DynaMorphPro: A New Diachronic and Multilingual Lexical Resource in the LLOD ecosystem

Matteo Pellegrini[1], Valeria Irene Boano[2], Francesco Gardani [3], Francesco Mambrini[1], Giovanni Moretti[1], Marco Passarotti[1]

[1]Università Cattolica del Sacro Cuore, Milano, [2]KU Leuven, [3]Universität Zürich

Correspondence: matteo.pellegrini@unicatt.it

Abstract

This paper describes the release as Linguistic Linked Open Data of DynaMorphPro, a lexical resource recording loanwords, conversions and class-shifts from Latin to Old Italian. We show how existing vocabularies are reused and integrated to allow for a rich semantic representation of these data. Our main reference is the OntoLex-lemon model for lexical information, but classes and properties from many other ontologies are also reused to express other aspects. In particular, we identify the CIDOC Concept Reference Model as the ideal tool to convey chronological information on historical processes of lexical innovation and change, and describe how it can be integrated with OntoLex-lemon.

1 Introduction

In the last decade, remarkable efforts have been made aiming to allow for a rich semantic modelling of linguistic information. Researchers and practitioners working in this framework have called attention to the need of data to be FAIR, i.e., Findable, Accessible, Interoperable and Reusable (Wilkinson et al., 2016), so as to maximise their exploitation for different purposes. To this aim, they have strived to follow the principles of the Semantic Web and of Linked Open Data (Berners-Lee et al., 2001), making machine-readable structured data available with open licenses in non-proprietary formats, following standards and guidelines developed by the World Wide Web Consortium (W3C) – namely, the RDF data model (Lassila and Swick, 1998) to encode information and the SPARQL query language (Prud'Hommeaux and Seaborne, 2008) to retrieve it – and linking data from different sources, so as to create a virtuous ecosystem of interoperability. Data of this kind have nowadays reached a remarkable coverage both in terms of number of languages and in terms of types of resources, as summarised by the graph provided in the Linguistic Linked Open Data cloud.[1]

Not only are many languages represented by virtue of individual resources, but recent years have witnessed the emergence of projects aiming at increasing the interconnection between the various resources available for a single language specifically. For instance, the LiLa (Linking Latin) project built a Knowledge Base of interoperable resources for Latin (Passarotti et al., 2020),[2] that currently includes 15 corpora pertaining to different epochs and 16 lexical resources documenting different aspects (such as semantics, etymology, polarity, morphology), and open to continuous additions and enrichments. In the wake of this effort, similar projects have been undertaken for other languages, e.g. LiITA (Linking Italian; cf. Litta et al., 2024)[3] and MOLOR (Morphologically Linked Old Irish Resource; cf. Fransen et al., 2024).

Best practices have been defined for many of the facets of language-related information that can be taken into consideration. A crucial prerequisite for such an enterprise is being able to model and harmonise the categories used for language description. To this aim, on the one hand, terminological repositories have been gathered including as many of the necessary categories as possible, and potentially expandable if needed, such as GOLD, (Farrar and Langendoen, 2003) and LexInfo (Cimiano et al., 2011); on the other hand, strategies have been devised to be able to accommodate the slightly different usage that can be made of such categories in different contexts, as can be done by means of the Ontologies of Linguistic Annotations (OLiA; cf. Chiarcos and Sukhareva, 2015).

These categories can then be used to represent the information provided in language resources of different kinds, including both textual and lexical

[1] `https://linguistic-lod.org/llod-cloud`

[2] `https://lila-erc.eu`.

[3] `https://www.liita.it`.

resources. For textual resources, there are RDF-compliant formats for their release such as CoNLL-RDF (Chiarcos and Fäth, 2017), and vocabularies that allow to model the annotations that can be added to corpora at different levels, such as POWLA (Chiarcos, 2012). On top of that, the NLP Interchange Format aims to achieve interoperability between resources of different kinds, their annotation, and NLP tools (Hellmann et al., 2013). For lexical resources, the *de-facto* standard is the OntoLex-lemon vocabulary (McCrae et al., 2017), that consists of a core model and several modules for more specific information (see Section 2).

In this work, we build on such previous efforts to release as Linguistic Linked Open Data (LLOD) the lexical database gathered and used by Gardani (2013) to explore the dynamics of morphological productivity in noun inflection from Latin to Italian – hence the name, DynaMorphPro. While these data are not very extensive in terms of number of entries, they provide rich and structured information on several aspects. They are multilingual: Latin and Italian are the primary object of inquiry, but many other languages appear as also the etymology of loanwords is provided. Morphological information is provided regarding both inflection classes and derivation – mostly, conversions. Diachrony is also involved as cases of shifts from one class to another are documented, and the time at which they are attested is specified (see Section 3).

We show how we exploit the potential of the LLOD ecosystem to offer a rich semantic modelling of these data. On the one hand, language-specific projects for Latin (LiLa) and Italian (LiITA) allow for interoperability with other resources for those languages. On the other hand, the OntoLex-lemon model gives us ways to represent many of the pieces of information provided, including morphology (with Morph, see Section 2) and attestation in texts (with FrAC, see Section 2). For other pieces of information, we make proposals to integrate other vocabularies, such as lemon-Ety (Khan, 2018) for etymology and CIDOC-CRM (Doerr, 2003) for time information.

The remainder of this paper is structured as follows. In Section 2, we review previous work and describe the existing vocabularies on which our own model is based. In Section 3, we describe the data, giving some background on the original aims and overall structure of the resource, and further details on the information it provides. In Section 4, we outline our model, showing how we reused existing vocabularies and the new classes and properties that we introduced. In Section 5, we describe the process of linking entries of our resource to lemmas of the Knowledge Bases available for Latin (LiLa) and Italian (LiITA). Section 6 concludes and highlights possibilities for future work.

2 Reference Vocabularies

2.1 Vocabularies for Lexical Information

The application of Semantic Web and Linked Open Data principles to linguistic data raised the issue of being able to provide a more expressive representation of lexical information related to ontology entities. To this aim, the Ontology Lexicon (OntoLex) community group of the W3C built upon a previously introduced Lexicon Model for Ontologies (lemon, McCrae et al., 2012) to release a new model, OntoLex-lemon (McCrae et al., 2017), which was published in 2016 as a W3C report.[4]

The model revolves around the class `ontolex:LexicalEntry`. Information can be provided on both form and meaning of lexical entries. For the former, there is a class `ontolex:Form` and a property `ontolex:lexicalForm`, with subproperties `ontolex:canonicalForm` for the citation form and `ontolex:otherForm` for other cases, and different variants of the same form are coded through a datatype property `ontolex:representation`, with subproperties `ontolex:writtenRep` and `ontolex:phoneticRep`. For the latter, there are classes and properties both for concepts (`ontolex:LexicalConcept`, `ontolex:evokes`) and for senses (`ontolex:LexicalSense, ontolex:sense`, and the two can be connected through the property `ontolex:lexicalizedSense`.[5]

Besides the core model, additional modules have been released to deal with specific aspects in more detail, including syntax and semantics (synsem module), decomposition of complex lexical entries (decomp module), variation and translation (vartrans module), metadata (lime module), and lexicographic information (lexicog[6] module). For our purposes, the most relevant modules have not been released yet, but are at an advanced stage of

[4] `https://www.w3.org/2016/05/ontolex/`.

[5] Inverse properties `ontolex:isEvokedBy`, `ontolex:isSenseOf`, `ontolex:isLexicalizedSenseOf` are also defined.

[6] `https://www.w3.org/2019/09/lexicog/`. The other modules are documented in the same web page of the core model.

development, namely Morph and FrAC.

The Morph module, in its latest draft[7] (cf. also Chiarcos et al. 2022b), has been devised to be able to express information on the one hand on inflection, including what is provided in our resource, namely inflection classes (class `morph:InflectionClass`, in the range of the property `ontolex:morphologicalPattern` in the core model); on the other hand on word formation, including what is provided in our resource, namely relations between words that are converted from one part of speech to another (class `morph:WordFormationRelation`, connected to the source and target word through the properties `vartrans:source` and `vartrans:target`).

The FrAC module, as recently described in Chiarcos et al. (2022a), provides a vocabulary to describe the actual usage of lexical items in texts, such as their attestations, frequencies and further information that can be found in corpora. For the purposes of our resource, we will be concerned only with attestations. In the module, there is a dedicated property `frac:attestation` that should be used to link lexical entries to usage examples provided in lexical resources about them. This is defined as a sub-property of `frac:citation`, that can be used for attestations from secondary sources, with a recommendation to use vocabularies for bibliographic information (on which see 2.3).

Etymologies are another piece of information that is frequently provided in lexical resources. For a modelling of this kind of information, an (external) extension of the OntoLex-lemon model has been proposed by Khan (2018), lemonEty, that provides classes and properties for etymologies themselves (`lemonEty:Etymology`, the reification of a scientific hypothesis about the history of a linguistic item, and the associated property `lemonEty:etymology`), for etymons involved in them (`lemonEty:Etymon` and the associated property `lemonEty:etymon`), and for the relation between two elements in an etymology (`lemonEty:EtyLink`, and the associated properties `lemonEty:hasEtyLink`, `lemonEty:etySource` and `lemonEty:etyTarget`).

2.2 Vocabularies for historical and chronological information

Immediately since the creation of the World Wide Web, political and cultural institutions operating in Cultural Heritage began to disseminate information and grant wider access to their data and collections on the WWW. The spread of information available online has inevitably raised the question of interoperability and standardisation (Doerr and Iorizzo, 2008). The CIDOC Concept Reference Model (CIDOC-CRM), an ontology developed since the end of the 1990's with the aim of providing a common model for the documentation of Cultural Heritage institutions, has emerged as a successful and widely adopted solution to this end. Originally curated by the International Committee for Documentation (CIDOC) of the International Council of Museums (ICOM), the CRM is recognized as an ISO standard since 2006 and the status was lastly renewed in 2023 (ISO 21127:2023).

Instead of focusing on producing metadata schemas to facilitate the data-capturing and data-entry process, the CRM attempts to represent the underlying meaning of the information. While the standards that are more oriented toward data entry, like e.g. the Dublin Core Metadata Initiative (DCMI),[8] aim to dictate what should be documented, the CRM emphasises how the data are conceptually related (Doerr, 2003). For these reasons, the CRM does not provide a 'flat' vocabulary for metadata or a fine-grained taxonomy of the different entity types, but is built as a high-level ontology that focuses on capturing the relations between entities (Doerr and Iorizzo, 2008).

This design choice is also the consequence of the fact that museum information is built from heterogenous data and may "virtually describe[s] the whole world as manifested in material objects from the past" (Doerr, 2003, 77). In its current official release (7.1.3), the CRM includes ca. 90 classes and 160 properties. Some of the most important of them are used to identify the basic concepts required to document the history of ideas, artifacts and environments. They include persons (`crm:E21_Person`), places (`crm:E53_Places`), human-made objects (`crm:E22_Human-made_Object`), conceptual objects (`crm:E28_Conceptual_Object`), and temporal entities (`crm:E2_Temporal_Entity`). A fundamental subclass of the latter is `crm:P5_Event`, which is used for delimited and coherent processes that affect all entities belonging to the class of `crm:E77_Persistent_Item`. Participation in events is in fact a crucial aspect to encode historical information in the CRM and to connect different

[7] `https://github.com/ontolex/morph`.

[8] `https://www.dublincore.org/`.

datasets in a semantic network.

The CRM is primarily designed for the documentation of institutions operating in the GLAM (Galleries, Libraries, Archives, and Museums) sector. However, thanks to its being designed as a high-level ontology, it is general enough to be applicable to any type of 'intangible' heritage, however broadly defined. This includes languages. Indeed, one of the central ideas of our paper is that the CRM is the ideal model to express what neither OntoLex-lemon nor FrAC are capable of capturing, i.e., the historical process of innovation and invention introduced by speakers in languages. In a previous discussion, Khan (2020) proposed a model based on OntoLex-lemon to integrate diachronic information about lexical entries. While that work included many important suggestions, we believe that the CRM is the right reference model to express this type of information, both because, while not integrating classes explicitly designed for linguistic concepts, it is capable of accommodating lexical data, and because it provides a general framework to document language change within its larger historical and social context, if researchers decide to do so. What our solution shares with Khan (2020) is the adoption of the ontology OWL-Time to encode relations between periods and their anchoring to a timeline (Gangemi et al., 2017).[9]

2.3 Vocabularies for Citations and References

While pioneering attempts to allow for a semantically rich representation of the domain of publishing – such as, among others, the Functional Requirements for Bibliographic Records (FRBR) by the International Federation of Library Association and Institution, later formalized as an ontology complementing the CIDOC-CRM (FRBRoo), and the OWL-native vocabulary of the Bibliographic Ontology (BIBO)[10] – deserve to be credited, in this work we refer to a more recent suite of complementary and orthogonal ontologies that have been developed for the modelling of Semantic Publishing and Referencing (SPAR, Peroni and Shotton, 2018), building on those previous efforts.

In particular, from that suite we use the FRBR-aligned Bibliographic Ontology (FaBiO),[11] designed to allow for the modelling of entities that are published or potentially publishable. FaBiO takes from the FRBR model the core distinction between classes corresponding to decreasing levels of abstraction, going from `Work` (e.g., Homer's Odissey), to its `Expression` (e.g., the English text of Homer's 'Odyssey' translated by Robert Fagles) through the property `realization`, to its `Manifestation` (e.g.,'The Illustrated Odyssey', published by Sidgwick & Jackson Ltd in 1980) through the property `embodiment`, to its `Item` (e.g., the copy of the latter at some library) through the property `exemplar`. Additionally, in FaBiO new properties are introduced to allow for a direct mapping between all levels (e.g., `fabio:hasManifestation` to map a `Work` to its `Manifestation`).

We also use the Bibliographic Reference Ontology (BiRO),[12] designed to allow for the modelling of bibliographic references and records, through the classes `biro:BibliographicReference` and `biro:BibliographicRecord`, and the property `biro:references` to map them to works.

3 The Data

The original data are extracted from a monograph by Gardani (2013), which explores the evolution of the productivity of the noun inflection classes of Latin and Old Italian. The goal of Gardani (2013) was to better understand the mechanisms that guide and constrain natural grammar, specifically what factors determine changes in the productivity of inflection classes, leading to the emergence of new ones, an increased or decreased degree of productivity through to the loss of extant ones. The object languages – Latin and Old Italian – were chosen among other reasons because they are well-documented and embody a diachronic development spanning almost 2,000 years: the Latin data range from the *Leges Duodecim Tabularum* (451-450 BCE) to Late Latin (200-600) and Early Medieval Latin (600-800); the data of Old Italian, as one of its continuers, range from *Indovinello veronese* (early 9th century) through 1375 (1400). The data were analyzed by applying a metric of productivity originally proposed by Dressler (2003) and there revised, based on a hierarchy of criteria reflecting the degree of impediment which a lexeme has to face when it is integrated into a specific inflection class. Productivity is here defined as "the force of attraction that inflectional patterns exert

[9] `https://www.w3.org/TR/owl-time/`.

[10] `https://writeanessayfor.me/bibliontology-com`.

[11] `https://sparontologies.github.io/fabio/current/fabio.html`.

[12] `https://sparontologies.github.io/biro/current/biro.html`.

on new lexemes (both foreign and native in origin) and on extant paradigms of native lexemes" (Gardani, 2013, p. 39). Inflection class productivity was measured on historical synchronic cuts, on the basis of the investigation of loanword integration, conversions, and class shift, with the data on the integration of loanwords being drawn from the contact languages Ancient Greek, Germanic, Arabic, Byzantine Greek, and Old French. The elaboration of the diachronic outline was encompassed by connecting the productivity degrees measured at each synchronic cut. The diachronic trajectory shows a progressive reduction in the number of the inflection classes from a total of at least 21 in Latin to a total of nine in Old Italian. Gardani (2013) showed that in the analyzed languages, the dynamics of growth and emergence of inflection classes are linked to the need of creating or restoring biunique relationships with respect to the realization of specific morphosyntactic features.

The resource provides rich and highly structured data on 2,434 lexical entries. All of them have been openly released on the basis of the model described in the rest of this paper. The primary data subdivision regards the language: entries are grouped into Latin (1,120) and Old Italian (1,314) items.[13] Each group is further divided into loanwords, conversions and class shifts. Additionally, each lexical entry is enriched with different types of further information: some pieces of information are shared by lexical entries of all types, while other ones are found only in relation with specific types.

All the entries regardless of their category are provided with a short definition of their meaning and with details pertaining to their first attestation. The latter may include information about the author and/or the document in which the word first appeared; for Old Italian entries, in many cases this is accompanied by a full reference to the text where the attestation is found; sometimes the geographical area and date of the attestation are supplied as well. For Old Italian entries, the language variety of the attestation is often provided (e.g. *fior.* for the variety spoken in Florence). Additionally, each lexical entry is classified by its inflection microclass, identified by an exemplary lexeme (e.g., *rosa rosae* 'rose' for Latin, or *casa case* 'house' for Old Italian), which is defined as a "set of paradigms which share exactly the same morphological and morphophonological generalizations" (Gardani, 2013, p. 26). Finally, all the lexical entries are grouped on a diachronic basis and are assigned to a specific chronological interval. For Latin, broad periods are defined, that correspond to the division into periods operated in studies on the history of the language: Archaic (451-240 BCE), Pre-Classical (240-75 BCE), Classical (75 BCE-14 CE), Post-Classical (14-200 CE), Late (200-600 CE), Early Medieval (600-800 CE) and Medieval Latin (800-1400 CE). For Old Italian, epochs consist of an indication of the interval of years in the range considered (ranging from 1000 CE to 1400 CE), using spans of 50, 100 or 150 years (e.g. "1101-1200").

Loanwords, conversions and class shifts are also provided with additional information specific to their own characteristics. For loanwords, the etymon is supplied, together with the specification of its language. With regard to Latin, this can be Etruscan, Ancient Greek or a Germanic language,[14] while for Old Italian, loanwords can be traced back to a Germanic language, Byzantine Greek, Arabic or Old French. Another relevant information provided in the case of loanwords concerns the (in)compatibility between the inflection microclass of the etymon, and that of the loanword itself (Gardani, 2013 pp. 39-41; see 4.1 for further details). For each conversion, the base verb from which it was derived and the latter's inflection microclass are provided. Finally, each class shift is provided with rich information about the entry in the original class, including its meaning, the etymon and the inflection microclass. In some cases, additional morphological information (e.g., the ending of the plural form or the genitive form) is also provided.

4 The Model

In this Section, we show how the reference vocabularies mentioned in Section 2 have been exploited to model the data described in Section 3, as well as the new classes and properties that we needed to introduce to allow for a complete representation of all the available information. We do that by providing examples and commenting on them in detail, with a focus on lexical information in Subsection 4.1, on historical and chronological information in Subsection 4.2 and on citations and references in Subsection 4.3.

[13] Sometimes, this requires a further specification, such as the identification of a more specific variety (e.g. Vulgar Latin), or an indication of the fact that the first attestation of an Old Italian loanword is still considered as a Latin form.

[14] The specific Germanic language is given only when the information is available.

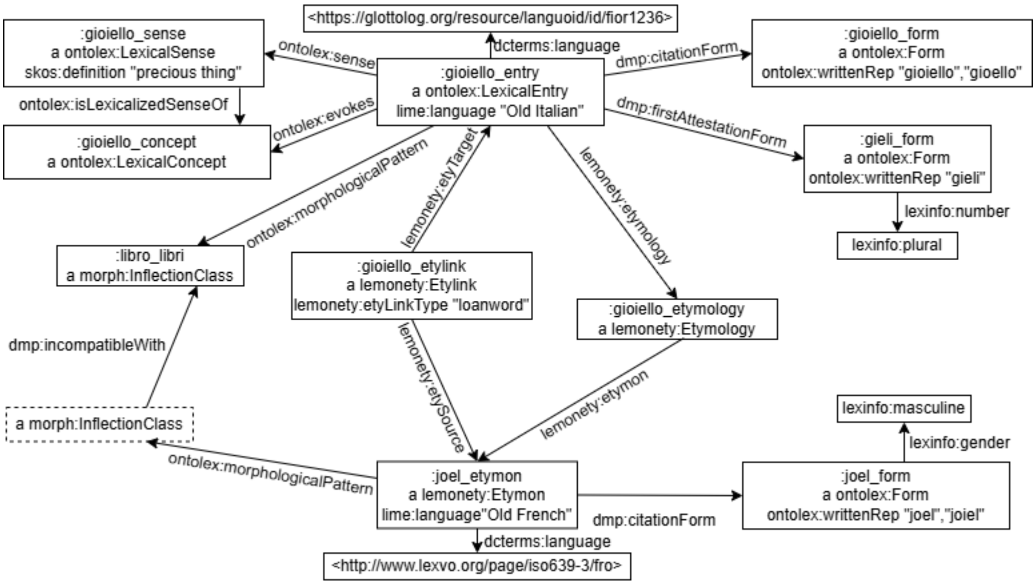

Figure 1: Modelling the It. loanword *gioiello* (< Fr. *joel*)

4.1 Modelling Lexical Information

To be able to provide a uniform treatment of the heterogeneous information provided by our resource, we consider it as a lexical resource providing etymological information of different kinds on its lexical entries. This is unproblematic for loanwords, but less obvious for conversions and inflection class shifts. However, if a wide definition of 'etymology' is assumed, such as the one implicit in the discussion of Mailhammer (2013), according to whom it can refer to anything that answers the question "where did that come from?", then it is reasonable to treat information of those kinds as etymological information too, in that it has to do with the history of words and their properties. In the case of conversions, a word is stated to come from another word belonging to another part of speech. In the case of class shifts, information is given on the fact that a word that used to be assigned to a specific inflection class starts being assigned to another one at some point in its history.

Consequently, words included in the resource are assigned to the class `ontolex:LexicalEntry`, and etymological information of different kinds is modelled using the lemonEty model. Lexical entries are connected through the property `lemonEty:etymology` to an instance of the class `lemonEty:Etymology`. Each etymology is linked through the property `lemonEty:etymon` to the `lemonEty:Etymon` provided by the resource for the lexical entry at hand: the corresponding lexical entry in the donor language for loanwords (see Figure 1), a lexical entry with a different part of speech in the same language for conversions (see Figure 2), a lexical entry with the same part of speech but a different inflection micro-class in the same language for class shifts. A `lemonEty:EtyLink` relation is also established that is connected through the property `lemonEty:etySource` to the etymon and through the property `lemonEty:etyTarget` to the lexical entry at hand. The property `lemonEty:etyLinkType` is used to distinguish between the different types of etymologies, namely "loanword", "borrowing" and "class shift".

Regarding conversions, the information provided by the resource can be given not only a diachronic, but also a synchronic interpretation: not only does the noun at hand comes from a corresponding word with a different part of speech, but there also exists a morphological relation between that noun and the word with different part of speech at some stage in the history of the language. To avoid neglecting this other interpretation, we redundantly code the same information also using the vocabulary of the emerging Morph module of OntoLex, following a strategy comparable to the one of previous works such as Pellegrini et al. (2021): we define a class `dmp:Conversion` as a sub-class of `morph:WordFormationRelation`, linked through the properties `vartrans:source` and `vartrans:target` to the input and output lexical entries, respectively. While the treatment of conversions as etymologies was crucial for uniformity with loanwords, their treatment as morphological information allows for interoperability with other resources providing information of that kind – like WFL for Latin (Litta and Passarotti, 2019). This makes it possible, for instance, to extract all the cases of conversions in Latin according to those different sources.

Due to the multilinguality of the resource, another important piece of information is the language of items of different kinds: indeed, such information is provided for both main lexical entries and their etymons on the one hand, and for the works from which attestations are taken on the other hand (see Subsection 4.3 below). As for lexical entries and etymons, following the recommendation of the OntoLex final model specifications,[15] on the one hand we code the name given to the language in the resource as a literal value using the datatype property `lime:language` from the lime module for metadata (Fiorelli et al., 2015); on the other hand, we link to URIs of controlled vocabularies through the property `dcterms:language`, from the DCMI. Whenever it is available, we use the URI provided for the ISO-639-3 code of the language on Lexvo.org (De Melo, 2015).[16] However, in some cases it is not possible to assign an ISO code corresponding to the information provided in the resource. For instance, some loanwords into Old Italian are only marked as coming from "Germanic", because it is difficult to decide from which specific Germanic language they have been borrowed. The Glottolog catalogue[17] also provides codes for families and their branches (in this case, `germ1287`[18]), thus allowing to express information at the appropriate level of granularity. Yet in other cases, it is excess, rather than lack of specificity that creates problems when looking for appropriate language codes. This is what happens for the languages of the works from which attestations of Old Italian forms are taken: in that case, the specific regional variety in which the work is written is specified (e.g., "Lombard Vulgar"), and sometimes even more detailed information is provided on the influence of other regional varieties (e.g. "Vulgar of Rome interfered by Tuscan"). Of course, this level of granularity is not achieved in any of the controlled vocabularies available for this purpose. As a consequence, we link to the closest match among the ISO and glottolog codes available (e.g., respectively, `lmo` and `lomb1257` for the former example), and we keep the original information as a literal, thus covering also cases where no corresponding code can be found (as happens for the latter example). Since in this case language information is predicated of works, to map to language names as literals we use the property `dcterms:language`, rather than `lime:language`, that could only be used for lexical entries.

Meaning is modelled using classes and properties from the core OntoLex model, i.e., `ontolex:evokes` to map to an instance of `ontolex:LexicalConcept` and `ontolex:sense` to map to a corresponding instance of `ontolex:LexicalSense`, with the gloss expressed as a literal using the property `skos:definition` from the SKOS vocabulary (Miles et al., 2005). Senses and concepts are related through the property `ontolex:isLexicalisedSenseOf`.

Also to record the forms listed in the resource for each entry we resort to core OntoLex vocabulary, where a property `ontolex:lexicalForm` is defined to map from entries to instances of `ontolex:Form`, alongside its sub-properties `ontolex:canonicalForm` and `ontolex:otherForm`. The former property is used for the linking to LiLa and LiITA (see Section 5 below). Since there is a cardinality restriction requiring at most 1 canonical form per lexical entry, it cannot be used in other cases. Consequently, we use the latter property for all other forms that are listed in the resource for each lexical entry. Furthermore, since there are subtle differences in the kinds of relations between lexical entries and forms in the resource, we define some new sub-properties of `ontolex:otherForm`, for specific cases, namely: `dmp:citationForm` for the citation form that is used in the resource; `dmp:modernItalianForm` when the resource also provides the corresponding form in contemporary Italian for Old Italian lexical entries; `dmp:latinForm` when the resource also provides the corresponding form in (Late) Latin for Old Italian lexical entries; `dmp:firstAttestationForm` for the form in which the lexical entry is first attested. The underspecified superproperty `ontolex:otherForm` is used in all cases that cannot be subsumed under one of the kinds just mentioned.

The most crucial piece of information for the original purpose for which the data were col-

[15] https://www.w3.org/community/ontolex/wiki/Final_Model_Specification#Metadata_.28lime.29.

[16] http://www.lexvo.org/.

[17] https://glottolog.org/.

[18] https://glottolog.org/resource/languoid/id/germ1287.

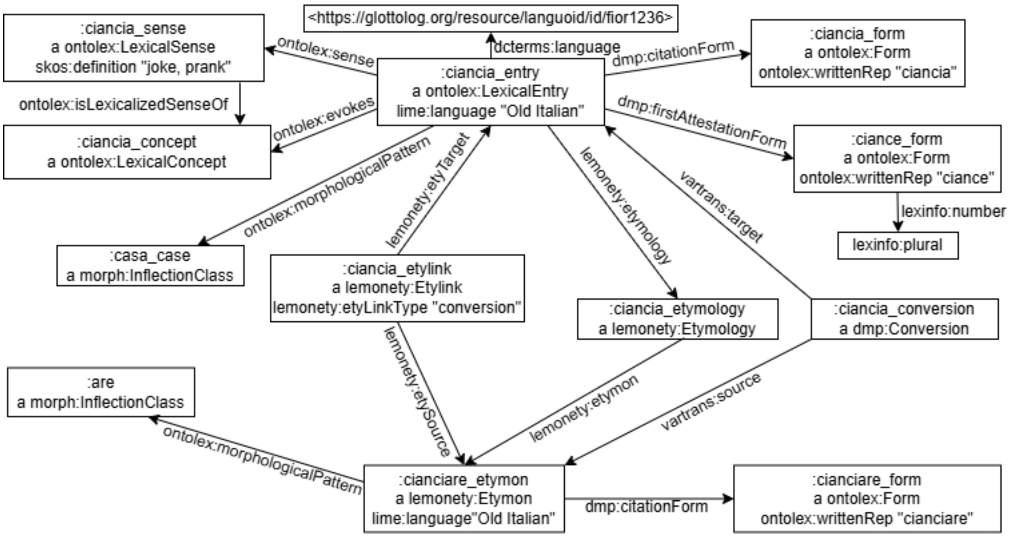

Figure 2: Modelling the Italian conversion from *cianciare* (V) to *ciancia* (N)

lected is the fine-grained inflectional behaviour ("inflection microclass") of lexical entries, as this is what informs users on differences in the degree of productivity: if an inflection class is frequently assigned to new items in the lexicon, such as loanwords from other languages and conversions from other parts-of-speech, or if it is frequently the new class assigned to nouns previously assigned to other classes, this indicates strong productivity. Information on the inflectional behaviour of entries is expressed using the property `ontolex:morphologicalPattern` of the core OntoLex model, that maps to instances of `morph:InflectionClass`, introduced in the Morph module.

In the case of loanwords, Gardani (2013) follows Dressler (2003) in distinguishing (i) cases in which a loanword is assigned to an inflection class in the recipient language based on compatibility of that inflection class with the one of the word in the donor language, from (ii) cases in which there is no such compatibility. For instance, the fact that the 1st-declension Ancient Greek noun *aithra* 'sky' is assigned to a micro-class of the 1st declension also when borrowed into Latin *aethra* is likely to be motivated by the fact that in some forms the endings that appear in the donor language are the same as the endings that would be used in the recipient language in the corresponding cell: e.g., the Greek NOM.SG *aithra* ends in *-a* exactly like 1st declension nouns in Latin. This is in turn due to the common diachronic source of the Greek and Latin 1st declension, that are both evolutions of Indo-European *-a-* stem nouns, thus producing a high degree of phonological and morphological comparability. Such an explanation cannot be invoked for the fact that Ancient Greek *lampas* 'torch' – belonging to the Greek 3rd declension and displaying NOM.SG in *-s* – is assigned to the 1st conjugation, and thus has NOM.SG in *-a-*, when borrowed into Latin *lampada*. This assignment cannot but be motivated by the attraction power of the inflection class in the recipient language, and can thus be taken as a stronger indication of its productivity. Accordingly, for each loanword recorded in the resource, there is an indication of the micro-class to which it is assigned in the recipient language on the one hand; on the other hand, the inflection class in the donor language is not always provided, but information is given on whether it is compatible with the class in the recipient language or not. To accurately reflect this state of affairs in RDF, we introduce blank nodes for the inflection class in the donor language when needed, and code compatibility (or lack thereof) between the inflection class in the recipient language and that blank node, as shown in Figure 1.

4.2 Modelling Historical and Chronological Information

Because of the diachronic spirit of the resource, it is crucial to be able to express the chronological information associated to items of different kinds in a semantically rich fashion. In the CIDOC-CRM, time information can be predicated of temporal entities – i.e., the class `crm:E2_Temporal_Entity` is in the domain of the property `crm:P4_has_time-span`. To accom-

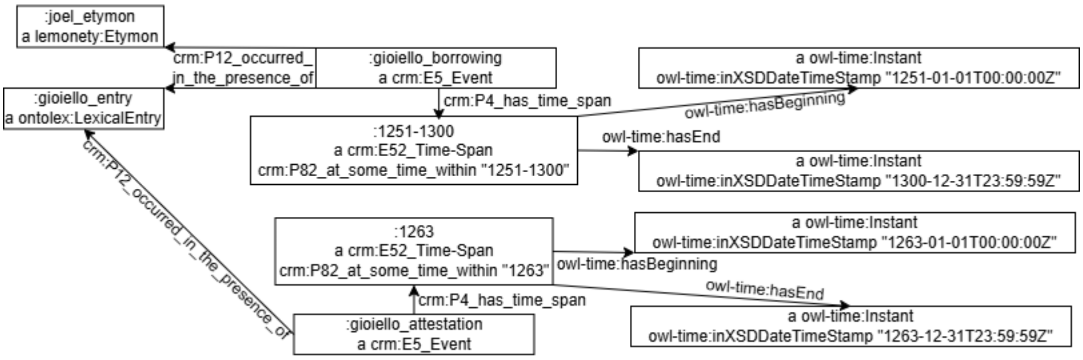

Figure 3: Modelling historical and chronological information on It. *gioiello*

modate for this requirement, we use a more specific sub-class of crm:E2_Temporal_Entity, and introduce a crm:E5_Event corresponding to the process by which each entry of our resource enters the lexicon of the language under consideration, or acquires different characteristics: both borrowing an item from one language to another and converting it from one part of speech to another can be considered as 'events', as well as shifts from one inflection class to another one. We then connect each event to both the entry itself and its etymon using the property crm:P12_occurred_in_the_presence_of, and associate it to the epoch when it occurred using the property crm:P4_has_time-span, pointing to an instance of crm:E52_Time-Span. For this purpose, we define time spans for each of the epochs mentioned above for Latin and Italian, as shown in Figure 3. According to the CIDOC-CRM specifications, the actual duration of time-spans can be expressed by means of the property crm:P82_at_some_time_within, that points to an instance of crm:E61_Time_Primitive, on its turn corresponding to a representation of the time span as a literal. To supplement this shallow coding with a semantically richer one that allows for queries exploiting the full potential of the information provided by the resource, we follow Khan (2020) and also express this using the OWL-Time ontology: each epoch is stated to begin (using the property owl-time:hasBeginning) and end (using the property owl-time:hasEnd) respectively at the owl-time:Instant corresponding to the first and last second of the years indicated in the resource, respectively.

For Old Italian entries, sometimes the coarse-grained information on the epoch at which a lexical entry can be approximately considered to have entered the lexicon is supplemented by a finer-grained information on the date at which it is first attested, on its turn based on the dating of the work in which it first appears. To express this additional piece of information, we introduce another instance of crm:E5_Event, this time corresponding to the event of the first documented usage of the lexical entry at hand. We then link this event to temporal information in the same way outlined before, using crm:P4_has_time_span pointing to a crm:E52_Time-Span further specified using the OWL-Time ontology. This accurately reflects the information provided in the resource: the date of the first attestation of a lexical item is more precise, but it cannot be taken as an indication of the time it became entrenched in the lexicon, which can have taken place before its documentation in texts, or even after if the first usage is just an occasionalism.

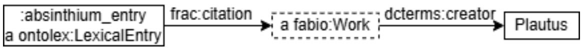

Figure 4: Modelling citations with blank nodes

4.3 Modelling Citations and References

The last piece of information that we need to cover concerns citations and references. Indeed, the resource provides information on the first attestations of entries. For Latin, most often, only an indication of the author who first used a form of the lexical entry at hand is given – e.g. the Ancient Greek borrowing *absinthium* is stated to be attested since Plautus. For Old Italian, in many cases this is accompanied by a reference to the text where the attestation is found – e.g., the borrowing of *veltro* into Italian is stated to be first documented in Dante's *Convivio*, also providing a full reference to the edition from which the variant has been taken.

Figures 4 and 5 show how we model those different possibilities. The property that we use is

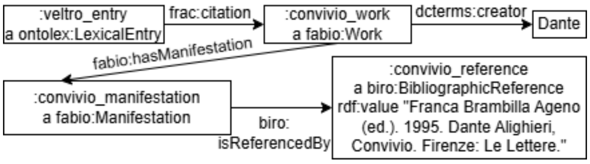

Figure 5: Modelling citations with actual works

frac:citation from the emerging FrAC module of OntoLex, connecting lexical entries to a representation of the works where they are attested.[19] For cases where there is a precise reference to a text, we introduce an actual instance of fabio:Work, and express additional information on it using properties from the DCMI – namely, dcterms:creator. For cases where there is no precise reference to a text, we use a blank node, about which we predicate the available pieces of information using the same properties.

It is reasonable to consider the first attestation of a lexical item as pertaining to the level of abstraction of fabio:Work – the relevant information is that, e.g., the borrowing *veltro* was first used in Dante's *Convivio* in the 14th century. However, the resource also provides a full citation of the modern edition where such usage is documented – in this case, namely, the one curated by Franca Brambilla Ageno in 1995. As a consequence, we also introduce a corresponding instance of the more concrete class fabio:Manifestation, and exploit the possibility of linking works to their manifestation directly by means of the property fabio:hasManifestation.[20]

To code the full bibliographic entry, we use the BiRO ontology, and predicate that each manifestation biro:isReferencedBy an instance of the class biro:BibliographicReference, with the full citation as the rdf:value.

5 Linking to LiLa and LiITA

In this Section, we detail the procedure that we followed for the linking of the entries of our resource to Knowledge Bases of interoperable resources available for the two languages – namely, LiLa for Latin and LiITA for Old Italian.

The architecture of the LiLa Knowledge Base is organised around the central class lila:Lemma, defined as a subclass of ontolex:Form that identifies forms that are potentially used to lemmatise a token in a corpus. Interoperability between different resources available for Latin is achieved by linking both tokens of textual resources and entries of lexical resources to the corresponding lemma, using the properties lila:hasLemma and ontolex:canonicalForm, respectively. Accordingly, we link entries of our resource to the LiLa Knowledge Base using the latter property. To find the corresponding lemmas, we take advantage of the list of forms provided by our resource on the one hand, and of the different form variants provided for each lemma in LiLa with the property ontolex:writtenRep on the other. Whenever there is a match between one of the forms of the resource and one of the written representations in LiLa, we record it. If at the end of the procedure there is only one match, we link our entry to the corresponding lemma. If there is more than one match, a process of semi-automatic disambiguation is performed, by first checking if there is also a match between the grammatical properties that are predicated of forms both in the resource and in LiLa, such as part of speech and inflection class, and then resolving remaining ambiguities manually. If no match is found, we enrich the Lemma Bank with new lemmas.

The more recent LiITA project (Litta et al., 2024) is strongly inspired to its predecessor. As a consequence, its overall architecture is very similar to the one just sketched for LiLa. This proves to be an important advantage in our effort to link a multilingual resource to the Knowledge Bases of both projects: the strategy that we adopt for linking to Italian is entirely parallel to the one just described for Latin, thus guaranteeing a high degree of uniformity in the treatment of lexical entries from different languages in that respect.

Table 1 gives statistics on the number and percentage of cases of single matches, multiple matches, and absence of matches between entries of our resources and lemmas in the Knowledge Bases of LiLa and LiITA.

Generally speaking, there are a fair amount of items that could be unambiguously matched to a single lemma (around 60 % in both languages). For Italian, there is a greater number of items for which no corresponding lemma could be found. This is likely to be motivated by the fact that the LiITA

[19] Note that we do not use frac:Attestation, since it should be used for a precise fragment of text, that however is not normally given in the resource.

[20] This is what motivates the use of Fabio rather than FRBR, as with the latter we would have needed to map works to manifestations through an instance of the class at intermediate level of concreteness, Expression, on which, however, we do not have any information.

	Latin	Italian
unambiguous match	1,130 (63.13 %)	860 (61.92 %)
ambiguous match	536 (29.94 %)	252 (18.14 %)
no available lemma	124 (6.93 %)	277 (19.94 %)

Table 1: Linking of entries in our resource to lemmas in LiLa and LiITA

Lemma Bank has been built mostly on the basis of resources for contemporary Italian, while our resource focuses on Old Italian, thus documenting a different variety displaying different form variants. For Latin, on the other hand, there is a greater number of items for which more than one lemma was available, which is mostly due to the availability of lemmas with the same form but different part of speech (e.g., common nouns, proper nouns and/or adjectives) or morphological properties (e.g., gender or inflection class). In those cases, however, disambiguation can be easily performed automatically, at least whenever we have information on the part of speech of lexical entries in our resource too. Indeed, in Latin, out of the 536 entries for which a match was found with more than one lemma in LiLa, 420 – i.e., almost 80 % – could be automatically disambiguated and assigned to a single lemma with this procedure. For Italian, automatic disambiguation based on part of speech information was only successful for about 20 % of entries with more than one match (53 out of 252), but the number of ambiguous matches was much lower to begin with.

6 Conclusions and Future Work

In this paper, we have described the release of the DynaMorphPro lexicon,[21] that documents loanwords, conversions and class-shifts from Latin to Old Italian, and located it within the LLOD ecosystem. By leveraging established models – such as OntoLex-lemon for lexical information, CIDOC-CRM and OWL-Time for historical and chronological information, FaBIO and BiRO for citations and references – and integrating them with specialised extensions – like Morph for morphology, FrAC for attestations and lemonEty for etymology – we have been able to provide a rich semantic modelling of the data recorded in the resource. Furthermore, the linking to the Knowledge Bases of LiLa for Latin and LiITA for Italian has ensured interoperability with other resources included in there, maximising the reusability of data for other purposes.

An interesting possibility for future work would be to extend such a strategy to entries in other languages – namely, those that are provided as etymons of the entries in the main languages. For instance, many Latin loanwords come from Ancient Greek. As a consequence, several pieces of information are provided for many etymons in that language. Since a project for the creation of a Wikibase for Ancient Greek is currently being undertaken,[22] it would be useful to link etymons to URIs in that project as soon as possible. A similar strategy could also be applied to all other languages for which similar projects will eventually arise.

References

Tim Berners-Lee, James Hendler, and Ora Lassila. 2001. The semantic web. *Scientific american*, 284(5):34–43.

Christian Chiarcos. 2012. POWLA: Modeling Linguistic Corpora in OWL/DL. In *The Semantic Web: Research and Applications*, pages 225–239, Berlin, Heidelberg. Springer Berlin Heidelberg.

Christian Chiarcos, Elena-Simona Apostol, Besim Kabashi, and Ciprian-Octavian Truică. 2022a. Modelling frequency, attestation, and corpus-based information with OntoLex-FrAC. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4018–4027, Gyeongju. International Committee on Computational Linguistics.

Christian Chiarcos and Christian Fäth. 2017. CoNLL-RDF: Linked Corpora Done in an NLP-Friendly Way. In *Language, Data, and Knowledge*, pages 74–88, Cham. Springer International Publishing.

Christian Chiarcos, Katerina Gkirtzou, Fahad Khan, Penny Labropoulou, Marco Passarotti, and Matteo Pellegrini. 2022b. Computational morphology with OntoLex-morph. In *Proceedings of the 8th Workshop on Linked Data in Linguistics within the 13th Language Resources and Evaluation Conference*, pages 78–86, Marseille. European Language Resources Association.

Christian Chiarcos and Maria Sukhareva. 2015. OLiA – Ontologies of Linguistic Annotation. *Semantic Web*, 6(4):379–386.

Philipp Cimiano, Paul Buitelaar, John McCrae, and Michael Sintek. 2011. LexInfo: A declarative model for the lexicon-ontology interface. *Journal of Web Semantics*, 9(1):29–51.

Gerard De Melo. 2015. Lexvo. org: Language-related information for the linguistic linked data cloud. *Semantic Web*, 6(4):393–400.

[21] At `https://lila-erc.eu/data/lexicalResources/DynaMorphPro`, under a CC BY-SA license.

[22] `https://kratylos-grc.wikibase.cloud/`.

Martin Doerr. 2003. The CIDOC Conceptual Reference Module: An Ontological Approach to Semantic Interoperability of Metadata. *AI Magazine*, 24(3):75–92.

Martin Doerr and Dolores Iorizzo. 2008. The dream of a global knowledge network—A new approach. *Journal on Computing and Cultural Heritage*, 1(1):1–23.

Wolfgang U Dressler. 2003. Degrees of grammatical productivity in inflectional morphology. *Italian Journal of Linguistics*, 15:31–62.

Scott Farrar and D Terence Langendoen. 2003. A linguistic ontology for the semantic web. *GLOT international*, 7(3):97–100.

Manuel Fiorelli, Armando Stellato, John P Mccrae, Philipp Cimiano, and Maria Teresa Pazienza. 2015. LIME: the metadata module for OntoLex. In *The Semantic Web. Latest Advances and New Domains*, pages 321–336. Springer.

Theodorus Fransen, Cormac Anderson, Sacha Beniamine, and Marco Passarotti. 2024. The MOLOR Lemma Bank: a New LLOD Resource for Old Irish. In *Proceedings of the 9th Workshop on Linked Data in Linguistics@ LREC-COLING 2024*, pages 37–43.

FRBR. 2009. IFLA Study Group on the Functional Requirements for Bibliographic Records: Functional Requirements for Bibliographic Records. Final report, International Federation of Library Associations and Institutions.

Aldo Gangemi, Sotiris Batsakis, Euripides G.M. Petrakis, Ilias Tachmazidis, and Grigoris Antoniou. 2017. Temporal representation and reasoning in owl 2. *Semantic Web*, 8(6):981–1000.

Francesco Gardani. 2013. *Dynamics of Morphological Productivity: The Evolution of Noun Classes from Latin to Italian*. Brill, Leiden, The Netherlands.

Sebastian Hellmann, Jens Lehmann, Sören Auer, and Martin Brümmer. 2013. Integrating NLP using Linked Data. In *Proceedings of the 12th International Semantic Web Conference*, Sydney.

ISO 21127:2023. 2023. Information and documentation – A reference ontology for the interchange of cultural heritage information. Standard, International Organization for Standardization, Geneva, CH.

Fahad Khan. 2018. Towards the Representation of Etymological Data on the Semantic Web. *Information*, 9(12):304.

Fahad Khan. 2020. Representing temporal information in lexical linked data resources. In *Proceedings of the 7th Workshop on Linked Data in Linguistics (LDL-2020)*, pages 15–22, Marseille. European Language Resources Association.

Ora Lassila and Ralph R. Swick. 1998. Resource Description Framework (RDF) Model and Syntax Specification.

Eleonora Litta and Marco Passarotti. 2019. (When) inflection needs derivation: a word formation lexicon for Latin. In Nigel Holmes, Marijke Ottink, Josine Schrickx, and Maria Selig, editors, *Words and Sounds*, pages 224–239. De Gruyter, Berlin, Boston.

Eleonora Litta, Marco Passarotti, Paolo Brasolin, Giovanni Moretti, Francesco Mambrini, Valerio Basile, Andrea Di Fabio, and Cristina Bosco. 2024. The Lemma Bank of the LiITA Knowledge Base of Interoperable Resources for Italian. In *Proceedings of the Tenth Italian Conference on Computational Linguistics (CLiC-it 2024)*. CEUR Workshop Proceedings.

Robert Mailhammer. 2013. *Lexical and structural etymology: Beyond word histories*. Walter de Gruyter, Berlin.

John McCrae, Guadalupe Aguado-de Cea, Paul Buitelaar, Philipp Cimiano, Thierry Declerck, Asunción Gómez-Pérez, Jorge Gracia, Laura Hollink, Elena Montiel-Ponsoda, and Dennis Spohr. 2012. Interchanging lexical resources on the semantic web. *Language Resources and Evaluation*, 46(4):701–719.

John P McCrae, Julia Bosque-Gil, Jorge Gracia, Paul Buitelaar, and Philipp Cimiano. 2017. The OntoLex-Lemon Model: Development and Applications. In *Electronic lexicography in the 21st century. Proceedings of eLex 2017 conference*, pages 587–597, Brno. Lexical Computing CZ s.r.o.

Alistair Miles, Brian Matthews, Michael Wilson, and Dan Brickley. 2005. SKOS core: simple knowledge organisation for the web. In *Proceedings of the International Conference on Dublin Core and Metadata Applications*. Dublin Core Metadata Initiative.

Marco Passarotti, Francesco Mambrini, Greta Franzini, Flavio Massimiliano Cecchini, Eleonora Litta, Giovanni Moretti, Paolo Ruffolo, and Rachele Sprugnoli. 2020. Interlinking through Lemmas. The Lexical Collection of the LiLa Knowledge Base of Linguistic Resources for Latin. *Studi e Saggi Linguistici*, LVIII(1):177–212.

Matteo Pellegrini, Eleonora Litta, Marco Passarotti, Francesco Mambrini, and Giovanni Moretti. 2021. The Two Approaches to Word Formation in the LiLa Knowledge Base of Latin Resources. In *Proceedings of the Third International Workshop on Resources and Tools for Derivational Morphology (DeriMo 2021)*, pages 101–109, Nancy, France. ATILF.

Silvio Peroni and David Shotton. 2018. The spar ontologies. In *The Semantic Web–ISWC 2018: 17th International Semantic Web Conference, Monterey, CA, USA, October 8–12, 2018, Proceedings, Part II 17*, pages 119–136. Springer.

E. Prud'Hommeaux and A. Seaborne. 2008. SPARQL query language for RDF. W3C Recommendation.

Mark D. Wilkinson et al. 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3(160018).

A Appendix

We expand here all the prefixes that appear in the CURIEs used in the text and figures of this paper.

```
: http://lila-erc.eu/data/lexicalResources/DynaMorphPro/
dmp: http://lila-erc.eu/ontologies/lila/DynaMorphPro/
biro: http://purl.org/spar/biro/
crm: http://www.cidoc-crm.org/cidoc-crm/
dcterms: http://purl.org/dc/terms/
fabio: http://purl.org/spar/fabio/
frac: http://www.w3.org/nl/lemon/frac#
lemonEty: http://lari-datasets.ilc.cnr.it/lemonEty#
lexinfo: http://www.lexinfo.net/ontology/3.0/lexinfo#
lila: http://lila-erc.eu/ontologies/lila/
lime: http://www.w3.org/ns/lemon/lime#
morph: http://www.w3.org/ns/lemon/morph#
ontolex: http://www.w3.org/ns/lemon/ontolex#
owl-time: http://www.w3.org/2006/time#
rdf: http://www.w3.org/1999/02/22-rdf-syntax-ns#
rdfs: http://www.w3.org/2000/01/rdf-schema#
vartrans: http://www.w3.org/ns/lemon/vartrans#
```