]https://gallosiciliani.unict.it/ns/

```
@prefix lime: <http://www.w3.org/ns/lemon/lime#>
```

Lexica are represented in OntoLex-lemon as instances of the lime:Lexicon class. Each lime:Lexicon instance is linked to its constituent lexical entries via the *lime:entry* property. Accordingly, the individual lexicon in our dataset is declared as an instance of lime:Lexicon.

As reported in [21], various Semantic Web vocabularies can be used to provide basic metadata for instances of lime:Lexicon, such as title, author, publication date, and more (see also [28, 29]). In fact, lime:Lexicon extends the Dataset class from *VoID*, a vocabulary for providing metadata about RDF datasets, as described in [30, 31]. *VoID*, in turn, recommends the use of *Dublin Core Metadata Terms* (see [29]). Accordingly, metadata for our dataset—such as title, description, authorship, license, versioning, and creation date—is provided as properties of the lexicon individual. Furthermore, we cite [7] as *source* of our ontology, since the lexical entries in our lexicon were extracted from it. Lastly, information about project funding is expressed using the *EUropean Research Information Ontology* (*EURIO*), available at [32].

However, OWL itself provides a native mechanism for embedding metadata: every OWL ontology or dataset must include an individual whose IRI matches that of the ontology itself, and metadata must be attached directly to this individual. Since this is an intrinsic feature of OWL, metadata specified in this way is recognized by all Linked Data agents and is commonly used, for example, by platforms such as the *Linked Open Data Cloud*[2] to index ontologies and datasets. For these reasons, metadata for our dataset is provided also as properties of the ontology individual, following the guidelines in [18].

The lime module provides the *lime:language* property to indicate the language of all entries in a lexicon. As is standard in OWL, languages are specified using *language tags*, as defined in [33]. These tags are built from subtags in the IANA Language Subtag Registry.[3] Unfortunately, this registry does not include a specific tag for Gallo-Italic as a whole—only for certain varieties such as *Piemontese*. Therefore, we use the tag `mis`, which denotes an unencoded language.

Alongside the *lime:language* datatype property, the OntoLex-lemon specification recommends the use of the corresponding object property from the Dublin Core vocabulary (see [29]) *dcterms:language*, where the *dcterms* stands for the Dublin Core terms namespace.

```
@prefix dcterms: <http://purl.org/dc/terms/>
```

In our case, this property is used to link the lexicon to the Glottolog entry for Gallo-Italic languages. *Glottolog*, described in [34], is a repository of language identifiers and metadata, with a particular focus on lesser-resourced and under-documented languages.

The language of lexical expressions in our lexicon is further specified through geographic characterization. To this end, our lexicon is linked, via the Dublin Core *dcterms:coverage* property, to entities representing the towns of Nicosia and Sperlinga in the *Geonames* ontology (see [35]).

Finally, the *lime:linguisticCatalog* property is used to indicate the catalog of linguistic categories adopted in the lexicon to describe the grammatical and lexical properties of its entries. Since OntoLex-lemon is agnostic with respect to category systems, it encourages the reuse of established external vocabularies. In our dataset, we adopt the category system provided by *LexInfo*, as described in [36] and reiterated in [21]. The following prefix declaration is used for referencing *LexInfo* elements:

```
@prefix lexinfo: <http://www.lexinfo.net/ontology/3.0/lexinfo#>
```
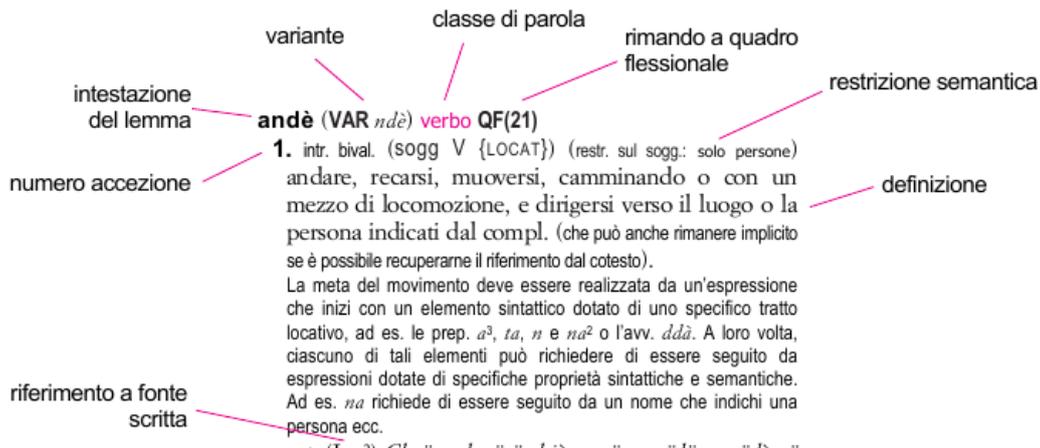
4. Gallo-Sicilian Lemmas

Dictionary entries are represented in OntoLex-lemon as instances of ontolex:LexicalEntry, where ontolex is the namespace for elements of the OntoLex-lemon core module

---

[2]https://lod-cloud.net
[3]https://www.iana.org/assignments/language-subtag-registry

Figure 2: Entry of the source PDF dictionary

```
@prefix ontolex: <http://www.w3.org/ns/lemon/ontolex#>
```

For brevity, we will refer to instances of this class simply as "entries". All entries in our dataset were retrieved from [7] by parsing the corresponding PDF file, excluding those that could not be classified as either nouns or verbs. An example of an entry from [7] is shown in Figure 2.

For our purposes, each entry must be associated with both a *lemma* and a *part of speech.*

A lemma is a specific *grammatical form* of a lexical expression, typically serving as the canonical or citation form in dictionaries. For instance, the lemma reported in the dictionary entry shown in Figure 2 is **andè**. The criteria for identifying lemmas vary by language. For example, in English, the singular form of a noun is conventionally used as the lemma.

Grammatical forms are represented in OntoLex-lemon as instances of the ontolex:Form class. Among the forms associated with a lexical entry, the one that corresponds to the lemma is specified using the *ontolex:canonicalForm* property. Each form must include at least one *written representation*, expressed via the datatype property *ontolex:writtenRep*. The range of this property is the datatype rdf:langString, as defined in [16].[4] As a consequence, every written representation must be provided as a string with an accompanying language tag consistent with the language declared for the lexicon (in our case, mis).

As noted in Section 3, our lexicon adopts the LexInfo category system. Accordingly, each entry is assigned a part of speech by connecting the ontolex:LexicalEntry individual to the appropriate LexInfo one—lexinfo:noun for nouns and lexinfo:verb for verbs—using the *lexinfo:partOfSpeech* property.

Parts of speech are retrieved from [7] by examining the *word class* field (*classe di parola* in Figure 2), which is visually distinguished in magenta. These word classes convey information beyond just part of speech—for example, they may also indicate mood, gender, number, and so on. The source dictionary contains approximately 400 distinct values for this field. Each of these has been manually recognized as indicating a noun, a verb, or another type of entry that falls outside the scope of this work (such as adjectives). For instance, the word class "verbo" in the entry shown in Figure 2 corresponds to the lexinfo:verb individual.

As a result of the parsing and conversion of the entries in [7], our dataset contains 10193 instances of ontolex:LexicalEntry. Figure 3 illustrates how the entry of Figure 2 has been encoded in our dataset.

---

[4]The prefix rdf refers to the canonical RDF namespace: http://www.w3.org/1999/02/22-rdf-syntax-ns#
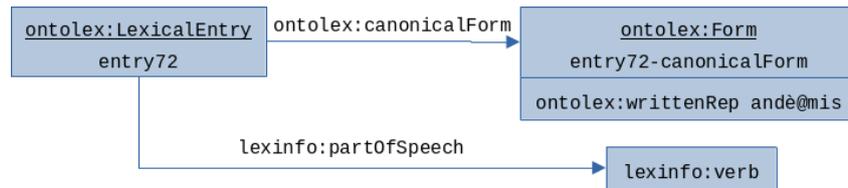
Figure 3: OntoLex-lemon encoding of **andè**

5. Etymologies

As anticipated in Section 2, etymologies for Gallo-Sicilian lemmas borrowed from Sicilian are represented using the OntoLex-lemon Etymological Extension, *lemonEty*, introduced in [27]. Elements of the LemonEty vocabulary are defined within the following namespace:

```
@prefix lemonEty: <http://lari-datasets.ilc.cnr.it/lemonEty#>
```

In this section, we briefly summarize the lemonEty representational features employed in our dataset. Etymologies are modeled as instances of the lemonEty:Etymology class. Lexical entries are linked to their corresponding etymologies via the *lemonEty:etymology* property. Note that multiple lemonEty:Etymology instances may be associated with the same entry, as they represent alternative "hypothesis about the history of a given lexical element," in the sense of [27].

In lemonEty, etymologies can be represented as ordered lists, enabling the reconstruction of the full genealogy of a lexical expression through a sequence of borrowing and inheritance steps. These lists are modeled using the class lemonEty:EtyLink. Each lemonEty:EtyLink instance describes a single etimological step and is characterized by a *source* (the etymon) and a *target* (the derived expression in the recipient language), both of which must be instances of ontolex:LexicalEntry. These are linked to the lemonEty:EtyLink individual via the properties *lemonEty:etySource* and *lemonEty:etyTarget*, respectively. An etymology, modeled as an instance of lemonEty:Etymology, is then connected to the first step of the sequence through the property *lemonEty:startingLink*.

In general, it is also appropriate to relate also the forms involved in the inheritance or borrowing process represented by a lemonEty:EtyLink. For example, the Italian lemma "padre" originates from the borrowing of the Latin accusative form "patrem," derived from "pater." To express this, lemonEty provides two additional properties for lemonEty:EtyLink, which can be used to indicate the specific forms involved: *lemonEty:etySubSource* and *lemonEty:etySubTarget*.

Following Goal [G4], each etymology in our dataset corresponds to a single borrowing from a Sicilian expression. Thus, each entry representing a Gallo-Sicilian lemma identified as a borrowing from Sicilian is associated with a corresponding lemonEty:Etymology individual, which in turn is linked to a lemonEty:EtyLink individual representing the borrowing process. Each such lemonEty:EtyLink individual is structured as follows:

- the dataset entry under consideration is specified as *lemonEty:etyTarget*;
- *lemonEty:etySubTarget* is assigned the canonical form of the entry;
- an ontolex:Form individual representing the Sicilian etymon is used as *lemonEty:etySubSource*.

Notice that the Sicilian etymon is not necessarily a Sicilian lemma, as borrowings may involve any grammatical form. For this reason, the lexical entry corresponding to the etymon is left unspecified, and only the relevant form is included to serve our purposes.

Figure 4 illustrates how the borrowing of the Gallo-Sicilian lemma "còpela" from the Sicilian etymon "còppula" is represented in our dataset using the lemonEty vocabulary.
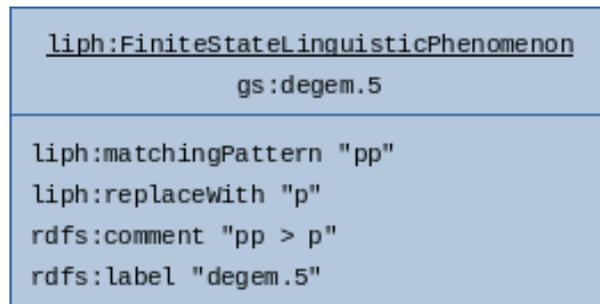
Figure 4: Etymology of "còpela" in lemonEty

6. Gallo-Sicilian Features

As discussed in Section 5, lexical expressions may undergo specific *linguistic phenomena* during borrowing processes (in the sense of [1]). *Gallo-Sicilian features*—that is, phenomena characterizing the borrowing of Sicilian expressions into Gallo-Sicilian languages—have been identified over the years (see [6, 5]) and encoded in the Gallo-Sicilian Features Ontology.[5] All the features in this ontology are represented as individuals within the following namespace:

```
@prefix gs: <https://gallosiciliani.unict.it/ns/gs-features#>
```

These features are expressed using the Linguistic Phenomena Ontology (LiPh), introduced in [1], which adopts the following namespace:[6]

```
@prefix liph: <https://gallosiciliani.unict.it/ns/liph#>
```

LiPh allows the association of each feature with a *regular relation* between strings (see [37]), by declaring the corresponding individual as an instance of the class liph:FiniteStateLinguisticPhenomenon, and specifying:

- a regular expression, via the property *liph:matchingPattern*, identifying the portion of the source string to be modified, and
- a replacement string, by means of *liph:replaceWith*, indicating how the matched part is to be transformed.

For example, the feature gs:degem.5 replaces every occurrence of "pp" with a single "p". The regular expression for gs:degem.5 is "pp" and the replacement "p". This representation in LiPh is shown in Figure 5.

---

[5]https://gallosiciliani.unict.it/ns/gs-features

[6]Note that the Gallo-Sicilian Features Ontology is based on the latest version of LiPh, which slightly differs from the one described in [1].

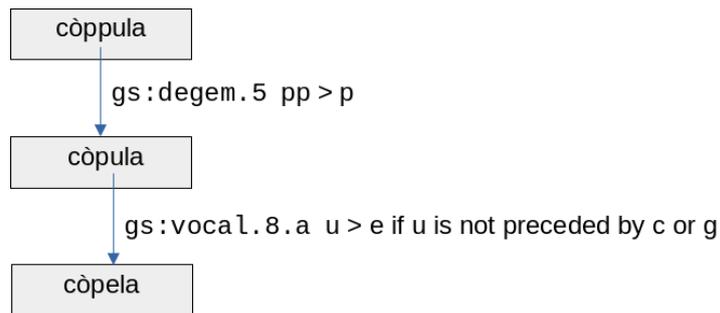**Figure 5:** Encoding of the Gallo-Sicilian feature degem.5

Figure 6: Derivation of "còpela" from "còppula"

This modeling approach enables automation tasks such as the semi-automatic identification of *candidate* etymologies and *derivations*—that is, sequences of linguistic phenomena applied to transform a source expression into its borrowed form. Consider, for example, the borrowing of the Sicilian "còppula", which becomes "còpela" in Gallo-Sicilian. This transformation involves the application of two Gallo-Sicilian features:

- gs:degem.5, which *degeminated* "pp" into a single "p", and
- gs:vocal.8.a, which modifies the vowel "u" into "e".

One of the (two) possible derivation paths of from "còppula" to "còpela" proceeds as follows: gs:degem.5 first transforms "còppula" into the intermediate form "còpula," and then gs:vocal.8.a changes "còpula" into the Gallo-Sicilian lemma "còpela." This derivation is illustrated in Figure 6.

We leveraged the encoding of Gallo-Sicilian features as regular relations to generate a set of candidate derivations capable of transforming Sicilian expressions into Gallo-Italic lemmas included in our dataset. Specifically, we adopted a brute-force approach: all possible combinations of Gallo-Sicilian features were applied to the Sicilian lexical expressions extracted from [38, 39, 40, 41, 42].[7] If a derivation resulted in a lemma present in our dataset, it was recorded as a candidate derivation for that lemma, and the originating Sicilian expression was identified as a candidate etymon. To reduce redundancy, only the *shortest* derivations—i.e., those with the minimal number of transformation steps—were retained for each lemma. All automatically generated derivations were subsequently reviewed by lexicographers, who evaluated their historical and linguistic plausibility, discarding those deemed implausible.

The set of verified derivations was incorporated into our dataset using LiPh. In addition to representing linguistic phenomena, LiPh provides classes and properties for describing, in fine detail, how these phenomena operate during borrowing and inheritance processes.

Each derivation step is modeled as an instance of the class liph:LinguisticPhenomenaOccurrence, which is connected to:

- the corresponding linguistic phenomena via the property *liph:isOccurrenceOf*, and

---

[7]For computational feasibility, Gallo-Sicilian features were applied in a fixed order.
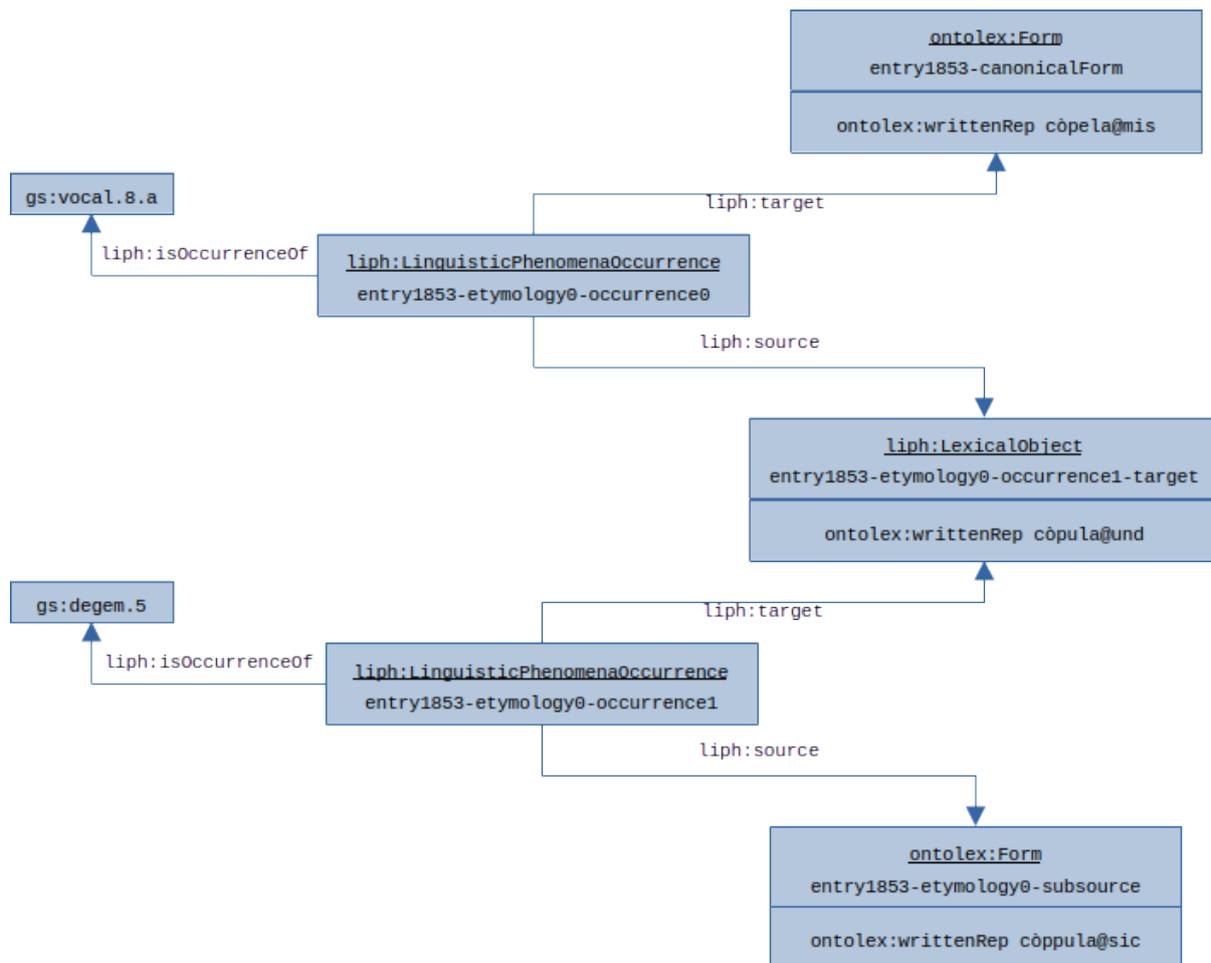
Figure 7: Derivation of "còpela" from "còppula" in LiPh

- the source and resulting lexical expressions—whether actual or intermediate forms—through the properties *liph:source* and *liph:target*, respectively.

Consider, for example, the first step in the derivation illustrated in Figure 6. The corresponding liph:LinguisticPhenomenaOccurrence instance is linked to the gs:degem.5 phenomenon. The *liph:source* property is set to the actual form representing the Sicilian etymon "còppula" (as shown in Figure 4), while the *liph:target* points to a new individual representing the intermediate form "còpula", modeled as an instance of the class liph:LexicalObject , which is used for representing such intermediate forms. Notably, the language tag associated with "còpula" is set to und (undetermined), following the IANA language subtag registry, as it is not possible to ascertain whether this intermediate form was ever used by actual speakers during a specific historical period.

Figure 7 illustrates how the full derivation from "còppula" to "còpula", as previously shown in Figure 6, is encoded in our dataset using LiPh.

7. Conclusions and Future Work

We have introduced a structured, machine-readable dataset of noun and verb lemmas from the Gallo-Italic variety spoken in Nicosia and Sperlinga, focusing on borrowings from Sicilian and the linguistic phenomena involved. The dataset leverages Semantic Web technologies to ensure FAIR compliance and adopts community-driven standards such as OntoLex-lemon, lemonEty, and LiPh to support rich, interpretable representations of lexical structure and etymological derivations.

By encoding relevant features as regular relations, we enable semi-automated derivation generation, which offers both practical utility and theoretical insight into contact-induced language changes. All candidate derivations were validated by expert lexicographers, ensuring their historical and linguistic plausibility.

This resource fills a critical gap in the digital documentation of under-resourced Gallo-Italic varieties and lays the groundwork for further computational and linguistic research on contact phenomena, diachronic processes, and dialectal variation within these speech communities in Sicily.

In addition, we argue that the methodology adopted to identify derivations, as well as the design choices made, can be effectively applied to the study of contact phenomena involving any pair of source and recipient languages.

Future work includes (i) extending the dataset to additional Gallo-Sicilian varieties and other grammatical categories, (ii) refining the brute-force derivation engine through heuristic pruning and statistical weighting, and (iii) integrating usage frequency and temporal data to distinguish attested forms from hypothetical intermediate.

We also plan to publish SPARQL endpoints, potentially equipped with a visual exploration tool such as that described in [43], to support comparative research in contact linguistics.

Finally, our dataset will be included in the Gallo-Sicilian project's data catalog, which is provided using the DCAT vocabulary as described in [28].

Acknowledgments

Declaration on Generative AI

During the preparation of this work, the authors used ChatGPT in order to grammar and spelling check, paraphrase and reword. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] D. Cantone, V. N. Di Caro, C. Longo, S. Menza, M. Nicolosi Asmundo, D. F. Santamaria, An OWL Ontology for linguistic phenomena with applications to Gallo-Italic dialects in Sicily, in: A. Bikakis, R. Ferrario, S. Jean, B. Markhoff, A. Mosca, M. Nicolosi Asmundo (Eds.), Proceedings of the fourth edition of the International Workshop on Semantic Web and Ontology Design for Cultural Heritage, Tours, France, October 30-31, 2024, volume 3809 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2024, pp. 12–24. URL: https://ceur-ws.org/Vol-3809/paper2.pdf.

[2] G. Marotta, Phonetics and Phonology, Cambridge Handbooks in Language and Linguistics, Cambridge University Press, 2022, p. 200.

[3] S. G. Thomason, T. Kaufman, Language contact, creolization and genetic linguistics, 1988.

[4] F. van Coetsem, A General and Unified Theory of the Transmission Process in Language Contact, Winter, 2000.

[5] S. C. Trovato, Galloitalische sprachkolonien. I dialetti galloitalici della Sicilia, Kontakt, Migration Und Kunstsprachen 7 (1998) 538–559.

[6] A. De Angelis, The strange case of the gallo-italic dialects of sicily: Preservation and innovation in contact-induced change, Languages 8 (2023). URL: https://www.mdpi.com/2226-471X/8/3/163. doi:10.3390/languages8030163.

[7] S. C. Trovato, S. Menza, Vocabolario del dialetto galloitalico di Nicosia e Sperlinga, number 39 in Materiali e ricerche dell'Atlante Linguistico della Sicilia, Centro di studi filologici e linguistici siciliani, 2020.

[8] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. B. da Silva Santos, P. E. Bourne, J. Bouwman, A. J. Brookes, T. Clark, M. Crosas, I. Dillo, O. Dumon, S. Edmunds, C. T. Evelo, R. Finkers, A. Gonzalez-Beltran, A. J. Gray, P. Groth, C. Goble, J. S. Grethe, J. Heringa, P. A. 't Hoen, R. Hooft, T. Kuhn, R. Kok, J. Kok, S. J. Lusher, M. E. Martone, A. Mons, A. L. Packer, B. Persson, P. Rocca-Serra, M. Roos, R. van Schaik, S.-A. Sansone, E. Schultes, T. Sengstag, T. Slater, G. Strawn, M. A. Swertz, M. Thompson, J. van der Lei, E. van Mulligen, J. Velterop, A. Waagmeester, P. Wittenburg, K. Wolstencroft, J. Zhao, B. Mons, The FAIR guiding principles for scientific data management and stewardship, Scientific Data 3 (2016) 160018–. URL: https://doi.org/10.1038/sdata.2016.18.

[9] C. Chiarcos, B. Klimek, C. Fäth, T. Declerck, J. P. McCrae, On the linguistic linked open data infrastructure, in: G. Rehm, K. Bontcheva, K. Choukri, J. Hajic, S. Piperidis, A. Vasiljevs (Eds.), IWLTP@LREC, European Language Resources Association, 2020, pp. 8–15. URL: http://dblp.uni-trier.de/db/conf/lrec/iwltp2020.html#ChiarcosKFDM20.

[10] Object Management Group, OMG Unified Modeling Language Specification Version 2.5.1, https://www.omg.org/spec/UML/2.5.1/, 2017.

[11] World Wide Web Consortium, Linked data - design issues, 2006. URL: https://www.w3.org/DesignIssues/LinkedData.html.

[12] C. Bizer, T. Heath, T. Berners-Lee, Linked data - the story so far, International Journal on Semantic Web and Information Systems 5 (2009) 1–22. URL: http://www.igi-global.com/article/linked-data-story-far/37496. doi:10.4018/jswis.2009081901.

[13] M. Duerst, M. Suignard, Internationalized Resource Identifiers (IRIs), RFC 3987, 2005. URL: https://rfc-editor.org/rfc/rfc3987.txt. doi:10.17487/RFC3987.

[14] R. Fielding, J. Reschke, Hypertext Transfer Protocol (HTTP/1.1): Message Syntax and Routing, RFC 7230, 2014. URL: https://rfc-editor.org/rfc/rfc7230.txt. doi:10.17487/RFC7230.

[15] F. Manola, E. Miller (Eds.), RDF Primer, W3C Recommendation, World Wide Web Consortium, 2004. URL: http://www.w3.org/TR/rdf-primer/.

[16] D. Brickley, R. Guha, RDF Schema 1.1 - W3C Recommendation 25 February 2014, Technical Report, World Wide Web Consortium (W3C), 2014. URL: http://www.w3.org/TR/rdf-schema/.

[17] P. Hitzler, M. Krötzsch, B. Parsia, P. Patel-Schneider, S. Rudolph, OWL 2 Web Ontology Language Primer, W3C Recommendation, World Wide Web Consortium, 2009. URL: http://www.w3.org/TR/owl2-primer/.

[18] D. Garijo, M. Poveda-Villalón, Best practices for implementing FAIR vocabularies and ontologies on the Web, CoRR abs/2003.13084 (2020). URL: http://dblp.uni-trier.de/db/journals/corr/corr2003.html#abs-2003-13084.

[19] D. Beckett, T. Berners-Lee, G. Carothers, E. Prud'hommeaux, RDF 1.1 Turtle, W3C Recommendation, W3C, 2014. URL: http://www.w3.org/TR/2014/REC-turtle-20140225/.

[20] A. F. Khan, C. Chiarcos, T. Declerck, D. Gifu, E. González-Blanco García, J. Gracia, M. Ionov, P. Labropoulou, F. Mambrini, J. P. McCrae, É. Pagé-Perron, M. Passarotti, S. Ros Muñoz, C.-O. Truica, When linguistics meets web technologies. Recent advances in modelling linguistic linked data, Semantic Web 13 (2022). URL: https://doi.org/10.5281/zenodo.7129494. doi:10.5281/zenodo.7129494.

[21] P. Cimiano, J. McCrae, P. Buitelaar, Lexicon Model for Ontologies: Community Report, Technical Report, W3C, 2016. URL: https://www.w3.org/2016/05/ontolex/.

[22] Y. M. Abgaz, Using OntoLex-lemon for representing and interlinking lexicographic collections of Bavarian dialects, in: M. Ionov, J. P. McCrae, C. Chiarcos, T. Declerck, J. Bosque-Gil, J. Gracia (Eds.), LDL@LREC, European Language Resources Association, 2020, pp. 61–69. URL: http://dblp.

uni-trier.de/db/conf/acl-ldl/acl-ldl2020.html#Abgaz20.

[23] T. Declerck, L. Bajcetic, M. Siegel, Adding pronunciation information to Wordnets, in: T. Declerk, I. Gonzalez-Dios, G. Rigau (Eds.), MMW@LREC, The European Language Resources Association (ELRA), 2020, pp. 39–44. URL: http://dblp.uni-trier.de/db/conf/lrec/mmw2020.html#DeclerckBS20.

[24] C. Chiarcos, C. Fäth, M. Ionov, The ACoLi dictionary graph, in: Proceedings of The 12th Language Resources and Evaluation Conference, 2020, pp. 3281–3290.

[25] S. Racioppa, T. Declerck, Porting the Latin WordNet onto Ontolex-lemon, in: I. Kosem, M. Cukr, M. Jakubíček, J. Kallas, S. Krek, C. Tiberius (Eds.), Electronic lexicography in the 21st century. Proceedings of the eLex 2021 conference, 2021, pp. 429–439.

[26] D. Lindemann, S. Ahmadi, A. F. Khan, F. Mambrini, F. Iurescia, M. C. Passarotti, When OntoLex meets Wikibase: Remodeling use cases, in: L.-A. Kaffee, S. Razniewski, K. Alghamdi, H. Arnaout (Eds.), Wikidata@ISWC, volume 3640 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2023, p. 14. URL: http://dblp.uni-trier.de/db/conf/wikidata/wikidata2023.html#LindemannAKMIP23.

[27] A. F. Khan, Towards the representation of etymological data on the Semantic Web, Information 9 (2018). URL: https://www.mdpi.com/2078-2489/9/12/304. doi:10.3390/info9120304.

[28] P. Archer, Data catalog vocabulary (dcat) (w3c recommendation), Online, 2014. URL: https://www.w3.org/TR/vocab-dcat/.

[29] DCMI Usage Board, DCMI Metadata Terms, DCMI Recommendation, Dublin Core Metadata Initiative, 2006. URL: http://dublincore.org/documents/2006/12/18/dcmi-terms/, published online on December 18th, 2006 at http://dublincore.org/documents/2006/12/18/dcmi-terms/.

[30] R. Cyganiak, M. Hausenblas, void guide - using the vocabulary of interlinked datasets, 2009. URL: http://rdfs.org/ns/void-guide, http://rdfs.org/ns/void-guide (Last visit 22/4/2010).

[31] K. Alexander, R. Cyganiak, M. Hausenblas, Describing Linked Datasets with the VoID Vocabulary, Technical Report, W3C, 2011.

[32] P. O. of the European Union, Eurio: European research information ontology, 2022. URL: http://publications.europa.eu/resource/dataset/eurio.

[33] A. Phillips, M. Davis, BCP 47 – Tags for Identifying Languages, BCP 47 Standard, see http://www.rfc-editor.org/rfc/bcp/bcp47.txt, 2006. URL: http://www.rfc-editor.org/rfc/bcp/bcp47.txt.

[34] S. Nordhoff, Linked data for linguistic diversity research: Glottolog/langdoc and ASJP, in: C. Chiarcos, S. Nordhoff, S. Hellmann (Eds.), Linked Data in Linguistics. Representing and Connecting Language Data and Language Metadata, Springer, Heidelberg, 2012, pp. 191–200. URL: http://www.springer.com/computer/ai/book/978-3-642-28248-5. doi:10.1007/978-3-642-28249-2.

[35] M. Wick, Geonames ontology, 2015. URL: http://www.geonames.org/about.html.

[36] P. Buitelaar, P. Cimiano, P. Haase, M. Sintek, Towards linguistically grounded ontologies, in: 6th Annual European Semantic Web Conference (ESWC2009), 2009, pp. 111–125. URL: http://www.cimiano.de/Publications/2009/eswc09/eswc09.pdf.

[37] S. Eilenberg, Automata, Languages, and Machines, Academic Press, Inc., USA, 1976.

[38] G. Piccitto, G. Tropea, S. C. Trovato, VOCABOLARIO SICILIANO I (A-E), number 1 in Vocabolario Siciliano, Centro di studi filologici e linguistici siciliani, 1977.

[39] G. Piccitto, G. Tropea, S. C. Trovato, VOCABOLARIO SICILIANO II (F-M), number 2 in Vocabolario Siciliano, Centro di studi filologici e linguistici siciliani, 1985.

[40] G. Piccitto, G. Tropea, S. C. Trovato, VOCABOLARIO SICILIANO III (N-Q), number 3 in Vocabolario Siciliano, Centro di studi filologici e linguistici siciliani, 1990.

[41] G. Piccitto, G. Tropea, S. C. Trovato, VOCABOLARIO SICILIANO IV (R-Sg), number 4 in Vocabolario Siciliano, Centro di studi filologici e linguistici siciliani, 1997.

[42] G. Piccitto, G. Tropea, S. C. Trovato, VOCABOLARIO SICILIANO V (Si-Z), number 5 in Vocabolario Siciliano, Centro di studi filologici e linguistici siciliani, 2002.

[43] T. Francart, Sparnatural: A visual knowledge graph exploration tool., in: C. Pesquita, H. Skaf-Molli, V. Efthymiou, S. Kirrane, A. Ngonga, D. Collarana, R. Cerqueira, M. Alam, C. Trojahn, S. Hertling (Eds.), ESWC (Satellite Events), volume 13998 of *Lecture Notes in Computer Science*, Springer, 2023, pp. 11–15. URL: http://dblp.uni-trier.de/db/conf/esws/eswc2023s.html#Francart23.