An OWL Ontology for Linguistic Phenomena with Applications to Gallo-Italic Dialects in Sicily

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

Each lexical expression in a language has its own history: elements of a language are the result of several linguistic processes and may descend from expressions of other languages. In this paper, we propose an ontology to represent *linguistic phenomena*, in particular those that have changed over time the pronunciation or morphology of lexical elements. Then, this ontology is applied to the study of some Gallo-Italic varieties spoken in Sicily. In more details, the linguistic phenomena that have generated the lexical expressions in these languages from Latin etymons have been encoded with our ontology. In addition, an operational *actionable* description, based on regular expressions, has been provided for a relevant amount of these phenomena.

Keywords

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

1. Introduction

Languages are continuously evolving over time. Usually, elements of a language are the result of linguistic processes that fall in the following two categories: *inheritance* and *borrowing*. That is, lexical expressions in a language could have been *inherited* from a parent language, as is the case with many lexemes in the so-called Romance languages, all of which derive from Latin. Alternatilvely, a recipient language can accept borrowings from a *source* language (see [1] for an exhaustive 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 one; 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]), as occurred for the Latin "patrem" that has become "padre" in Italian, is typical of the Italian language.

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

In addition, we say that a linguistic phenomenon is a *feature* of a specific language if it occurred in the generation of a relevant amount of elements of this language.

SWODCH'24: International Workshop on Semantic Web and Ontology Design for Cultural Heritage, October 30–31, 2024, Tours, France

domenico.cantone@unict.it (D. Cantone); vincenzo.dicaro@unict.it (V. Di Caro); cristianolongo@opendatahacklab.org (C. Longo); salvatore.menza@unict.it (S. Menza); marianna.nicolosiasmundo@unict.it (M. Nicolosi Asmundo); daniele.santamaria@unict.it (D. F. Santamaria)

Eliciting linguistic phenomena and language features is worthwhile for historical-linguistics, in particular for studying *contact-induced changes*, i.e., changes 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 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. Indeed, as reported in [4], due to the long-term contact with the Sicilian language, these dialects can represent an excellent testing ground for the language contact theory.

The rest of the paper is structured as follows. In Section 2, we provide a precise definition to designate our intended meaning for linguistic phenomena. From this definition, our ontology for linguistic phenomena is presented and explained in Section 3. Section 4 describes the use-case for the ontology concerning Gallo-Italic varieties. In Section 5, we discuss related works that may suggest additional application fields for our ontology. Finally, in Section 6, we conclude with final observations and directions for future work.

2. Definitions

As mentioned earlier, a linguistic phenomenon occurs when an expression from a *parent* language is transformed into one in a *target* language. At first glance, these phenomena sound like just relations between lexical expressions. For example, one could define a relation *truncation* which drops the final "m" from Latin expressions. Then, as shown in Figure 1, such *truncation* may be used to relate the Latin expression "lucem" with the Italian one "luce".

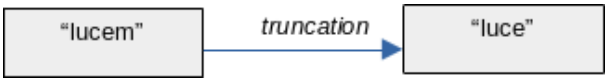

Figure 1: Application of *truncation*, deriving the Italian "luce" from the Latin "lucem"

However, several changes may have occurred in the linguistic process that generated the target expression. Let us consider, for example, the Latin expression "patrem", which has turned into the Italian "padre". The latter may be derived from the former by means of two different phenomena: a relation representing changes of "t" to "d", which we name *lenition* $t > d$, and a *truncation*, as previously defined. In addition, "padre" can be derived from "patrem" by means of these two phenomena in two different ways, as shown in Figure 2.

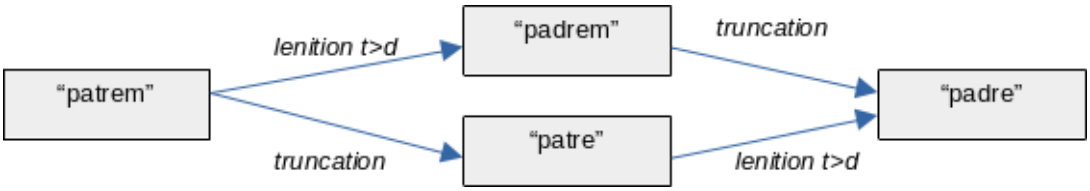

Figure 2: Two possible derivations of "padre" from "patrem"

Notice that intermediate expressions, such as "padrem" and "patre" of the example, are not necessarily lexical expressions of any language. In light of this, linguistic phenomena should be defined more generically as **relations between strings**, where *strings* are understood, as usual, to mean finite sequences of characters from any alphabet.

Following this definition, both *lenition* $t > d$ and *truncation* in Figure 2 are linguistic phenomena. In addition, also their compositions *lenition* $t > d \circ$ *truncation* and *truncation* $\circ$ *lenition* $t > d$ are themselves linguistic phenomena. This shows that *the composition of linguistic phenomena is a linguistic phenomenon as well.*

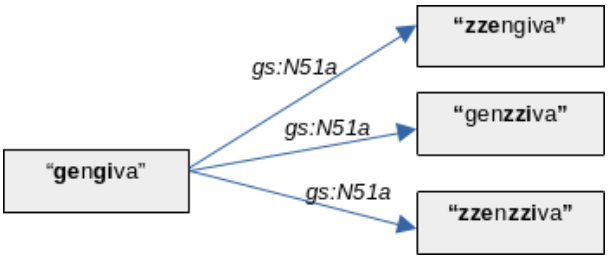

Figure 3: Multiple occurrences of a language feature

It is worth pointing out that linguistic phenomena are defined as relations rather than as functions as they may affect different parts of the same expression, and they may apply to all of these rather than to just some of them. For example, let us consider the linguistic phenomenon typical of languages of northern Italy, where every occurrence of 'g' followed by 'e' or 'i' is changed to 'zz', referred to as `gs:N51a` (see Section 4). It applies to two different parts of the Latin expression "**gengi**va". Thus, the values associated to "gengiva" by `gs:N51a` are "**zze**ngiva", "gen**zzi**va", and "**zze**n**zzi**va", as shown in Figure 3.

Once the notion of linguistic phenomenon is established, the next section provides an OWL [5] encoding for these types of phenomena.

3. The Linguistic Phenomena Ontology

Digital encoding for lexical information has been a thriving research topic in the last decades (see, for example, [6, 7, 8, 9]). In these years, *OntoLex-lemon*, introduced in [10], has become a well-established standard for representing lexical information (see, for example, [11, 12, 13, 14, 15]).

In this section, we present the *Linguistic Phenomena Ontology* (in short, *LiPh*), an OntoLex-lemon extension devoted to representing linguistic phenomena occurred during inheritance and borrowing of lexical expressions. To this end, we first need to recall some core features of OntoLex-lemon.

In the rest of the paper, we will use the Turtle syntax [16] for namespaces and corresponding prefix abbreviations.

3.1. OntoLex-lemon

OntoLex-lemon is an OWL [5] vocabulary that provides rich linguistic grounding for ontologies, 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`, which must have (at least one) written representation, specified by the `ontolex:writtenRep` datatype property, and may have one or more phonetic representations, indicated through the `ontolex:phoneticRep` datatype property. In addition, an `ontolex:LexicalEntry` instance can be associated to other grammatical form variants (plural form, feminine form, etc.) and again represented by `ontolex:Form`, by means of the property `ontolex:otherForm`.

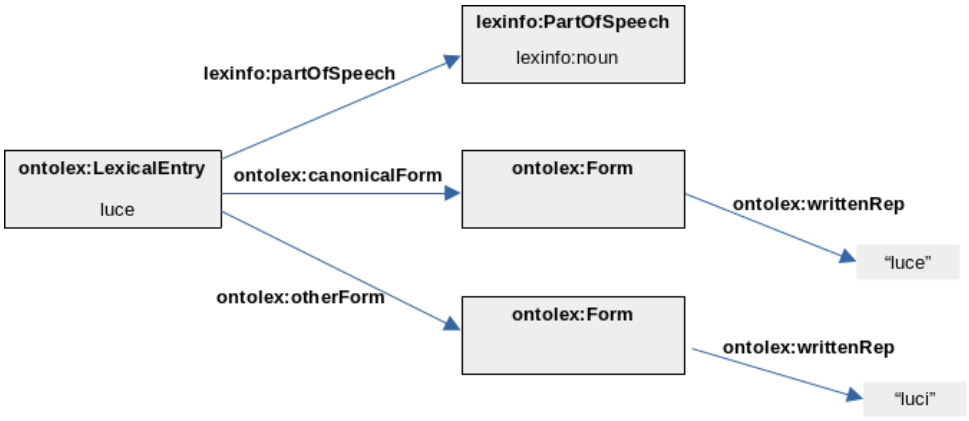

Figure 4: OntoLex-lemon representation for the Italian lexeme "luce"

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 Figure 4, we present a representation of the Italian lexeme "luce" in OntoLex-Lemon, using the *LexInfo* category system introduced in [17], which includes the lemma and its plural form.

Now that we have reviewed the core features of OntoLex-lemon, let us introduce our ontology in the next section.

3.2. Linguistic Phenomena

The Linguistic Phenomena Ontology provides representational features to describe linguistic phenomena (in the sense of Section 2). All ontology entities are in the following namespace:

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

The core property of the ontology is `liph:linguisticPhenomenon`, intended as a super-property for all linguistic phenomena.

Following Section 2, this property must be suitable for connecting instances of `ontolex:Form`. For example, `ontolex:Form` instances representing the Italian expression "luce"' and the Latin expression "lucem" are illustrated in Figure 1. However, we remark that the class `ontolex:Form` represents grammatical realizations of lexical entries, whereas linguistic phenomena may involve also strings that are not expressions of any language such as, for example, "padrem" and "patre" in Figure 2. To cope with this, a more generic class `liph:LexicalObject`, encompassing `ontolex:Form` and preserving the constraints stated in OntoLex-lemon about it, has been created, so that `liph:linguisticPhenomenon` is defined as an object property having `liph:LexicalObject` as domain and range. In other words, `liph:LexicalObject` represents expressions involved in linguistic phenomena, regardless of whether they are part of any language–such as the endpoints of a lexical derivation–or not, as is the case with certain *intermediate forms*.

Porting the existential restriction on `ontolex:writtenRep` (stated in OntoLex-lemon) from `ontolex:Form` to `liph:LexicalObject` ensures that `liph:linguisticPhenomenon`, and its sub-properties, relate only individuals having at least one representation.

We remark that, following Section 2, linguistic phenomena are relations over strings. Given a linguistic phenomenon ρ, let $p_\rho \sqsubseteq$ `liph:linguisticPhenomenon` be an object property dedicated to representing it. For two instances i and j of `liph:LexicalObject`, we say that i is related to j through p_ρ if and only if there exists representations r_i and r_j of i and j, respectively, such that $\rho(r_i) = r_j$.

The definition of domain and range of `liph:linguisticPhenomenon` clarifies that this property concerns forms and cannot be used for phenomena that occur at the level of lexical entries (in the sense

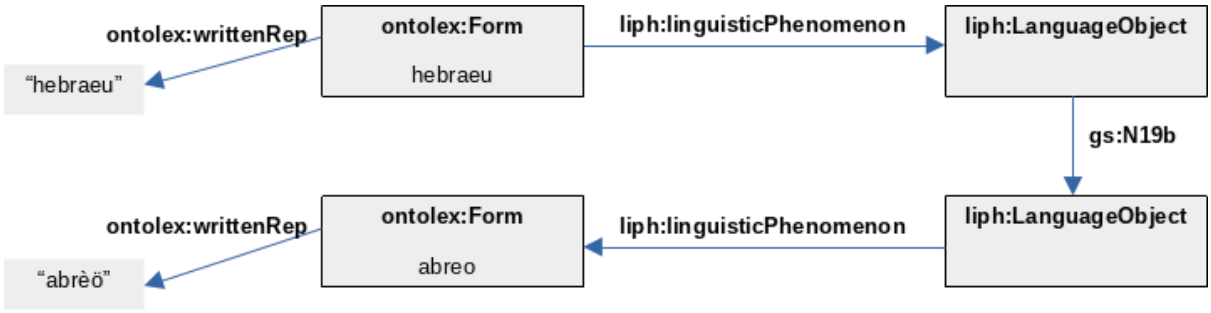

Figure 5: Partial representation of the derivation of "abrèö" from the Latin "hebraeu"

of OntoLex-lemon) such as, for example, sense-shift.

Furthermore, `liph:linguisticPhenomenon` has been declared as transitive in order to include chains of linguistic phenomena, because, as shown in Section 2, linguistic phenomena obtained by composition of other linguistic phenomena are linguistic phenomena as well.

Finally, also reflexivity is imposed on this property, i.e., `liph:linguisticPhenomenon` incorporates the identity relation. Although it is not strictly necessary, it is not inappropriate either as it does not contradict any of the definitions provided so far. Instead, it is helpful for partial modelling of derivations, which can be expressed by chaining specific linguistic phenomena with the general `liph:linguisticPhenomenon` property, used as a place holder for the unspecified portions. Let us consider, for example, the lexeme of the variety spoken in Nicosia "abrèö" (see Section 4), descending from the Latin "hebraeu". Here, the linguistic phenomenon `gs:N19b`, which involves replacing the last part of a lexical expression with "ö", is illustrated by the alteration of the last vowel in "hebraeu". However, it is certain that other phenomena contributed to this derivation, some of which may be unknown. This setup can be represented, as shown in Figure 5, by using `gs:N19b` in conjunction with two instances of `liph:linguisticPhenomenon`, which model all the linguistic phenomena that occurred before and after `gs:N19b`, respectively.

All entities of Linguistic Phenomena Ontology defined up to now can be concisely summarized using the Description Logic syntax [18] through the constraints presented in (1), where `rdf` is an abbreviation for the RDF namespace [19].

$$\begin{aligned} \texttt{ontolex:Form} &\sqsubseteq \texttt{liph:LexicalObject} \\ \texttt{liph:LexicalObject} &\sqsubseteq \exists(\texttt{ontolex:writtenRep}).(\texttt{rdf:langString}) \\ \texttt{liph:linguisticPhenomenon} &\sqsubseteq \texttt{liph:LexicalObject} \times \texttt{liph:LexicalObject} \\ \texttt{liph:linguisticPhenomenon} &\sqsubseteq \texttt{liph:linguisticPhenomenon}^* \end{aligned} \tag{1}$$

As an example, Figure 6 illustrates how to represent the two possible derivations of the Italian "padre" in Figure 2, given that *lenition* $t > d$ and *truncation* are defined as subproperties of `liph:linguisticPhenomenon`.

Subproperties of `liph:linguisticPhenomenon` can be further organized into hierarchies to provide a more fine-grained classification of phenomena. For instance, one may define a superproperty *lenition* for properties indicating specific forms of lenition. In Section 4, property hierarchies are utilized to categorize Gallo-Sicilian language features based on linguistic and geographical information.

3.3. Operational Descriptions

Usually, the linguistic phenomena studied by linguistic researchers can be described *operationally*. For example, the language feature `gs:S22a`, typical of Southern Italy languages, can be described in human-readable way as follows

`gs:S22a replace the occurrences of ‘‘tj’’ with ‘‘sg’’ .`

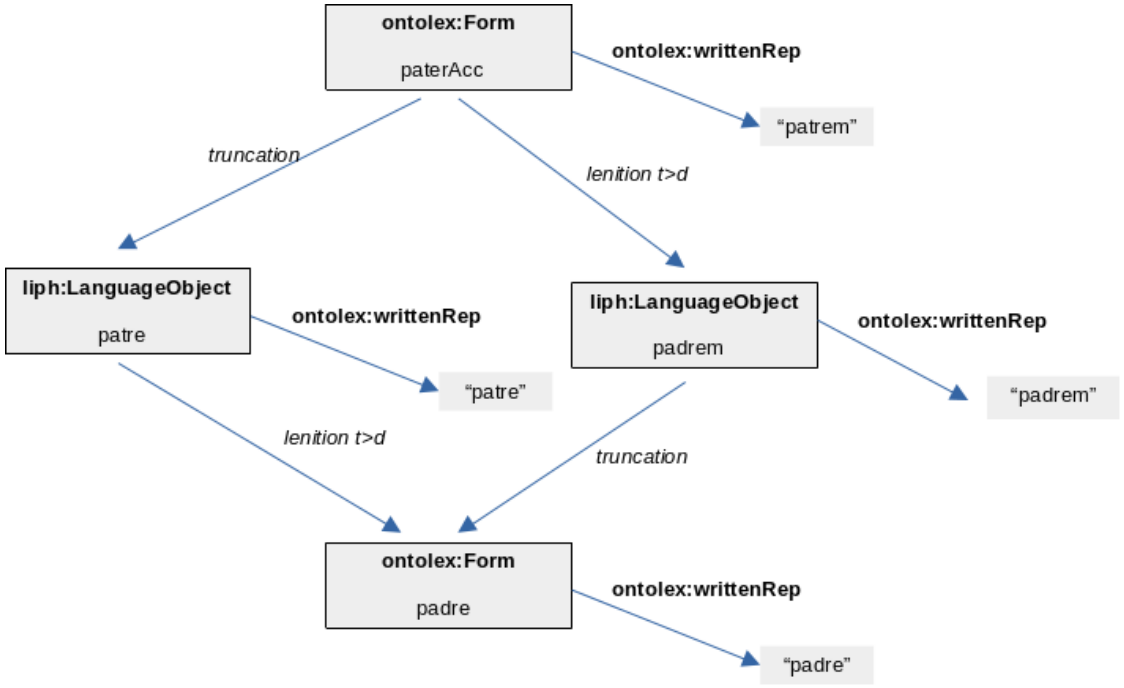

Figure 6: Derivations of "padre" from "patrem" with Linguistic Phenomena Ontology

However, to foster information reuse, a machine-readable description is more desirable when available. To this end, the Linguistic Phenomena Ontology defines two annotation properties, `liph:regex` and `liph:replacement`, which allow for describing a linguistic phenomenon in terms of:

- a *regular expression* (see [20]), used to recognize the parts of the parent strings that may be affected by the phenomenon, and
- one or more *replacements*, indicating how these parts will be modified.

The studies on Gallo-Italic varieties spoken in Sicily, which are the subject of Section 4, show that a relevant amount of language features can be described in this way, and several examples in this section will make use of features recognized in this context.

Note that `liph:regex` and `liph:replacement` are declared as annotation properties, so that they are ignored in reasoning. This approach avoids the use of metamodeling functionalities, which would lead to undecidability, as demonstrated in [21]. Instead, our ontology adheres to OWL-DL, a fragment of OWL whose decidability has been proven in [22].

In short, a regular expression is a sequence of characters (following the well-known syntax of regular expressions) that can match some or none of the substrings within a string.[1] For example, the regular expression associated with `gs:S22a` is simply the plain string "sg". Naturally, more complex ones can also be used. For example, consider

```
gs:N28b replace ‘‘cl’’ and ‘‘tl’’, with ‘‘ghj’’ .
```

Here the associated regular expression is `[ct]l`, matching every "c" or "t" followed by "l".

If the regular expression matches either the entire parent expression or just a substring of it, the linguistic phenomenon can be *applied* by replacing the matched parts as specified by the `liph:replacement` property.

The only replacement provided for `gs:N28b` is "ghj". However, there may be more than one replacement. An example is

[1] For the remainder of the paper, we will use regular expression defined for Java 8, available at https://docs.oracle.com/javase/8/docs/api/java/util/regex/Pattern.html .

```
gs:D04a replace ‘‘ptj’’ with ‘‘zzi’’ or ‘‘zzï’’
```

For this linguistic phenomenon, the regular expression is `ptj`, while the associated replacements are “zzi” and “zzï”. The application of `gs:D04a` to a parent expression containing “ptj” will generate two different target expressions for each occurrence of “ptj”: one where the occurrence is replaced with “zzi”, and another where it is replaced with “zzï”.

Regular expressions may contain sub-expressions enclosed in parentheses, known as *capturing groups*, which are numbered sequentially, starting at 1. Capturing groups can be used in replacements: every occurrence of $\$n$ in a replacement string will be substituted with the substring matched by the n-th capturing group.

Capturing groups are a well-known feature of regular expressions in Java. However, to clarify their usage in our context, let us provide an example related to the following feature:

```
sg:N29a replace ‘‘ct’’ with ‘‘it’’, provided that ‘‘ct’’ is
     internal in the parent string.
```

The regular expression associated to `gs:N29a` is `(\S)ct(\S)`, which contains two capturing groups intended to detect the characters preceding and following “ct”, respectively.[2] These captured characters are then used in the replacement “$1it$2”, so that the preceding and the following characters are correctly reported in the target string.

Let us note that `liph:regex` and `liph:replacement` are defined as annotation properties, which means no constraints (in the sense of OWL) can be enforced on these properties. However, declaring multiple `liph:regex` for the same linguistic phenomenon should be considered inappropriate. Analogously, a linguistic phenomenon with a `liph:regex` annotation but with no `liph:replacement` specified has to be considered inappropriate as well.

Moreover, as annotation properties, `liph:regex` and `liph:replacement` are not taken into account by *reasoners* (see [23, 24] as examples of classical reasoners). However, if a linguistic phenomenon is described using these properties, an automated agent could easily verify whether any statement in an ontology is compliant with it or not. Let us describe how such an agent should operate by means of some examples. First, let us consider the following definition for the *truncation* phenomenon of Section 2:

```
@prefix ex: <https://gallosiciliani.unict.it/examples/validation#> .

ex:truncation a liph:linguisticPhenomenon;
        liph:regex "^(.*)m$";
        liph:replacement "$1".
```

Now, let `ex:lux`, `ex:luce`, and `ex:lucem` be three instances of `liph:LexicalObject` with written representations “lux”, “luce” and “lucem”, respectively.

It can be automatically determined that the following statement is not compliant with the definition of `ex:truncation` just provided as the regular expression defined in it does not match any substring of “lux”:

```
ex:lux ex:truncation ex:luce .
```

Instead, the written form of `ex:lucem` matches the regular expression provided for `ex:truncation`. The only string that can be obtained from “lucem” by replacing the `$1` part of the replacement (which, in this case, is the entire replacement) with the captured string “luce”, corresponding to the capturing group `(.*)`, is “luce”. Thus, the following statement complies with the definition of `ex:truncation`:

```
ex:lucem ex:truncation ex:luce .
```

Conversely, any statement relating `ex:lucem` to any other individual whose written representation differs from “luce” through the property `ex:truncation` would violate the definition of `ex:truncation`, such as the following example:

[2] In regular expressions, \S matches any alphabetic character.

```
ex:lucem ex:truncation ex:lux .
```

4. Linguistic phenomena for Gallo-Sicilian varieties

As mentioned in Section 1, some Gallo-Italic varieties are spoken in Sicily as a consequence of a medieval immigration from north-western Italy. These varieties descend from the original Gallo-Italic varieties of northern Italy, but have been affected by a long-term contact with Sicilian. However, since both Gallo-Italic and Sicilian varieties are Romance languages, most of the terms in these languages have been inherited from Latin.

Let us 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 of these features, in particular those concerning the varieties spoken in Nicosia, Sperlinga, San Fratello, and Novara di Sicilia, have been identified [4, 25]. Furthermore, some of these features have been classified as typical of northern or southern Italy. This may be helpful in determining whether a Gallo-Sicilian lexeme has been inherited from a source Gallo-Italian variety, or it has been borrowed from Sicilian.

Then, these features have been reported in the *Gallo-Sicilian Features Ontology*, which is based on the Linguistic Phenomena Ontology described in Section 3, and that is structured as follows.

The Gallo-Sicilian Features Ontology chiefly consists of a set of subproperties of `liph:linguisticPhenomenon`, each one corresponding to a Gallo-Sicilian Feature. All properties in the ontology are in the following namespace:

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

First, four specializations of `liph:linguisticPhenomenon` are defined, each one devoted to one of the varieties under consideration:

$$\begin{aligned} \texttt{gs:nicosiaFeature} &\sqsubseteq \texttt{liph:linguisticPhenomenon} \\ \texttt{gs:sperlingaFeature} &\sqsubseteq \texttt{liph:linguisticPhenomenon} \\ \texttt{gs:sanFratelloFeature} &\sqsubseteq \texttt{liph:linguisticPhenomenon} \\ \texttt{gs:novaraFeature} &\sqsubseteq \texttt{liph:linguisticPhenomenon} \end{aligned} \tag{2}$$

Note that these properties are not mutually disjoint, as the same feature may span several languages. For example, the `gs:S22a` property mentioned in Section 3.3 is a subproperty of both `gs:nicosiaFeature` and `gs:novaraFeature`.

Orthogonally, two `liph:linguisticPhenomenon` specializations are provided for features typical of northern and southern Italy:

$$\begin{aligned} \texttt{gs:northernItalyFeature} &\sqsubseteq \texttt{liph:linguisticPhenomenon} \\ \texttt{gs:southernItalyFeature} &\sqsubseteq \texttt{liph:linguisticPhenomenon} \end{aligned} \tag{3}$$

Finally, properties for all the recognized Gallo-Sicilian features are reported as subproperties of at least one of `gs:nicosiaFeature`, `gs:sperlingaFeature`, `gs:sanFratelloFeature` and `gs:novaraFeature`, and, when appropriate, as subproperties of one of `gs:northernItalyFeature` and `gs:southernItalyFeature`. In addition, operational descriptions are provided in terms of `liph:regex` and `liph:replacement`, as described in Section 3.3, for a relevant amount of features.

As an example, we report a turtle fragment with the definition of `gs:S22a` as a southern Italy feature typical of the variety spoken in Nicosia.

```
gs:S22a a gs:southernItalyFeature, gs:nicosiaFeature;
        liph:regex "tj";
```

```
        liph:replacement "sg".
```

Examples of derivations of Gallo-Sicilian lexical expressions from Latin etymons can be retrieved as indicated in Appendix A.

Based on (1) and (2), and provided that some language expressions and their derivations are encoded using the Linguistic Phenomena Ontology described in Section 3 alongside the Gallo-Sicilian Features Ontology presented here, one can retrieve, for example, all expressions exhibiting features typical of one of the varieties under consideration, assuming reasoning functionalities are available.

In fact, all the instances of `ontolex:Form` with features typical of the variety spoken in Nicosia can be retrieved by the following SPARQL [26] query, based on the prefixes defined thus far:

```
SELECT ?expression WHERE {
        ?etymon liph:linguisticPhenomenon ?x .
        ?x gs:nicosiaFeature ?y .
        ?y liph:linguisticPhenomenon ?expression
}
```

Moreover, thanks to (3), expressions with features typical of northern and southern Italy can be retrieved as well. For example, the following query can be used to retrieve all those lexical expressions with northern Italy features:

```
SELECT ?expression WHERE {
        ?etymon liph:linguisticPhenomenon ?x .
        ?x gs:northernFeature ?y .
        ?y liph:linguisticPhenomenon ?expression
}
```

5. Related Works

Connecting etymologies to the linguistic phenomena that characterized them can be valuable for historical-linguistic researchers. The *OntoLex-lemon Etymological Extension*, abbreviated as *lemonEty*, provides specific representational features for etymological information, as detailed in [27]. It defines the following namespace:

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

In lemonEty, etymologies are defined as *hypotheses about the history of lexical elements*. In particular, simple and complex etymologies, tracing back the hypothetical provenance of a lexeme through inheritance and borrowings, can be described in a graph shaped fashion with the class `lemonEty:EtyLink`.

A `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 that refer to the entries involved in the `lemonEty:EtyLink`.

We claim that the Linguistic Phenomena Ontology of Section 3 may be used here to provide more detailed and automatically-verifiable derivations between forms, thus providing solid evidences to etymologies. For example, explicitly stating that “luce” can be derived from “lucem” through the `ex:truncation` feature in Section 3.3, as illustrated in Figure 7, would reinforce the etymology presented in the previous example.

In this paper, we focused on phenomena relevant to historical linguistics; however, other types of phenomena should also be considered. The linguistic phenomena related to word formation, declension, and inflection are core subjects of study in *morphology*. [28] describes an ongoing effort to develop an OntoLex-lemon module for linguistic morphology, called *OntoLex-Morph*. This module provides a class `morph:Rule` to describe how grammatical rules apply to generate new lexical expressions

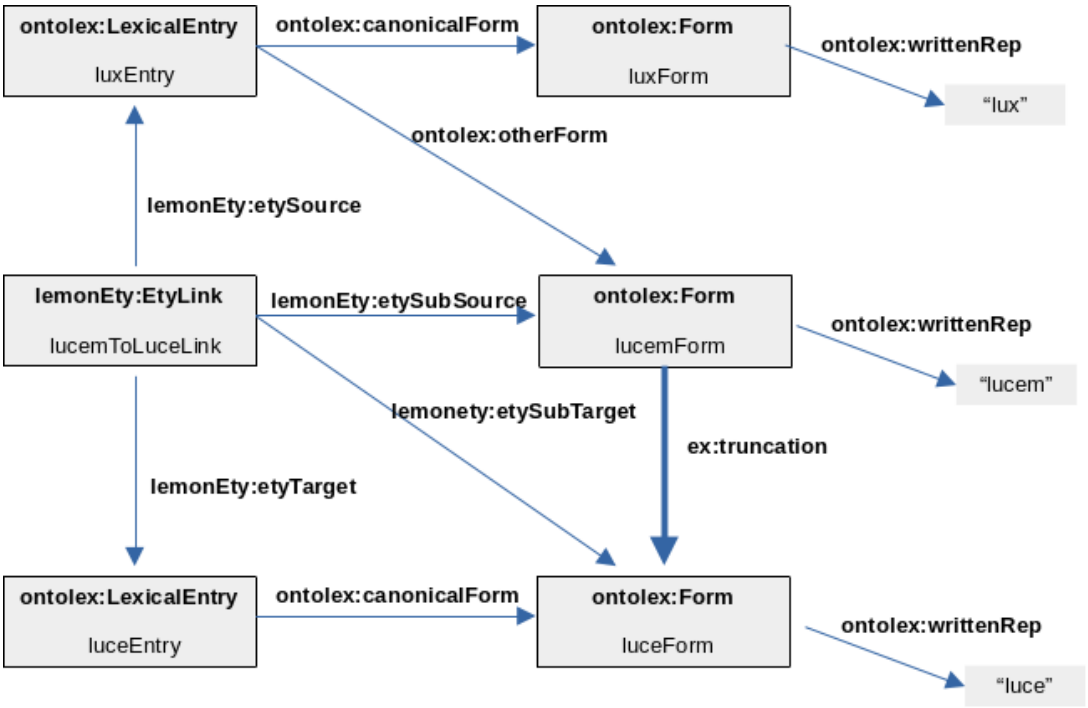

Figure 7: Etymology for the Italian lexeme "luce" expressed using etyLink and LiPh

from some source 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 3.3. However, at the current stage of development of the draft, derivation details cannot be described, in particular those involving intermediate forms.

More in general, several *Rule-based language technologies*, which could benefit of an OWL encoding, have been devised during the years (see [29, 30]). In particular, modelling *phonological rules*, as described in [30], using the representational features introduced in Section 3.3 appears to be quite straightforward.

We conclude this section by noting that linguistic phenomena modelled using our ontology are *perdurant* properties of forms and, more generally, of strings. In light of this, representation models concerning *diachronic* data, such those summarized in [31], has not been considered.

6. Conclusions and Future Works

In this paper, we presented the Linguistic Phenomena Ontology, an OWL ontology based on OntoLex-lemon, designed to represent changes in the morphology and pronunciation of lexemes. We further extended it to model the linguistic phenomena characterizing the Gallo-Italic varieties spoken in Sicily, aiming to study the etymology of lexemes in these varieties.

Some aspects of the OWL encoding of linguistic phenomena proposed here require further study. As discussed above, the `liph:linguisticPhenomenon` property connects two forms, specifically instances of `liph:LexicalObject`, which is defined as a superclass of `ontolex:Form`. However, forms may be associated with different written representations, and they may also have phonetic representations. Thus, the issue of determining which are the representations of two lexical objects involved in a linguistic phenomenon has to be assessed.

In addition, applications of the Linguistic Phenomena Ontology to morphology and connections to rule based technologies has to be investigated.

Acknowledgments

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", founded by the European Union – Next Generation EU, Mission 4, Component 1, CUP J53D23017360001 - ID P2022YWS8T; Research Unit of the University of Catania.

References

[1] S. G. Thomason, T. Kaufman, Language Contact, Creolization, and Genetic Linguistics, University of California Press, 1988. URL: http://dx.doi.org/10.1525/9780520912793. doi:10.1525/9780520912793.

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

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

[4] 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.

[5] 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/.

[6] C. Fellbaum (Ed.), WordNet: an electronic lexical database, Massachusetts: The MIT Press, 1998. P.423.

[7] S. Farrar, T. Langendoen, A linguistic ontology for the semantic web, GLOT International 7 (2003).

[8] B. Pedersen, J. McCrae, C. Tiberius, S. Krek, ELEXIS - a European infrastructure fostering cooperation and information exchange among lexicographical research communities, in: F. Bond, P. Vossen, C. Fellbaum (Eds.), Proceedings of the 9th Global Wordnet Conference, Global Wordnet Association, Nanyang Technological University (NTU), Singapore, 2018, pp. 335–340. URL: https://aclanthology.org/2018.gwc-1.40.

[9] T. Tasovac, L. Romary, P. Banski, J. Bowers, J. de Does, K. Depuydt, T. Erjavec, A. Geyken, A. Herold, V. Hildenbrandt, M. Khemakhem, S. Petrović, A. Salgado, A. Witt, TEI Lex-0: A baseline encoding for lexicographic data, Technical Report, DARIAH Working Group on Lexical Resources, 2018. URL: https://dariah-eric.github.io/lexicalresources/pages/TEILex0/TEILex0.html.

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

[11] 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.

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

[13] 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.

[14] 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.

[15] 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. URL: http://dblp.uni-trier.de/db/conf/wikidata/wikidata2023.html#LindemannAKMIP23.
[16] 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/.
[17] 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.
[18] F. Baader, D. Calvanese, D. McGuinness, D. Nardi, P. Patel-Schneider, The Description Logic Handbook: Theory, Implementation and Applications, Cambridge University Press, 2003.
[19] G. Klyne, J. J. Carroll, Resource Description Framework (RDF): Concepts and Abstract Syntax, Technical Report, W3C, 2004. URL: http://www.w3.org/TR/rdf-concepts/.
[20] S. Kleene, Representation of Events in Nerve Nets and Finite Automata, in: C. Shannon, J. McCarthy (Eds.), Automata Studies, Princeton University Press, 1956, pp. 3–41.
[21] B. Motik, On the Properties of Metamodeling in OWL, Journal of Logic and Computation 17 (2007) 617–637.
[22] I. Horrocks, P. F. Patel-Schneider, Reducing OWL Entailment to Description Logic Satisfiability, Journal of Web Semantics 1 (2004) 345–357.
[23] B. Glimm, I. Horrocks, B. Motik, G. Stoilos, Z. Wang, HermiT: An OWL 2 Reasoner, Journal of Automated Reasoning 53 (2014) 245–269.
[24] E. Sirin, B. Parsia, B. C. Grau, A. Kalyanpur, Y. Katz, Pellet: A practical owl-dl reasoner, Web Semantics: Science, Services and Agents on the World Wide Web 5 (2007) 51 – 53. URL: http://www.sciencedirect.com/science/article/pii/S1570826807000169. doi:http://dx.doi.org/10.1016/j.websem.2007.03.004, <ce:title>Software Engineering and the Semantic Web</ce:title>.
[25] S. C. Trovato, Galloitalische sprachkolonien. i dialetti galloitalici della sicilia, Kontakt, Migration Und Kunstsprachen 7 (1998) 538–559.
[26] World Wide Web Consortium, SPARQL 1.1 Query Language, 2013. URL: http://www.w3.org/TR/sparql11-query.
[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] C. Chiarcos, K. Gkirtzou, A. F. Khan, P. Labropoulou, M. Passarotti, M. Pellegrini, Computational morphology with ontolex-morph., in: T. Declerck, J. P. McCrae, E. Montiel-Ponsoda, C. Chiarcos, M. Ionov (Eds.), LDL@LREC, European Language Resources Association, 2022, pp. 78–86. URL: http://dblp.uni-trier.de/db/conf/acl-ldl/acl-ldl2022.html#ChiarcosGKLPP22.
[29] A. Hurskainen, K. Koskenniemi, T. Pirinen, L. Antonsen, E. Axelson, E. Bick, B. Gaup, S. Hardwick, K. Hiovain, F. Karlsson, K. Lindén, I. Listenmaa, I. Mikkelsen, S. Moshagen, A. Ranta, J. Rueter, D. Swanson, T. Trosterud, L. Wiechetek, Rule-Based Language Technology, 2023.
[30] R. M. Kaplan, M. Kay, Regular models of phonological rule systems, Computational Linguistics 20 (1994) 331–378. URL: https://aclanthology.org/J94-3001.
[31] H.-U. Krieger, 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, 2014.

A. Online Resources

The Linguistic Phenomena Ontology and the Gallo-Sicilian Features Ontology are stored into the GIT repository available at https://github.com/Gallosiciliani/languagefeatures. In addition, the repository contains some example derivations of lexical expressions of the Gallo-Sicilian variety of Nicosia from their Latin etymons at `/examples/nicosia`, and a prototypical implementation of a derivations validator.