

Modeling Linguistic Phenomena in Gallo-Italic Dialects: The Linguistic Phenomena Ontology

DOMENICO CANTONE ^a, MICHELE COSENTINO ^c
VINCENZO NICOLÒ DI CARO ^b, CRISTIANO LONGO ^{a,1},
SALVATORE MENZA ^b, MARIANNA NICOLOSI ASMUNDO ^a, and
DANIELE FRANCESCO SANTAMARIA ^a

^a *University of Catania, Department of Mathematics and Computer Science, Italy*

^b *University of Catania, Department of Human Sciences, Italy*

^c *University of Messina, Department of Ancient and Modern Civilizations, Italy*

ORCID ID: Domenico Cantone <https://orcid.org/0000-0002-1306-1166>, Vincenzo Nicolò Di Caro <https://orcid.org/0000-0003-2604-2172>, Marianna Nicolosi Asmundo <https://orcid.org/0000-0003-4456-5110>, Daniele Francesco Santamaria <https://orcid.org/0000-0002-4273-6521>

Abstract. Each lexical expression in a language has its own history. The elements of a language are the result of several linguistic processes and may originate from expressions of other languages. In this paper, we present an ontology designed to describe *linguistic phenomena*, in particular, those that have influenced the pronunciation or morphology of lexical elements over time. The ontology, called *The Linguistic Phenomena Ontology*, is applied to the study of some Gallo-Italic varieties spoken in Sicily (Italy), with a focus on the linguistic phenomena that have generated some lexical expressions in these languages by borrowing them from Sicilian. Furthermore, an operational description based on regular expressions is provided for many of these phenomena.

Keywords. Semantic Web, OWL, Historical Linguistics, Language Contact Theory, Sicilian, Gallo-Italic Languages

1. Introduction

Languages constantly evolve. Typically, the elements of a language result from linguistic processes that fall in two main categories: *inheritance* and *borrowing*.

Inheritance refers to lexical expressions passed down from a parent language, as seen in many Romance languages, which descend from Latin. Borrowing, in contrast, occurs when a recipient language adopts lexical items from a source language (see [1] for a comprehensive discussion on this topic).

In other words, each lexical expression in a language has its own *history*: it may have been inherited from a parent language or may have been borrowed from another;

¹Corresponding Author: Cristiano Longo, cristianolongo@opendatahacklab.org.

it may be derived from other expressions by mixing or truncation; its morphology or pronunciation may have changed over time.

However, languages spoken by a specific population or in a specific geographic area are strongly characterized by some recurring linguistic phenomena. For example, the modification of “t” to “d”, which is a particular form of *lenition* (see [2]), occurred, for example, for the Latin “patrem” that has become “padre” in Italian, a typical phenomenon in the Italian language.

We denote as *linguistic phenomena* those kinds of phenomena that cause modifications of language expressions, in particular, those that occurred during the inheritance and borrowing of lexical expressions, which are of interest for historical linguistics and which changed the expressions from a morphological, phonetic, or phonological point of view. In contrast, phenomena concerning language expressions’ semantics (e.g., *sense shift*) are not considered here.

Moreover, we say that a linguistic phenomenon is a *feature* of a specific language if it occurred in the generation of a significant number of lexical elements of that language.

Eliciting linguistic phenomena and language features is worthwhile for historical-linguistics, in particular for studying *contact-induced changes*, i.e., changes that occurred in a recipient language spoken by a population that came into contact with a population speaking the source language (see also [3]).

In this paper, we propose *The Linguistic Phenomena Ontology*, an ontology to represent linguistic phenomena, thus enabling automated analysis on languages and linguistic phenomena in them. This ontology has been used to encode features of some Gallo-Italic varieties that are spoken in Sicily as a consequence of medieval immigration from North-Western Italy. As reported in [4], these dialects represent an excellent testing ground for language contact theory, due to their long-standing interaction with the Sicilian language.

This work extends a previous contribution [5], in which the first version of the ontology was presented. A detailed comparison between that version and the current one, namely version 2.0.0, is presented in Section 4.

The rest of the paper is structured as follows. Section 2 provides a brief overview of technologies and studies in the field of computational linguistics that could be related to the representation and usage of linguistic phenomena. Section 3 presents notions and concepts required to understand the proposed ontology. *The Linguistic Phenomena Ontology* is presented and explained in Section 4. Section 5 describes the use-cases for the ontology concerning Gallo-Italic varieties. Finally, in Section 6, we conclude with final observations and directions for future work.

2. Related Works

Digital encoding of lexical information has been a thriving research topic over the last decades (see, for example, [6,7,8,9]). During this time, *OntoLex-lemon*, introduced in [10] and briefly described in Section 3.1, has become a well-established standard for representing lexical information (see, for example, [11,12,13,14,15]). Among all the extensions proposed for *OntoLex-lemon*, the *OntoLex-lemon Etymological Extension*, described in [16] and recalled in Section 3.2, is of particular interest for historical linguistics.

The linguistic phenomena related to word formation, declension, and inflection are core topics of study within the field of *morphology*. In [17], the authors describe an ongoing effort to develop an *OntoLex-lemon* module for linguistic morphology, called *OntoLex-Morph*. This module provides a class, `morph:Rule`, which describes how grammatical rules are applied to generate new lexical expressions from existing ones. Such rules can be described with regular expressions through the class `morph:Replacement`, using a representational pattern substantially equivalent to the one presented in Section 4.6. However, at the current stage of development, derivation details cannot be described, in particular, those involving intermediate forms.

More in general, several *Rule-based language technologies*, which could benefit from an OWL encoding, have been devised during the years (see [18,19]). In particular, modeling *phonological rules*, as described in [19], using the representational features introduced in Section 4.6, appears to be relatively straightforward.

In this paper, we focus on phenomena relevant to historical linguistics; however, other types of phenomena should also be considered, such as those studied in morphology and briefly mentioned above.

It is worth noting that the Turtle syntax [20] is adopted in this paper for namespaces and their corresponding prefix abbreviations.

3. Preliminaries

This section is devoted to the presentation of notions, concepts, and ontologies required to understand *The Linguistic Phenomena Ontology*. Specifically, we shortly introduce *OntoLex-lemon* and the *OntoLex-lemon Etymological Extension*.

3.1. *OntoLex-lemon*

OntoLex-lemon is an OWL [21] ontology that provides rich linguistic grounding for ontological models, including the representation of morphological and syntactic properties of lexical expressions, as well as the meaning of these expressions with respect to an ontology. It is divided into several modules, each one defining its own namespace. Here we restrict our attention to the core module `ontolex`:

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

Essentially, *lexica* (e.g., dictionaries) are defined in *OntoLex-lemon* as sets of *lexical entries*. These are, in turn, represented as instances of the class `ontolex:LexicalEntry`, or of one of its subclasses `ontolex:Word`, `ontolex:MultiwordExpression`, and `ontolex:Affix`. A lexical entry is characterized by a *lemma* and a *part-of-speech*, as a lemma may take on different meanings when used in different parts of speech (consider, for example, the English lemma “love”). An `ontolex:LexicalEntry` is associated with the corresponding lemma through the *ontolex:canonicalForm* property. Lemmas are represented as instances of `ontolex:Form`, each of which must have at least one written representation, specified by the *ontolex:writtenRep* datatype property, and may have zero or more phonetic representations, indicated by the *ontolex:phoneticRep* datatype property. In addition, an `ontolex:LexicalEntry` instance can be associated with other grammatical form

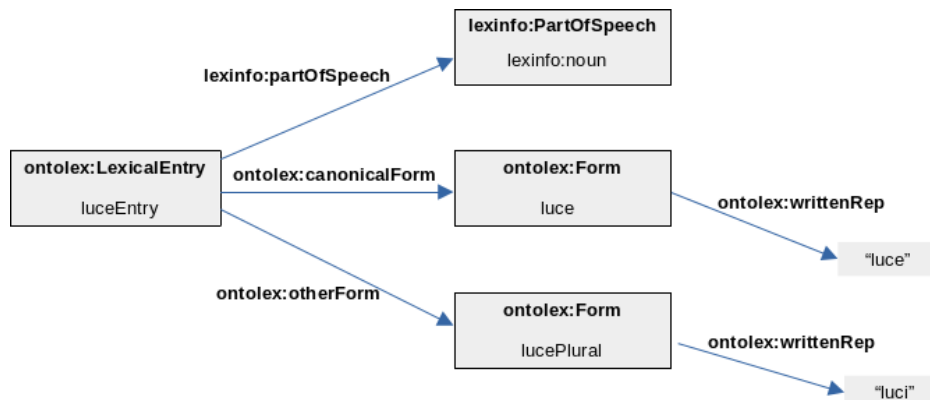


Figure 1. OntoLex-lemon representation for the Italian lexical expression “luce”

variants (e.g., plural form, feminine form), also represented by `ontolex:Form` by means of the `ontolex:otherForm` property.

OntoLex-lemon does not prescribe a specific representation mechanism for parts of speech; instead, it requires the use of a category system defined for the lexicon.

In this paper we will consider the category system provided by the OWL ontology *LexInfo*, described in [22] and recalled in [10], which defines the following namespace:

```
@prefix lexinfo: <http://www.lexinfo.net/ontology/3.0/lexinfo#> .
```

The *LexInfo* ontology provides the property `lexinfo:partOfSpeech` to associate a lexical entry with a specified part of speech. Parts of speech are, in turn, provided explicitly in *LexInfo* as individuals such as `lexinfo:noun`, `lexinfo:verb`, etc.

In Figure 1, we present a representation of the Italian expression “luce” in *OntoLex-lemon*, which includes the lemma and its plural form.

Now that we have reviewed the core features of *OntoLex-lemon*, we turn to its extension for representing etymologies.

3.2. The *OntoLex-lemon Etymological Extension*

The *OntoLex-lemon Etymological Extension*, abbreviated as *lemonEty*, provides specific representational features for etymological information, as detailed in [16]. It defines the following name-space:

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

In *lemonEty*, etymologies are defined as *hypotheses about the history of lexical elements*. In particular, both simple and complex etymologies, which trace the hypothetical provenance of a lexical expression through inheritance and borrowings, can be described in a graph-shaped fashion using the class `lemonEty:EtyLink`.

Notice that *lemonEty* also provides an alternative pattern for representing etymologies, which models them as ordered lists of derivations using the class `lemonEty:EtyList`;



Figure 2. Application of *truncation*, deriving the Italian “luce” from the Latin “lucem”

this pattern is not considered in the present work. For further details, the reader is referred to [16].

An `lemonEty:EtyLink` instance connects a source `ontolex:LexicalEntry` instance with a target one (i.e., the *derived* entry) by means of the properties `lemonEty:etySource` and `lemonEty:etyTarget`, respectively. In addition, the properties `lemonEty:etySubSource` and `lemonEty:etySubTarget` can be used to establish more specific links between lexical elements, such as `ontolex:Form` individuals referring to the entries involved in the `lemonEty:EtyLink`.

For example, consider the etymology of the Italian lexical expression “luce”, illustrated in Figure 2, which descends from the accusative form “lucem” of the Latin lemma “lux”. As shown in Figure 3, an `lemonEty:EtyLink` representing this derivation would include:

- an `ontolex:LexicalEntry` corresponding to “lux” as `lemonEty:etySource`,
- an `ontolex:Form` with written representation “lucem” as `lemonEty:etySubSource`,
- an `ontolex:LexicalEntry` corresponding to “luce” as `lemonEty:etyTarget`, and
- its canonical form, with written representation “luce”, as `lemonEty:etySubTarget`.

Thus, in brief, with *lemonEty* one can state that a specific form descends from another form of a different language. However, with this ontology one can just provide a textual justification, intended for humans, for etymologies.

We claim that the *Linguistic Phenomena Ontology*, which will be introduced in the next section, can be used to provide more detailed and automatically verifiable derivations between forms, thereby offering stronger empirical support for etymologies. For example, explicitly stating that “luce” can be derived from “lucem” through the *truncation* feature of Section 4.1, as illustrated in Figure 3, would reinforce the etymology presented in the previous example.

4. The Linguistic Phenomena Ontology

Connecting etymologies to the linguistic phenomena that characterized them can be valuable for historical-linguistic researchers. This would foster the realization of tools for the automatic discovery of language features from etymologies, or even to discover possible etymologies of a lexical expression. Above all, representing in a detailed and automatically verifiable way the derivations between the forms involved in an etymology in terms of linguistic phenomena would provide solid evidence for the etymology itself.

For these reasons, we developed the *Linguistic Phenomena Ontology*, which provides representational features for describing linguistic phenomena and derivations, in the sense of Section 4.1.

To begin with, we provide a precise definition to clarify our intended use of the term *linguistic phenomena*. We argue that the definition of linguistic phenomena presented in Section 4.1 is sufficiently general to encompass phenomena studied in other areas of linguistics, such as inflection and declension. However, in this work, we primarily

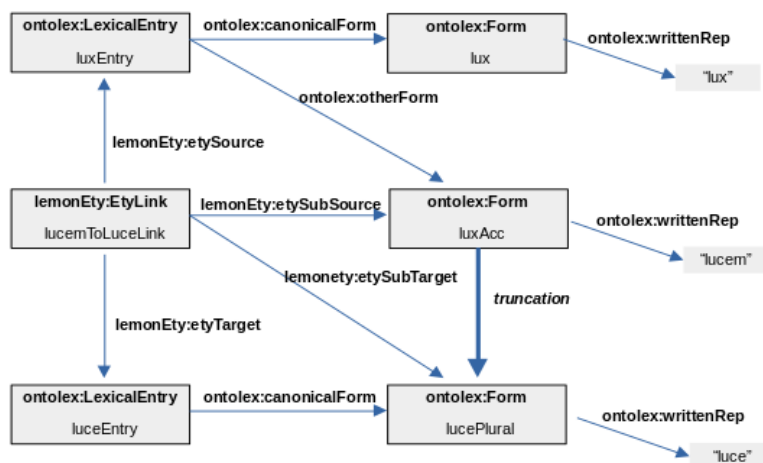


Figure 3. Etymology for the Italian lexical expression “luce” expressed using etyLink and LiPh

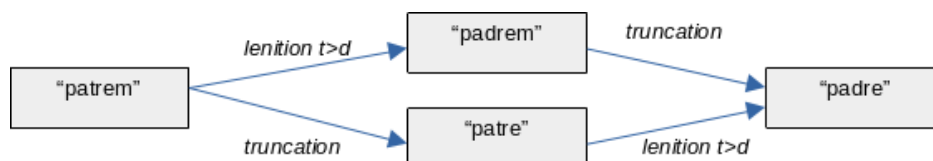


Figure 4. Two possible derivations of “padre” from “patrem”

focus on linguistic phenomena arising in the contexts of inheritance and borrowing, while leaving open the possibility of future applications to other linguistic domains.

4.1. Definitions

As mentioned earlier, a linguistic phenomenon occurs when an expression from a *source* language is transformed into one in a *recipient* language. At first glance, these phenomena sound like just relations between lexical expressions.

For example, one could define a relation *truncation*, which removes the final “m” from Latin expressions. As shown in Figure 2, this *truncation* can then be used to relate the Latin expression “luce” to the Italian “luce”.

In general, multiple modifications may occur in the linguistic process that generated the target expression. Consider, for example, the Latin expression “patrem”, which evolved into the Italian “padre”. The latter may be derived from the former through two distinct phenomena: a relation representing changes of “t” to “d”, which we name *lenition t > d*, and a *truncation*. Moreover, “padre” can be derived from “patrem” via these two phenomena in two different ways, as illustrated in Figure 4.

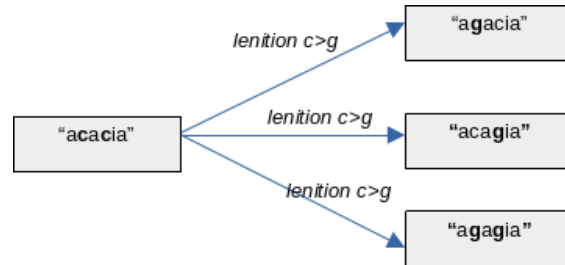


Figure 5. Multiple occurrences of a language feature

Notice that intermediate expressions, such as “padrem” and “patre” in the example, are not necessarily lexical expressions of any language. For expressions that do not belong to an actual *intermediate* language, it would be inappropriate to consider their *pronunciation*. In contrast, if such an expression was adopted by a significant number of speakers during a specific time period, its pronunciation could be inferred from the grapheme-to-phoneme pronunciation rules of the source and recipient language. In light of this, our linguistic phenomena should be defined more as **relations between strings**, where *strings* are understood, as usual, to mean finite sequences of characters from any alphabet.

In other words, our linguistic phenomena are relations over orthographical forms. As a consequence, phenomena which just alter the pronunciation of symbols without affecting the written representation are out of the scope of this paper.

Following this definition, both *lenition $t > d$* and *truncation* in Figure 4 are linguistic phenomena.

It is worth noting that linguistic phenomena are defined as relations rather than functions, as they may affect multiple parts of the same expression and can apply to all of them, not just a subset. For example, consider another form of lenition, typical of languages of northern Italy, in which every occurrence of ‘c’ is changed to ‘g’; we refer to this as *lenition $c > g$* . It applies to two different parts of the Latin expression “acacia”. Thus, the values associated to “acacia” by *lenition $c > g$* are “agacia”, “acagia”, and “agagia”, as shown in Figure 5.

Building on this notion of linguistic phenomena, we define the concept of **derivation** through linguistic phenomena as follows: we say that x *derives* y if there exists a (possibly empty) chain of occurrences of linguistic phenomena starting with x and ending with y .

We conclude this section by pointing out that linguistic phenomena modeled using our ontology are *perdurant* properties of forms and, more generally, of strings. In light of this, representation models concerning *diachronic* data, such as those summarized in [23], have not been considered.

Once the notions of linguistic phenomenon and derivation have been established, we outline the scope and representational goals of our ontology in the following section.

4.2. Scope

Our ontology primarily aims to represent linguistic phenomena that occur during inheritance and borrowing processes and that affect the written form of an expression, while remaining general enough to allow applications to other types of phenomena.

In order to precisely define the scope of our ontology, we define some *competency questions*, i.e., a set of questions in natural language that a dataset using our ontology will need to answer. In Section 4.5, our ontology will be tested with presenting some SPARQL [24] queries to fulfill these questions.

[CQ1] Given two forms l and e , what linguistic phenomena, if any, allow l to be derived from e ?

[CQ2] Given a linguistic phenomenon p , what are the pairs (l, e) of forms such that there exists a derivation from e to l that involves p ?

[CQ1] is informative about the modeling purposes of our ontology. When our ontology is used in conjunction with *lemonEty*, it may be rephrased in terms of etymologies as follows: “Given an etymology from *etymon* to *lemma*, is there a suitable set of linguistic phenomena that derives *lemma* from *etymon*?”. Instead, **[CQ2]** is a seminal example of the potential usages of our ontology for quantitative analysis.

The next sections are devoted to presenting and discussing in full detail all the features of our ontology.

4.3. The Ontology

In this section we describe the *Linguistic Phenomena Ontology* (in short, *LiPh*). Before presenting the ontology’s classes and properties, we first introduce the relationships of *LiPh* with other OWL ontologies.

LiPh targets lexical forms, which are defined in *OntoLex-lemon* as members of the class `ontolex:Form`. As a consequence, *LiPh* has been devised as an *OntoLex-lemon* extension. Conversely, the ontology does not extend *lemonEty*, but an extension of *lemonEty* in *LiPh* can be easily framed.

All the *LiPh* entities are defined within the following name-space:

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

The core of *LiPh* is the object property *liph:derives*, which serves as a super-property encompassing all linguistic phenomena. This property encodes the notion of derivation through linguistic phenomena introduced in Section 4.1.

In addition, it is defined as transitive and reflexive: transitivity just enforces the modeling of linguistic phenomena chains; instead, due to reflexivity, every lexical expression derives first from itself. As will be shown in Section 4.5, this property is particularly helpful for query tasks focused on specific phenomena but potentially involving others, such as those disclosed by **[CQ2]**. In such cases, *liph:derives* can serve as a placeholder for those parts of the derivation that involve phenomena outside the scope of the task.

To allow for the modeling of other types of phenomena (e.g., those related to lexical expression semantics) and to preserve the generality implied by its name, we do not restrict the applicability of this property to any specific class of objects. However, as we will see, its use is constrained when applied to lexical forms.

Following Section 4.1, this property must be suitable for connecting instances of `ontolex:Form`. For example, it may connect `ontolex:Form` instances representing the Italian expression “luce” and the Latin expression “lucem”, as illustrated in Figure 2. However, we emphasize that the class `ontolex:Form` represents grammatical realizations of

Table 1. Linguistic Phenomena Ontology base constraints

$\text{ontolex:Form} \sqsubseteq \text{liph:LexicalObject}$ $\text{ontolex:writtenRep} \sqsubseteq \text{liph:writtenRep}$ $\text{liph:LexicalObject} \sqsubseteq \exists(\text{liph:writtenRep}).(\text{rdf:langString})$ $\text{liph:derives} \equiv \text{liph:derives}^*$ $\text{liph:LexicalObject} \sqsubseteq \forall(\text{liph:derives}).(\text{liph:LexicalObject})$
--

lexical entries, which are, in turn, characterized as members of specific languages. In contrast, linguistic phenomena may also involve objects that are not expressions of any particular language—such as “padrem” and “patre” in Figure 4. Notice that these intermediate forms may have been used by some speakers during a limited period of time. In this case, the usage of `ontolex:Form` could be considered appropriate, as they would be expressions of a sort of an intermediate language. However, to cope with cases of intermediate forms whose effective adoption cannot be firmly established, we introduced the more general class `liph:LexicalObject`. This class represents expressions involved in linguistic phenomena, regardless of whether they belong to any language—such as the endpoints of a lexical derivation—or not, as in the case of certain intermediate forms. This class encompasses `ontolex:Form` and preserves the constraints stated in *OntoLex-lemon* about it, in particular the existential restriction concerning the `ontolex:writtenRep` datatype property. As anticipated before, `ontolex:writtenRep` has to be generalized as well, as *OntoLex-lemon* restricts its usage to `ontolex:Form`. In addition, applications of `liph:derives` to `liph:LexicalObject` individuals are restricted to individuals of the same class. This determines that this property, when applied to forms, concerns just the form level and not the lexical entries one.

Notice that `ontolex:Form` individuals may have several written representations. The following definition clarifies how to deal with linguistic phenomena occurring between objects of this kind: given an object property p_ρ dedicated to represent a linguistic phenomenon ρ , and given two instances i and j of `liph:LexicalObject`, i is related to j through p_ρ if and only if there exist representations r_i and r_j of i and j , respectively, such that $\rho(r_i) = r_j$.

All entities of *LiPh* defined up to now can be concisely summarized using the *Description Logic* syntax [25] through the constraints presented in Table 1, where `rdf` is an abbreviation for the RDF namespace [26].

4.4. Reified Linguistic Phenomena

Linguistic phenomena can be defined just as subproperties of `liph:derives`. For example, Figure 6 illustrates how to represent the two possible derivations of the Italian “padre” shown in Figure 4, assuming that *lenition* $t > d$ and *truncation* are defined as subproperties of `liph:derives`.

While this approach is fully compliant with our ontology, it has significant limitations. Specifically, it avoids providing *semantically relevant* descriptions—i.e., statements which affect reasoning—for phenomena and their occurrences, as doing so would lead to undecidability (see [27]). Instead, our ontology adheres to OWL-DL, a fragment of OWL for which decidability has been proven (see [28]). Examples of semantically relevant descriptions of phenomena will be discussed in Section 4.6. Concerning phenomena occurrences, one may, for instance, report the time period or geographic region

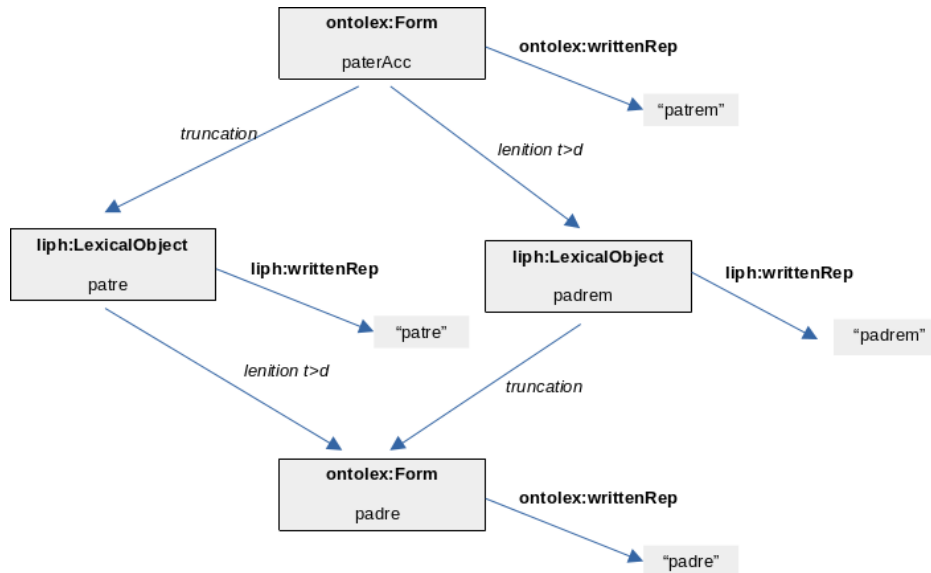


Figure 6. Derivations of “padre” from “patrem” with Linguistic Phenomena Ontology

in which the phenomenon presumably took place. We addressed this issue by adopting the *reification* approach presented in [25], which closely resembles the modeling of `lemonEty:EtyLink` individuals in *lemonEty*. In brief, both linguistic phenomena and their occurrences are reified as distinct individuals. This allows them to become first-class citizens of the knowledge domain, enabling further connections to other individuals and classes to enrich their descriptions. To this end, we created two classes:

- `liph:LinguisticPhenomenon` for all the reified phenomena, and
- `liph:LinguisticPhenomenonOccurrence` for the individuals representing their occurrences.

Then, occurrences of a phenomenon are connected to the individual representing it via the functional property `liph:occurrenceOf`. Finally, each occurrence is bound to the source and produced `liph:LexicalObject` through the functional properties `liph:source` and `liph:target`, respectively. All of these classes and properties are summarized by the constraints listed in Table 2.

This modeling pattern is exemplified in Figure 7, which presents an encoding of the derivation of Figure 2 in *LiPh*. This example illustrates an inferred `liph:derives` property assertion—indicated with a dashed arrow—relating the source and target of a phenomenon occurrence. This inference results from the property chain constraint in the ontology, which enforces our reification mechanism. In this way, `liph:LinguisticPhenomenonOccurrence` instances can act as derivation members.

A straightforward consequence of this constraint, in conjunction with those concerning `liph:LexicalObject` in Table 1, is that any linguistic phenomenon occurrence with an `liph:LexicalObject` individual specified as source must have a target of the same class.

Further examples of this modeling pattern will be examined in the next section.

Table 2. Linguistic Phenomena Ontology reification constraints

$\text{liph:LinguisticPhenomenon} \sqcap \text{liph:LexicalObject} \equiv \perp$
$\text{liph:LinguisticPhenomenonOccurrence} \sqcap \text{liph:LexicalObject} \equiv \perp$
$\text{liph:LinguisticPhenomenon} \sqcap \text{liph:LinguisticPhenomenonOccurrence} \equiv \perp$
$\text{Func}(\text{liph:occurrenceOf})$
$\text{Func}(\text{liph:source})$
$\text{Func}(\text{liph:target})$
$\text{liph:LinguisticPhenomenonOccurrence} \sqsubseteq \forall(\text{liph:occurrenceOf}).(\text{liph:LinguisticPhenomenon})$
$\text{liph:LinguisticPhenomenonOccurrence} \sqsubseteq \exists(\text{liph:occurrenceOf}).\top$
$\text{liph:LinguisticPhenomenonOccurrence} \sqsubseteq \exists(\text{liph:source}).\top$
$\text{liph:LinguisticPhenomenonOccurrence} \sqsubseteq \exists(\text{liph:target}).\top$
$\text{liph:source}^{-1} \circ \text{liph:target} \sqsubseteq \text{liph:derives}$

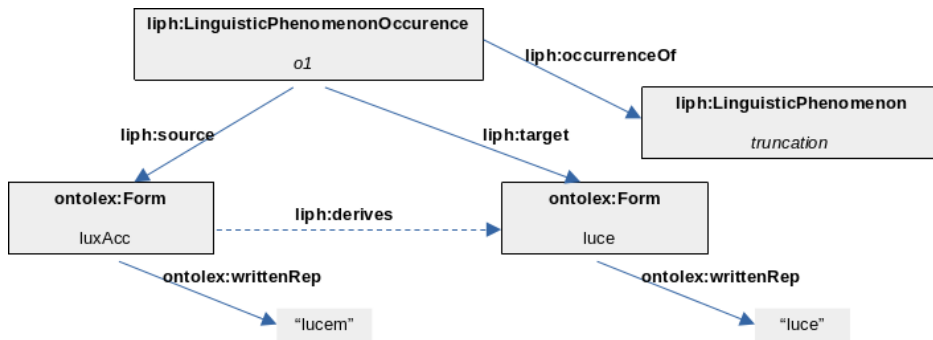


Figure 7. Reified representation of an occurrence of truncation

4.5. SPARQL Queries

This section is devoted to evaluating *LiPh* against the competency questions outlined in Section 4.2. To this end, we present a set of SPARQL query examples that address them.

Query executions are illustrated with respect to a test dataset, shown in Figure 8, which models the first derivation from Figure 4. In particular, two `liph:LinguisticPhenomenon` individuals representing the *lenition* $t > d$ and *truncation* phenomena are reported. Then, two `liph:LinguisticPhenomenonOccurrence` individuals are used to represent the modification of “patrem” into “padrem” through *lenition* $t > d$, and of “padrem” into “padre” through *truncation*, respectively. Finally, `liph:LinguisticPhenomenonOccurrence` individuals are appropriately bound to the corresponding `liph:LinguisticPhenomenon` ones.

Figure 9 lists the property assertions inferred from the test dataset, together with the constraints present in Table 2 and Table 1. First, `paterAcc` is related to `padrem`, which is in turn related to `padre`, through the *liph:derives* property, as a result of the property chain constraint in Table 2. Next, `paterAcc`, `padrem` and `padre` are each related to themselves due to the reflexivity of *liph:derives*. Finally, *liph:derives* also connects `paterAcc` and `padre`, since it is transitive.

Now, consider the query [CQ1], which asks for the linguistic phenomena involved in the derivations that transforms a specific etymon into a given lemma. The SPARQL query below answers query [CQ1] for the case where `paterAcc` is the etymon and `padre`

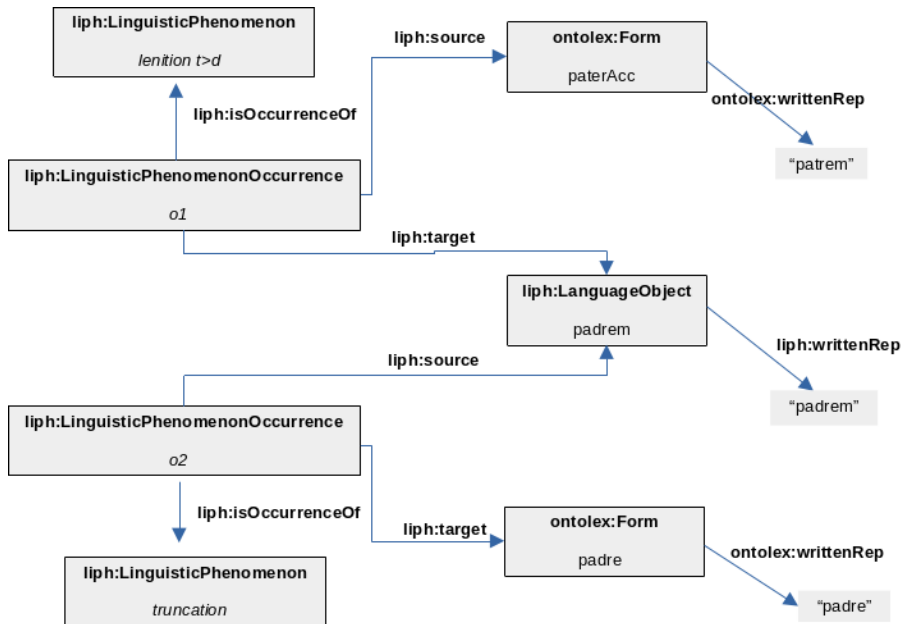


Figure 8. An example representation of the derivation from “padrem” to “padre”

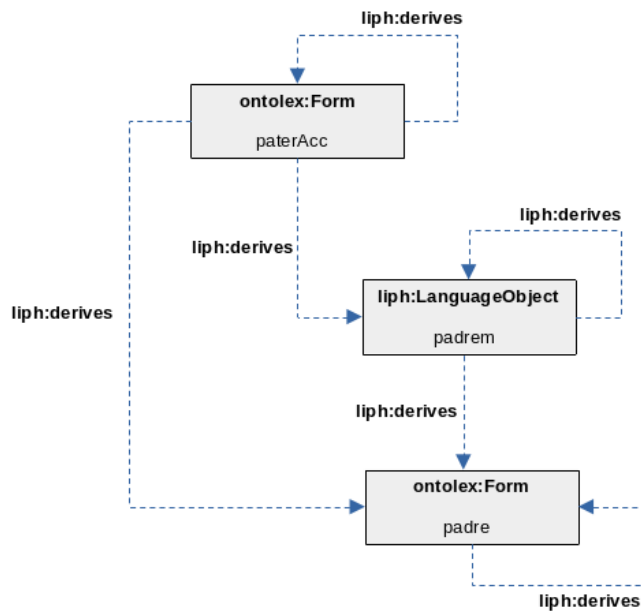


Figure 9. Inferred property assertions

Table 3. SPARQL query matches

?x	?o	?p	?y
paterAcc	o1	<i>lenition t > d</i>	patrem
patrem	o2	<i>truncation</i>	padre

is the lemma, assuming that a reasoner is available and the constraints defined in Table 2 and Table 1 are enforced. This query can be easily generalized to any pair of etymon e and lemma l by replacing `paterAcc` and `padre` with e and l , respectively.

```
PREFIX : <https://gallosiciliani.unict.it/examples/derivation#>
PREFIX liph: <https://gallosiciliani.unict.it/ns/liph#>
```

```
SELECT ?p WHERE {
  :paterAcc liph:derives ?x .
  ?o liph:source ?x ;
    liph:occurrenceOf ?p ;
    liph:target ?y .
  ?y liph:derives :padre
}
```

When executed on the test dataset, this query returns the two matches listed in Table 3 and also depicted in Figure 10, corresponding to the expected linguistic phenomena *lenition t > d* and *truncation*.

This query highlights that *liph:derives* is suitable to be used as a place-holder for derivation portions. This feature can be applied to [CQ2] to devise a SPARQL query to retrieve all the etymon and lemma pairs such that the derivation from etymon to lemma involves a specific phenomenon, say *truncation*.

```
PREFIX : <https://gallosiciliani.unict.it/examples/derivation#>
PREFIX liph: <https://gallosiciliani.unict.it/ns/liph#>
```

```
SELECT ?e ?l WHERE {
  ?e liph:derives ?x .
  ?o liph:source ?x ;
    liph:occurrenceOf :truncation ;
    liph:target ?y .
  ?y liph:derives ?l
}
```

By virtue of the constraints defined in our ontology, this query is capable of extracting all pairs $(?e, ?l)$ such that there exists a derivation from $?e$ to $?l$ involving *truncation*. Let us consider a generic derivation from $?e$ to $?l$ consisting of n linguistic phenomenon occurrences, where *truncation* is the i -th occurrence. This can be summarized in Turtle as follows:

```
:o1 liph:source ?e;
    liph:target :x1;
    liph:occurrenceOf :p1.
```

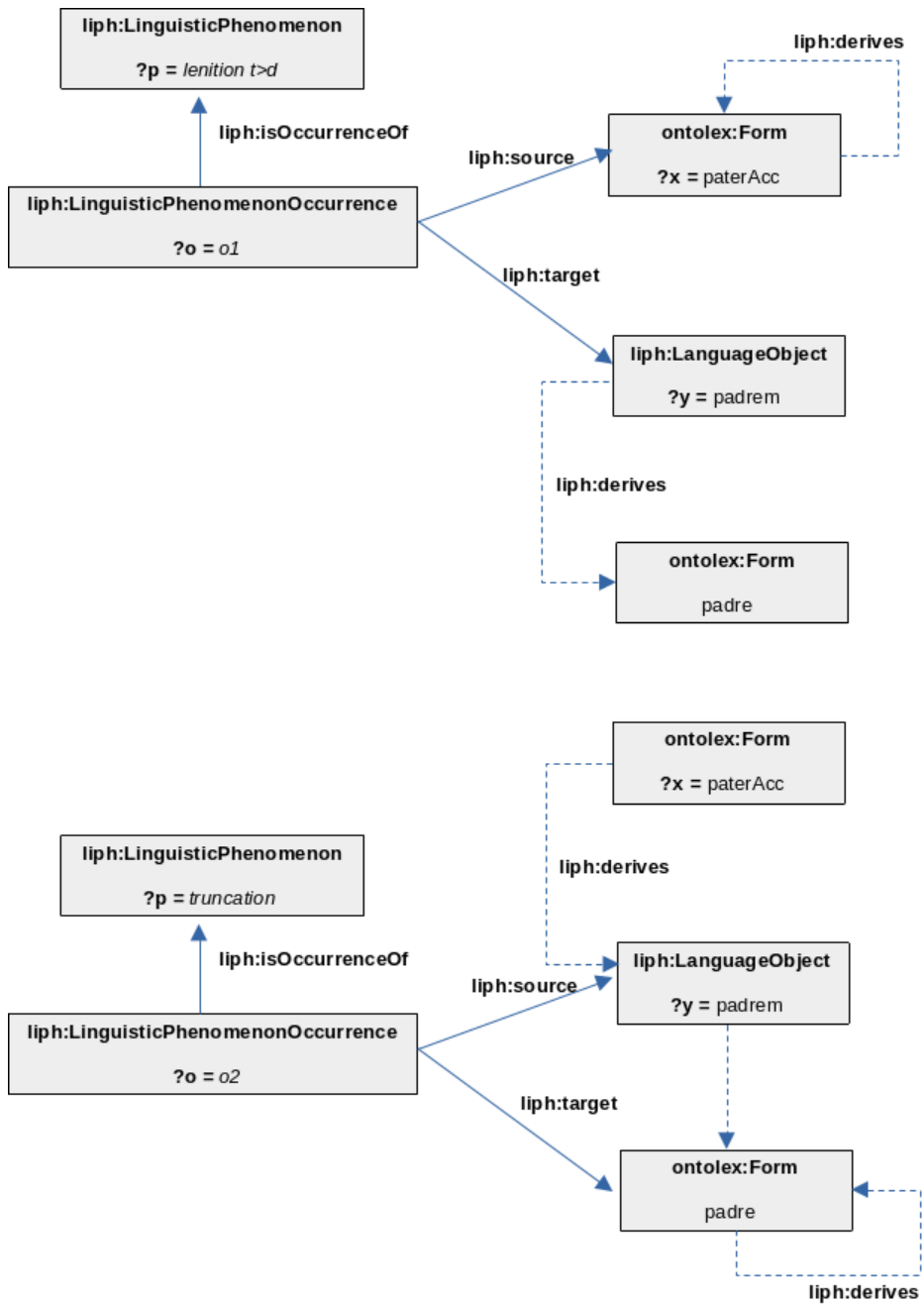


Figure 10. SPARQL query matches

```

...

:oi liph:source :x(i-1);
    liph:target xi;
    liph:occurrenceOf :truncation.
...

:on liph:source :x(n-1);
    liph:target :?l;
    liph:occurrenceOf :pn.

```

The derivation paths from $?e$ to $x(i-1)$ and from xi to $?l$ are condensed by the following two assertions:

```

?e liph:derives :x(i-1).
:xi liph:derives ?l.

```

Thus, the SPARQL query is satisfied by associating $?x = x(i-1)$, $?o = oi$, and $?y = xi$. In addition, the cases in which the considered phenomenon occurs at the beginning or at the end of the derivation are handled by the very same query in light of the reflexivity of *liph:derives*.

4.6. Finite-State Phenomena

Linguistic phenomena studied by researchers can often be described *operationally*. For example, the linguist language feature *truncation* illustrated in Figure 2 can be described in a human-readable way by the following phrase:

remove the final “m”.

However, to foster information reuse, a machine-readable description is more desirable when available. In particular, automatically-verifiable phenomena would enable automated compliance checks of linguistic phenomenon occurrences against their definitions.

Finite-state technologies have been proficiently applied to the modeling of phonological phenomena, as shown in [29,19] and discussed in Section 2.

These technologies are grounded in the notion of *regular relations* (also known as *rational relations*, see [30]), i.e., a generalization of the notion of *regular language* to n -ary relations. For our purposes, it suffices considering binary relations, as our subject of study are just phenomena relating pairs of lexical expressions. Thus, in what follows we will use the term *regular relation* to indicate a regular binary relation. Analogously to regular languages,² which can be recognized by *finite-state automata*, regular relations are those that can be recognized by *finite-state transducers*, i.e., finite-state automata equipped with two tapes, whose transitions are labeled with pairs of symbols rather than single ones.

Our ontology provides the class `liph:FiniteStateLinguisticPhenomenon`, a subclass of `liph:LinguisticPhenomenon`, to represent and provide executable descriptions of linguistic phenomena that correspond to regular relations. Such a finite-state phenomenon has to be specified in terms of

²Here the word “language” is used strictly to denote a set of strings of any alphabet.

Table 4. Finite-state Linguistic Phenomena constraints

<code>liph:FiniteStateLinguisticPhenomenon</code>	\sqsubseteq	<code>liph:LinguisticPhenomenon</code>
<code>liph:matchingPattern</code>	\sqsubseteq	<code>liph:FiniteStateLinguisticPhenomenon</code> \times <code>xsd:string</code> <code>Func(liph:matchingPattern)</code>
<code>liph:replaceWith</code>	\sqsubseteq	<code>liph:FiniteStateLinguisticPhenomenon</code> \times <code>xsd:string</code> <code>Func(liph:replaceWith)</code>
<code>liph:FiniteStateLinguisticPhenomenon</code>	\sqsubseteq	$\exists(liph:matchingPattern).(xsd:string)$
<code>liph:FiniteStateLinguisticPhenomenon</code>	\sqsubseteq	$\exists(liph:replaceWith).(xsd:string)$

- a *regular expression* (see [31]), i.e, a sequence of characters that can *match* some or none of the substrings of the written representation of a lexical expression, and
- a *replacement*, indicating how these parts will be modified.

The regular expression and the replacement characterizing a phenomenon are specified by the two functional data-type properties `liph:matchingPattern` and `liph:replaceWith`, respectively. The regular expression syntax must conform the specification reported in [32]. In contrast, replacements are plain strings, but may contain references to *captured subexpressions*, as described in [33] and reviewed later in the text.

In short, if the regular expression matches either the entire parent expression or just a substring of it, the phenomenon can be *applied* by replacing the matched portion as specified by the `liph:replaceWith` property. For example, the phenomenon *lenition* $t > d$ presented in Section 4.1 could be described by a `liph:FiniteStateLinguisticPhenomenon` individual with “t” and “d” as values for the `liph:matchingPattern` and `liph:replaceWith` properties, respectively.

As anticipated before, regular expressions may contain subexpressions enclosed in parentheses, known as *captured expressions*, which are numbered sequentially, starting from 1. Captured expressions can be used in replacements: every occurrence of $\$n$ in a replacement string will be substituted with the substring matched by the n -th captured expression.

Captured expressions are a well-known feature of finite-state transducers. However, to clarify their usage in our context, we provide an example, in Turtle syntax, related to the phenomenon *truncation* introduced earlier:

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

:truncation a liph:FiniteStateLinguisticPhenomenon;
  liph:matchingPattern "^(.*)m$";
  liph:replaceWith "$1".
```

Here, the regular expression contains a captured expression “(.*)” intended to detect all the characters preceding the final “m”. These captured characters are then used as replacement, so that the resulting string is the original minus the final “m”.

The modelling features introduced in this section are summarized using Description Logic notation in Table 4, provided that `xsd` is the abbreviation for the prefix `http://www.w3.org/2001/XMLSchema#`.

If a finite-state linguistic phenomenon is specified using these features, an automated agent could easily verify whether any of its occurrences is compliant with its definition or not. Let us describe how such an agent should operate by means of some examples.

Let `lux`, `luce`, and `luxAcc` be three instances of `liph:LexicalObject` with written representations “lux”, “luce”, and “lucem”, respectively. It can be automatically determined that the following phenomenon occurrence is not compliant with the definition of the phenomenon *truncation* provided above, as the regular expression associated to *truncation* does not match any substring of “lux”:

```
:o1 a liph:LinguisticPhenomenonOccurrence ;
    liph:occurrenceOf :truncation ;
    liph:source :lux ;      # writtenRep "lux"
    liph:target :luxAcc . # writtenRep "lucem"
```

Instead, the written form of `luxAcc` matches the regular expression characterizing the phenomenon *truncation*. The only string that can be obtained from “lucem” by replacing the `$1` placeholder in the replacement string (which, in this case, constitutes the entire replacement) with the captured string “luce”, matched by the captured expression `(.*)`, is “luce”. Thus, the following statement complies with the definition of *truncation*:

```
:o2 a liph:LinguisticPhenomenonOccurrence;
    liph:occurrenceOf :truncation;
    liph:source :luxAcc ; # writtenRep "lucem"
    liph:target :luce .   # writtenRep "luce"
```

Conversely, any statement relating `luxAcc` to an individual whose written representation differs from “luce” via the phenomenon *truncation* would violate the definition of the phenomenon, as in the following example:

```
:o3 a liph:LinguisticPhenomenonOccurrence;
    liph:occurrenceOf :truncation;
    liph:source :luxAcc ; # writtenRep "lucem"
    liph:target :lux .    # writtenRep "lux"
```

4.7. Comparison with the previous version

As mentioned in Section 1, a previous version of *LiPh* was presented in [5] and is available at the following URL:

<https://gallosiciliani.unict.it/ns/liph/1.2.0>

In this section, we briefly discuss the differences and compatibility matters between the previous version, in short called *LiPh 1*, and the version presented in this paper, which we refer to as *LiPh 2*.

The core property of *LiPh 1* was *liph:linguisticPhenomenon*, intended as a super-property for all the linguistic phenomena. However, the property name is somewhat misleading: as noted in Section 1 and reiterated in Section 4.1, the linguistic phenomena of interest in our context are those that necessarily produce changes in lexical expressions. Hence, they should be irreflexive. On the contrary, *liph:linguisticPhenomenon* has been defined as a transitive and reflexive property just for technical reasons.

In *LiPh 2*, the property *liph:linguisticPhenomenon* has been replaced by the most specific property *liph:derives*, representing chains of linguistic phenomena applications.

In contrast with *liph:linguisticPhenomenon*, the reflexivity of *liph:derives* is appropriate, as it serves for modeling empty chains.

The *liph:LexicalObject* class remains mostly unchanged, but some issues related to the *ontolex:writtenRep* domain constraint stated in *OntoLex-lemon* have been solved.

The chief innovation introduced by *LiPh 2* is the reification of linguistic phenomena occurrences described in Section 4.4. Indeed, linguistic phenomena can be still represented as object properties, so that linguistic phenomena occurrences have to be stated as object property assertions. However, as reported in Section 4.4, the reification-based approach is more general and beneficial.

As a consequence, the specification of finite-state phenomena differs in that, in *LiPh 1*, regular expressions were attached to linguistic phenomena properties through the annotation properties *liph:regex* and *liph:replacement*, respectively, whereas in *LiPh 2* they are attached to the individual representing the phenomenon by means of the two object properties *liph:matchingPattern* and *liph:replaceWith*. This allowed us to provide a more precise and valuable definition of finite-state phenomena thanks to the *liph:LinguisticPhenomenon* class, whose instances have exactly one regular expression and replacement (see Table 2). On the contrary, in *LiPh 1*, *liph:matchingPattern* and *liph:replaceWith* are annotation properties, hence, they are excluded by reasoning tasks.

5. Linguistic phenomena for Gallo-Sicilian varieties

As already mentioned in Section 1, some Gallo-Italic varieties are spoken in Sicily as a consequence of a medieval immigration from Northwestern Italy. These varieties descend from the original Gallo-Italic varieties of Northern Italy, but have been affected by a long-term contact with Sicilian.

We denote by *Gallo-Sicilian varieties* the Gallo-Italic varieties spoken in Sicily, and refer to *Gallo-Sicilian features* as the set of linguistic phenomena that characterize these varieties.

Several features of Gallo-Sicilian varieties, in particular those spoken in Nicosia, Sperlinga, San Fratello, and Novara di Sicilia, have been identified [4,34]. Furthermore, some of these features have been classified as typical of either Northern or Southern Italy. This may be helpful in determining whether a Gallo-Sicilian lexical expression has been inherited from a source Gallo-Italian variety, or it has been borrowed from Sicilian.

Using *LiPh*, a set of linguistic phenomena that occurred in borrowings of Sicilian lexical expressions into Gallo-Sicilian varieties, relevant to the analysis of contact-induced changes, have been collected in the *Gallo-Sicilian Features Ontology*, which will be detailed in Section 5.3. To explain the methodology used to identify these linguistic phenomena, we first provide some background notions from linguistics.

5.1. Linguistic Background

We begin by briefly recalling some linguistic notions necessary to describe the methodology that will be presented in the next section (Section 5.2) for identifying and selecting language features. As will become clear, this methodology is closely tied to phonetics and phonology. As a consequence, most of the concepts and definitions introduced in this section are expressed using the *International Phonetic Alphabet* (in short, *IPA*), as described in [35].

In linguistics, “feature” is a generic term that identifies a recognizable characteristic of a certain linguistic entity (phonemes, words, phrases, etc.) [36].

For instance, inflected words have different grammatical categories or *features* such as gender, number, case, tense, mood, aspect, and so on, each of them having a set of potential values. Nominative, genitive, accusative and singular, plural, dual are some of the available values for the case and number features, respectively (see [37] for an exhaustive discussion about grammatical features).

Similarly, at the phonological level, phonemes are distinguished from each other and grouped into sound categories on the basis of their *phonetic features*. To give an example, in Italian and English we can distinguish /p/ and /b/ through the *voice* feature, which is specified by the values ‘voiced’ (or, in binary terms, [+voice]) and ‘unvoiced’ ([-voice]): both segments are *bilabial plosives consonants*, but while /p/ has the value [-voice], /b/ is described as [+voice]. Since, in this case, the voice feature is relevant to the phonemic definition of a phonic segment – namely, a segment that serves to differentiate meanings in a particular language and is therefore opposed to another phoneme in the same syntagmatic position (as in the Italian minimal pair [*pasta*(*pasta*);*basta*(*enough!*)] – we can say that this feature is a *distinctive feature* (see [38,39,40]).

However, it must be said that not all phonetic features are phonologically relevant. For example, English has two types of voiceless plosives: [p^h t^h k^h] [+aspirate] and [p t k] [-aspirate]. The former can only occur in initial word position (as in *pin* [p^hm]), whereas the latter is found after a word-initial [s] (as in *spin* [spɪn]). The two different [p] appear in different contexts – for this reason, [p^h] is an *allophone*, i.e., a contextual variant of the phoneme /p/ – and, consequently, the aspiration feature can be predicted depending on the context of occurrence of the sound. Hence, aspiration, in contrast to voice, does not serve to differentiate meanings in English and is considered a *non-relevant* or *non-distinctive feature*, that is, a feature relevant only at the phonetic level (see [38,39]).

The difference between distinctive and non-distinctive features, as well as the difference between phonemes and phones allows us to introduce another crucial concept, that of *phonological rule*. Since the same phoneme could correspond to several different phonetic realizations (as in the case of the English /p/, which corresponds to both [p^h] and [p]), according to generative phonology (see [41]), there exist processes transforming phonological representations into phonetic representations, e.g., to determine the form of a phoneme in a certain phonotactic context. These are precisely the phonological rules or processes.

For example, in Italian there is a phonological rule of assimilation, described in [42], whereby /n/ → [ŋ] before a velar consonant, as in *anca* [aŋka] (hip) and *finco* [fiŋgo] (I pretend).

Similarly, in the transition from Latin to Romance languages, the clusters -mb-, -nd- have become -[mm]-, -[nn]-, respectively, in many Southern Italo-Romance dialects. For example, the Latin *gamba* (leg) has become [jamm] in the Apulian dialect of Altamura [43]. There was, thus, a diachronic rule, whereby /b/ and /d/ after the nasal consonants /n/ and /m/ have totally assimilated to the preceding consonant (this rule, in fact, is known as *total progressive assimilation*, see [44]).

The only distinction between the two rules just described is that the former is still *active* in Italian, while the latter is no longer active in most dialects in which it is documented. This means that if new words containing the clusters /mb/ and /nd/ enter in these varieties, these clusters will remain unchanged.

5.2. Gallo-Sicilian Features for Contact-Induced Changes

The Gallo-Sicilian Features Ontology presented here has been developed as part of a broader effort to study contact-induced changes in Gallo-Sicilian varieties resulting from long-term contact with Sicilian. At this stage, the study focuses on the written representations of canonical forms (i.e., lemmas) of both nouns and verbs, reported in [45], which were borrowed from Sicilian, along with their corresponding Sicilian etymons, retrieved from [46,47,48,49,50].

For these reasons, giving due consideration to the fact that ontology considers only lemmas, not entire dictionary entries, we first selected phonological features that can highlight phonological changes. In addition, as Gallo-Sicilian varieties have proved to be very conservative at a phonetic-phonological level (see [4]), we did not test (at least for now) the presence of Sicilian phonological traits into Gallo-Sicilian, but the preservation of Gallo-Italic phonological traits, despite centuries of contact with Sicilian. These traits are ultimately what we call *Gallo-Sicilian features*.

Specifically, we did not simply select features found in the inherited Gallo-Italic lexicon (i.e., the results of diachronic phonological processes that occurred in Northern Italo-Romance varieties), since this would have only phonetic relevance, showing a lexicalized phenomenon – namely, one that is not synchronically active. Instead, we focus on those Gallo-Italic features found in lexical borrowings from Sicilian (e.g., *nic. viadörö* ‘pack-leader’ ; *sic. abbiaturi* ‘id.’), where we observe two clearly Gallo-Italic traits: the lenition of /t/ to /d/ and the preservation of a Northern Italo-Romance vowel system, as shown by the development of tonic /u/ into /o/. In fact, only these features are phonologically relevant, since they allow us to demonstrate the presence of active phonological processes – no matter whether still active today or only at one stage in the linguistic history of Gallo-Sicilian – as well as the existence a phonological inventory different from that of Sicilian. In this way, ultimately, it is possible to see how Gallo-Sicilian phonological systems operate synchronically in contact with those of Sicilian.

5.3. Ontology Structure

All the items in the ontology belong to the following namespace:

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

The Gallo-Sicilian Features Ontology chiefly consists of individuals representing Gallo-Sicilian features. We characterize these features as members of the novel subclass of `liph:LinguisticPhenomenon` `gs:GalloSicilianFeature`. This would foster the reuse of these individuals in scenarios involving also other feature sets.

Each of these phenomena is equipped with a finite-state executable description, as detailed in Section 4.6. In this way, derivation chains involving these phenomena can be automatically generated, and phenomena occurrences compliance can be automatically verified.

Gallo-Sicilian features are grouped into ten families, encoded by the subclasses of `liph:LinguisticPhenomenon` enumerated below:

- `gs:Leniz` for *lenition*, i.e., the softening of the intervocalic voiceless (both fricative and plosive) consonants;³

³It can also be considered as “an increase in the vocalic nature of a segmen” (cf. [51]).

- gs:Degem for *degemination*, i.e., the reduction of all long (or double) consonants into short (or single) ones;
- gs:Assib for *assibilantion*, i.e., the process by which non-sibilant consonants (typically plosives or affricates) change into sibilant sounds;
- gs:Dissim for *dissimilation*, i.e., the process by which “two adjacent segments that share some features change to become less like each other” (cf. [51]);
- gs:Ditt for *diphthongization*, i.e., the turning of a stressed vowel into a diphthong, which is “a (functionally) single vowel that starts out in the position of one monophthong and ends up in the position of another” (cf. [51]);
- gs:Vocal for the adaptation of a loanword to the vocalic system of the receiving language;
- gs:Afer for *apheresis*, i.e., the loss or deletion of one or more sounds at the beginning of a word;
- gs:Palat for *vocalic palatalization*, i.e., the phonological process by which the low, central vowel /a/ shifts towards a more fronted and sometimes slightly raised position, typically resulting in a vowel like /æ/, /ɛ/, or /e/, in a palatal or fronting phonetic environment;
- gs:Elim for the *elimination* of some allophonic variants according to a specific phonological context, such as the unstressed position;
- gs:Deretr for *deretroflexion*, i.e., the elimination of the retroflexed pronunciation of a consonant.

All the aforementioned phenomena families are, to different degrees, well-known families of linguistic phenomena, so that defining the corresponding classes as subclasses of gs:GalloSicilianFeature would be inappropriate. In addition, notice that, for the purposes of the present study, the family labeled as gs:Leniz also includes the nasalization of the vowels in final position, which is another well documented phenomenon among different languages of the world.

This arrangement is an example of how linguistic phenomena can be organized into hierarchies to provide a fine-grained classification. Other well-known vocabularies such as, for example, SKOS (see [52]) could be employed to organize phenomena into complex classification systems. However, we preferred the canonical approach based on class hierarchy, which brought with it all the consequences in terms of reasoning and inferred information provided by the OWL model-theoretic semantics.

As an example of Gallo-Sicilian feature encoding, we report a turtle fragment with the definition of the phenomenon changing “tt” to “t”.

```
gs:degem.2 a gs:GalloSicilianFeature ,
    gs:Degem ;
    rdfs:label "degem.2" ;
    rdfs:comment "tt > t";
    liph:matchingPattern "tt" ;
    liph:replaceWith "t" .
```

In this fragment, the phenomenon is represented by an individual gs:degem.2 belonging to both gs:GalloSicilianFeature, characterizing all the Gallo-Sicilian features, and gs:Degem, representing the phenomena family of degeminations. The regular expression, associated with the phenomena by means of *liph:matchingPattern*, restricts this

feature to all “tt”. The value provided for *liph:replaceWith* indicates that each instance of “tt” should be replaced with a single “t”. Notice that, despite it is not stated explicitly, *gs:degem.2* belongs to the class *liph:FiniteStateLinguisticPhenomenon* as well, in force of the constraints in Table 4.

Based on the constraints in Table 1 and 2, and provided that some language expressions and their derivations are encoded using the Linguistic Phenomena Ontology described in Section 4 alongside the Gallo-Sicilian Features Ontology presented here, one can retrieve, for example, all the expressions exhibiting features belonging to one of the families enumerated above, assuming reasoning functionalities are available. In this case, all the individuals corresponding to lexical expressions with features of a specific family, say *gs:Leniz*, can be retrieved by the following SPARQL query:

```
PREFIX liph: <https://gallosiciliani.unict.it/ns/liph#>
PREFIX gs: <https://gallosiciliani.unict.it/ns/gs-features#>

SELECT ?expression WHERE {
  ?o liph:target ?y;
    liph:occurrenceOf ?p .
  ?p a gs:Leniz .
  ?y liph:derives ?expression
}
```

This query closely resembles the one presented in Section 4.5 concerning [CQ2], with the only difference that it calls for all the phenomena in a specific family instead of requesting a specific one.

6. Conclusions and Future Works

In this paper, we presented the *Linguistic Phenomena Ontology*, an OWL ontology based on *OntoLex-Iemon*, designed to represent changes in the morphology and pronunciation of lexical expressions. We further extended it to model the linguistic phenomena characterizing the Gallo-Italic varieties spoken in Sicily, with the aim of studying the etymology of expressions borrowed from Sicilian.

Some aspects of the OWL encoding of linguistic phenomena proposed here require further investigation.

First, as discussed in Section 4.1, we restricted our attention to phenomena that alter the written representation of a lexical expression. However, changes in the written form necessarily affect pronunciation, thereby altering the phonetic form as well. When not explicitly provided, the phonetic representation of a lexical expression in any well-known language can often be derived from its written form using language-specific grapheme-to-phoneme conversion maps, when available. Examples include resources for English and French as described in [53], or the conversion rules for the Gallo-Italic variety spoken in Nicosia and Sperlinga, summarized in [45].

Despite ongoing research in automated grapheme-to-phoneme conversion (see, e.g., [53,54,55,56,57,58]), to the best of our knowledge, no attempt has been made so far to encode such conversion systems in OWL.

Now consider a derivation (as defined in Section 4.1) from a source lexical expression in one language to a target expression in another. Even when grapheme-to-phoneme conversion rules exist for both the source and recipient languages, the question remains: how should we handle the pronunciation of intermediate forms within the derivation? This becomes particularly challenging if the two languages assign different phonetic values to the same graphemes.

In addition, the application of our ontology to *Computational Morphology* (as mentioned in Section 2) appears promising and deserves further exploration.

Finally, future work may also include extending the ontology to cover semantic phenomena, such as *sense-shift*, which represent an important dimension.

Acknowledgements

This work has been created in the scope of the project PRIN 2022 PNRR “Contact-induced change and sociolinguistics: an experimental study on the Gallo-Italic dialects of Sicily”, funded by the European Union – Next Generation EU, Mission 4, Component 1, CUP J53D23017360001 - ID P2022YWS8T; Research Unit of the University of Catania.

References

- [1] Thomason SG, Kaufman T. *Language Contact, Creolization, and Genetic Linguistics*. University of California Press; 1988. Available from: <http://dx.doi.org/10.1525/9780520912793>.
- [2] Marotta G. II. In: *Phonetics and Phonology*. Cambridge Handbooks in Language and Linguistics. Cambridge University Press; 2022. p. 200.
- [3] van Coetsem F. *A General and Unified Theory of the Transmission Process in Language Contact*. Winter; 2000.
- [4] De Angelis A. The Strange Case of the Gallo-Italic Dialects of Sicily: Preservation and Innovation in Contact-Induced Change. *Languages*. 2023;8(3). Available from: <https://www.mdpi.com/2226-471X/8/3/163>.
- [5] Cantone D, Di Caro VN, Longo C, Menza S, Nicolosi Asmundo M, Santamaria DF. An OWL Ontology for Linguistic Phenomena with Applications to Gallo-Italic Dialects in Sicily. In: Bikakis A, Ferrario R, Jean S, Markhoff B, Mosca A, Nicolosi Asmundo M, editors. *Proceedings of the fourth edition of the International Workshop on Semantic Web and Ontology Design for Cultural Heritage, Tours, France, October 30-31, 2024*. vol. 3809 of *CEUR Workshop Proceedings*. CEUR-WS.org; 2024. p. 12-24. Available from: <https://ceur-ws.org/Vol-3809/paper2.pdf>.
- [6] Fellbaum C, editor. *WordNet: an electronic lexical database*. Massachusetts: The MIT Press; 1998. P.423.
- [7] Farrar S, Langendoen T. A Linguistic Ontology for the Semantic Web. *GLOT International*. 2003;7.
- [8] Pedersen B, McCrae J, Tiberius C, Krek S. ELEXIS - a European infrastructure fostering cooperation and information exchange among lexicographical research communities. In: Bond F, Vossen P, Fellbaum C, editors. *Proceedings of the 9th Global Wordnet Conference*. Nanyang Technological University (NTU), Singapore: Global Wordnet Association; 2018. p. 335-40. Available from: <https://aclanthology.org/2018.gwc-1.40>.
- [9] Tasovac T, Romary L, Banski P, Bowers J, de Does J, Depuydt K, et al. TEI Lex-0: A baseline encoding for lexicographic data. *DARIAH Working Group on Lexical Resources*; 2018. Available from: <https://dariah-eric.github.io/lexicalresources/pages/TEILex0/TEILex0.html>.
- [10] Cimiano P, McCrae JP, Buitelaar P. *Lexicon Model for Ontologies: Community Report*. W3C; 2016. Available from: <https://www.w3.org/2016/05/ontolex/>.
- [11] Abgaz YM. Using OntoLex-Lemon for Representing and Interlinking Lexicographic Collections of Bavarian Dialects. In: Ionov M, McCrae JP, Chiarcos C, Declerck T, Bosque-Gil J, Gracia J, editors.

- LDL@LREC. European Language Resources Association; 2020. p. 61-9. Available from: <http://dblp.uni-trier.de/db/conf/acl-ldl/acl-ldl2020.html#Abgaz20>.
- [12] Chiarcos C, Fäth C, Ionov M. The ACoLi Dictionary Graph. In: Proceedings of The 12th Language Resources and Evaluation Conference; 2020. p. 3281-90.
- [13] Racioppa S, Declerck T. Porting the Latin WordNet onto OntoLex-Lemon. In: Kosem I, Cukr M, Jakubiček M, Kallas J, Krek S, Tiberius C, editors. Electronic lexicography in the 21st century. Proceedings of the eLex 2021 conference; 2021. p. 429-39.
- [14] Khan AF, Chiarcos C, Declerck T, Gifu D, González-Blanco García E, Gracia J, et al. When linguistics meets web technologies. Recent advances in modelling linguistic linked data. *Semantic Web*. 2022 Sep;13(6). Available from: <https://doi.org/10.5281/zenodo.7129494>.
- [15] Lindemann D, Ahmadi S, Khan AF, Mambrini F, Iurescia F, Passarotti MC. When OntoLex Meets Wikibase: Remodeling Use Cases. In: Kaffee LA, Razniewski S, Alghamdi K, Arnaout H, editors. *Wikidata@ISWC*. vol. 3640 of CEUR Workshop Proceedings. CEUR-WS.org; 2023. p. 14. Available from: <http://dblp.uni-trier.de/db/conf/wikidata/wikidata2023.html#LindemannAKMIP23>.
- [16] Khan AF. Towards the Representation of Etymological Data on the Semantic Web. *Information*. 2018;9(12). Available from: <https://www.mdpi.com/2078-2489/9/12/304>.
- [17] Chiarcos C, Gkirtzou K, Khan AF, Labropoulou P, Passarotti M, Pellegrini M. Computational Morphology with OntoLex-Morph. In: Declerck T, McCrae JP, Montiel-Ponsoda E, Chiarcos C, Ionov M, editors. LDL@LREC. European Language Resources Association; 2022. p. 78-86. Available from: <http://dblp.uni-trier.de/db/conf/acl-ldl/acl-ldl2022.html#ChiarcosGKLP22>.
- [18] Hurskainen A, Koskenniemi K, Pirinen T, Antonsen L, Axelson E, Bick E, et al. Rule-Based Language Technology. *Northern European Association for Language Technology*; 2023.
- [19] Kaplan RM, Kay M. Regular Models of Phonological Rule Systems. *Computational Linguistics*. 1994;20(3):331-78. Available from: <https://aclanthology.org/J94-3001>.
- [20] Beckett D, Berners-Lee T, Carothers G, Prud'hommeaux E. RDF 1.1 Turtle. W3C; 2014. Available from: <http://www.w3.org/TR/2014/REC-turtle-20140225/>.
- [21] Hitzler P, Krötzsch M, Parsia B, Patel-Schneider PF, Rudolph S. OWL 2 Web Ontology Language Primer. World Wide Web Consortium; 2009. Available from: <http://www.w3.org/TR/owl2-primer/>.
- [22] Buitelaar P, Cimiano P, Haase P, Sintek M. Towards Linguistically Grounded Ontologies. In: 6th Annual European Semantic Web Conference (ESWC2009); 2009. p. 111-25. Available from: <http://www.cimiano.de/Publications/2009/eswc09/eswc09.pdf>.
- [23] Krieger HU. A Detailed Comparison of Seven Approaches for the Annotation of Time-Dependent Factual Knowledge in RDF and OWL. In: Proceedings of the 10th Joint ACL-ISO Workshop on Interoperable Semantic Annotation (held in conjunction with LREC 2014), European Language Resources Association. Northern European Association for Language Technology (NEALT); 2014. p. 1-8.
- [24] World Wide Web Consortium. Prud'hommeaux E, Harris S, Seaborne A, editors. SPARQL 1.1 Query Language. W3C; 2013. Available from: <http://www.w3.org/TR/sparql11-query>.
- [25] Baader F, Calvanese D, McGuinness D, Nardi D, Patel-Schneider P. *The Description Logic Handbook: Theory, Implementation and Applications*. Cambridge University Press; 2003.
- [26] Klyne G, Carroll JJ. *Resource Description Framework (RDF): Concepts and Abstract Syntax*. W3C; 2004. Available from: <http://www.w3.org/TR/rdf-concepts/>.
- [27] Motik B. On the Properties of Metamodeling in OWL. *Journal of Logic and Computation*. 2007;17(4):617-37.
- [28] Horrocks I, Patel-Schneider PF. Reducing OWL Entailment to Description Logic Satisfiability. *Journal of Web Semantics*. 2004;1(4):345-57.
- [29] Karttunen L. Finite-State Constraints. In: Goldsmith J, editor. *The Last Phonological Rule*. University of Chicago Press; 1993. p. 173-94.
- [30] Eilenberg S. *Automata, Languages, and Machines*. USA: Academic Press, Inc.; 1976.
- [31] Kleene SC. Representation of Events in Nerve Nets and Finite Automata. In: Shannon CE, McCarthy J, editors. *Automata Studies*. Princeton University Press; 1956. p. 3-41.
- [32] Biron PV, Malhotra A. XML Schema Part 2: Datatypes Second Edition. W3C; 2024. [Http://www.w3.org/TR/2004/REC-xml-20040204](http://www.w3.org/TR/2004/REC-xml-20040204).
- [33] Kay M. XPath and XQuery Functions and Operators 3.1. W3C; 2017. [Http://www.w3.org/TR/2004/REC-xml-20040204](http://www.w3.org/TR/2004/REC-xml-20040204).
- [34] Trovato SC. Galloitalische Sprachkolonien. I dialetti galloitalici della Sicilia. *Kontakt, Migration Und*

Kunstsprachen. 1998 jan;7:538–559.

- [35] Association IP. Handbook of the International Phonetic Association. Cambridge, New York, Madrid: Cambridge University Press; 1999.
- [36] Corbett GG. Features. Cambridge Textbooks in Linguistics. Cambridge University Press; 2012.
- [37] Kibort A, Kibort A, Corbett G, editors. A typology of grammatical features.. Surrey Morphology Group; 2008. Available from: <http://www.grammaticalfeatures.net/inventory.html>.
- [38] Trubeckoj NS. Fondamenti di fonologia. Giulio Einaudi Editore; 1971.
- [39] HYMAN LM. Fonologia : Teoria e analisi. Il Mulino; 1981.
- [40] Lopporcaro M. In: Beccaria GL, editor. Tratto distintivo. Einaudi; 2004. p. 777-8.
- [41] Chomsky N, Halle M. The sound pattern of English. Studies in language. New York: Harper and Row; 1968.
- [42] Renzi L, Andreose A. Manuale di linguistica e filologia romanza. Manuali . Filologia e critica letteraria. Bologna: Il Mulino; 2009.
- [43] Lopporcaro M. Grammatica storica del dialetto di Altamura. Giardini; 1988.
- [44] Graffi G, Scalise S. Le lingue e il linguaggio. Introduzione alla linguistica. Il Mulino; 2002.
- [45] Trovato SC, Menza S. Vocabolario del dialetto galloitalico di Nicosia e Sperlinga. No. 39 in *Materiali e ricerche dell'Atlante Linguistico della Sicilia*. Centro di studi filologici e linguistici siciliani; 2020.
- [46] Piccitto G, Tropea G, Trovato SC. VOCABOLARIO SICILIANO I (A-E). No. 1 in *Vocabolario Siciliano*. Centro di studi filologici e linguistici siciliani; 1977.
- [47] Piccitto G, Tropea G, Trovato SC. VOCABOLARIO SICILIANO II (F-M). No. 2 in *Vocabolario Siciliano*. Centro di studi filologici e linguistici siciliani; 1985.
- [48] Piccitto G, Tropea G, Trovato SC. VOCABOLARIO SICILIANO III (N-Q). No. 3 in *Vocabolario Siciliano*. Centro di studi filologici e linguistici siciliani; 1990.
- [49] Piccitto G, Tropea G, Trovato SC. VOCABOLARIO SICILIANO IV (R-Sg). No. 4 in *Vocabolario Siciliano*. Centro di studi filologici e linguistici siciliani; 1997.
- [50] Piccitto G, Tropea G, Trovato SC. VOCABOLARIO SICILIANO V (Si-Z). No. 5 in *Vocabolario Siciliano*. Centro di studi filologici e linguistici siciliani; 2002.
- [51] Davenport M, Hannahs SJ. *Introducing Phonetics and Phonology* (3rd edition). London: Hodder Education; 2010.
- [52] Isaac A, Summers E. Isaac A, Summers E, editors. SKOS Simple Knowledge Organization System Primer. W3C; 2008. World Wide Web Consortium, Working Draft WD-skos-primer-20080829.
- [53] Divay M, Vitale AJ. Algorithms for Grapheme-Phoneme Translation for English and French: Applications for Database Searches and Speech Synthesis. *Comput Linguistics*. 1997;23:495-523. Available from: <https://api.semanticscholar.org/CorpusID:61159418>.
- [54] Pagel V, Lenzo KA, Black AW. Letter to Sound Rules for Accented Lexicon Compression. *CoRR*. 1998;cmp-1g/9808010. Available from: <http://arxiv.org/abs/cmp-1g/9808010>.
- [55] Yolchuyeva S, Németh G, Gyires-Tóth B. Transformer Based Grapheme-to-Phoneme Conversion. In: *Interspeech 2019*. Interspeech. ISCA; 2019. p. 2095–2099. Available from: <http://dx.doi.org/10.21437/Interspeech.2019-1954>.
- [56] Cheng S, Zhu P, Liu J, Wang Z. A Survey of Grapheme-to-Phoneme Conversion Methods. *Applied Sciences*. 2024;14(24). Available from: <https://www.mdpi.com/2076-3417/14/24/11790>.
- [57] van den Bosch A, Weijters T, Daelemans W. Modularity in Inductively-Learned Word Pronunciation Systems. In: *New Methods in Language Processing and Computational Natural Language Learning*; 1998. p. 185-94. Available from: <https://aclanthology.org/w98-1223/>.
- [58] Lee JL, Ashby LFE, Garza ME, Lee-Sikka Y, Miller S, Wong A, et al. Massively Multilingual Pronunciation Modeling with WikiPron. In: Calzolari N, Béchet F, Blache P, Choukri K, Cieri C, Declerck T, et al., editors. *Proceedings of the Twelfth Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association; 2020. p. 4223-8. Available from: <https://aclanthology.org/2020.lrec-1.521/>.