# Linking CompL-it to the LiITA Knowledge Base

Eleonora Litta[1], Marco Passarotti[1], Giovanni Moretti[1], Paolo Brasolin[1], Francesco Mambrini[1], Valerio Basile[2], Andrea Di Fabio[2], Eliana Di Palma[2], Emiliano Giovannetti[3,*], Simone Marchi[3], Andrea Bellandi[3] and Flavia Sciolette[3]

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

[2] *Università di Torino, Via Verdi 8, 10124 Torino, Italia*

[3] *Cnr-Istituto di Linguistica Computazionale "A. Zampolli", Via G. Moruzzi 1, 56124 Pisa, Italia*

**Abstract**

This paper presents the integration of CompL-it, a Linked Open Data (LOD) computational lexicon for contemporary Italian, into LiITA (Linking Italian), a Knowledge Base (KB) designed for linguistic interoperability. CompL-it contains over 101k lexical entries enriched with detailed morphological and semantic information, derived from multiple authoritative sources and modelled using the OntoLex-Lemon vocabulary. The linking process involved aligning lexical entries with lemmas in the LiITA's Lemma Bank (LB), addressing both exact and ambiguous matches through systematic and semantically informed strategies. Moreover, 12,739 new lemmas were added to the LiITA LB. This integration enhances the expressiveness and interoperability of LiITA, enabling complex SPARQL queries that exploit the semantic network encoded in CompL-it. Examples are provided to demonstrate the advantages of querying interlinked resources.

**Keywords**

Linked Open Data, Italian, language resources

## 1. Introduction

During the past two decades, the landscape of digital linguistic resources has experienced exponential growth. Among the many languages benefiting from this expansion, Italian has emerged as a particularly well-resourced language in terms of both lexical and textual resources. These range from semantic lexicons, such as ItalWordNet [1], to treebanks in the Universal Dependencies initiative,[1] as well as diachronic and synchronic corpora like TLIO-OVI,[2] Midia,[3] and CORIS/CODIS [2]. Such diversity and depth of resources position Italian as a highly promising candidate for advanced linguistic research and computational applications. Nevertheless, the abundance of linguistic data presents a double-edged sword. Although the sheer volume of resources is an asset, their heterogeneity in structure, encoding formats, and annotation schemes often impedes their effective integration. Different projects employ different lemmatisation practices, tagsets, and annotations at different granularity levels. This inconsistency leads to significant challenges in interoperability, preventing researchers from leveraging the full empirical potential of the available datasets. Without harmonisation, the possibility of conducting federated searches, comparative analyses, or constructing large-scale linguistic knowledge graphs remains limited.

In response to these challenges, the linguistic data community has coalesced around the principles of Linked Open Data (LOD) [3] and the broader paradigm of the Semantic Web. Driven by initiatives such as the recently concluded COST Action Nexus Linguarum,[4] scholars have collaborated to create and promote shared vocabularies, ontologies, and modelling practices for the publication of interoperable linguistic resources. These developments have been instrumental in establishing a foundation for representing linguistic knowledge in ways that are both machine-readable and semantically robust.

A pioneering example of applying LOD principles to linguistic data is the LiLa (Linking Latin) Knowledge Base,[5] a project designed to interlink Latin lexical and textual resources through a shared, lemma-centred ar-

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

[*] Corresponding author.

[†] These authors contributed equally.

eleonoramaria.litta@unicatt.it (E. Litta); marco.passarotti@unicatt.it (M. Passarotti); giovanni.moretti@unicatt.it (G. Moretti); paolo.brasolin@unicatt.it (P. Brasolin); francesco.mambrini@unicatt.it (F. Mambrini); valerio.basile@unito.it (V. Basile); andrea.difabio@unito.it (A. D. Fabio); eliana.dipalma@unito.it (E. D. Palma); emiliano.giovannetti@ilc.cnr.it (E. Giovannetti); simone.marchi@ilc.cnr.it (S. Marchi); andrea.bellandi@ilc.cnr.it (A. Bellandi); flavia.sciolette@ilc.cnr.it (F. Sciolette)

0000-0002-0499-997X (E. Litta); 0000-0002-9806-7187 (M. Passarotti); 0000-0001-7188-8172 (G. Moretti); 0000-0003-2471-7797 (P. Brasolin); 0000-0003-0834-7562 (F. Mambrini); 0000-0001-8110-6832 (V. Basile); 0000-0002-3290-8158 (A. D. Fabio); 0000-0003-2154-2696 (E. D. Palma); 0000-0002-0716-1160 (E. Giovannetti); 0000-0003-4320-6466 (S. Marchi); 0000-0002-1900-5616 (A. Bellandi); 0000-0002-7998-9768 (F. Sciolette)

[1] https://universaldependencies.org/

[2] http://www.ovi.cnr.it/en/Il-Corpus-Testuale.html

[3] https://www.corpusmidia.unito.it/

[4] https://nexuslinguarum.eu

[5] https://lila-erc.eu

chitecture, by following the LOD principles. In LiLa, lemmas act as pivots between textual data (composed by tokenised texts) and lexical metadata (compiled by lexical entries). Lemmas are collected in a Lemma Bank (LB) to serve as the nexus for integrating distributed linguistic resources and enabling seamless connections across heterogeneous datasets [4]. This architecture has not only proven effective in unifying Latin resources, but has also demonstrated its adaptability to other languages. Building upon the LiLa framework, the LiITA (Linking Italian) Knowledge Base has been conceived as a Knowledge Base for Italian linguistic resources[5]. LiITA inherits the lemma-centric design, constructing a LB for Italian. This LB, initially comprising over 113,000 entries extracted from the Nuovo De Mauro dictionary,[6] is meticulously curated to support interoperability, particularly in the context of divergent lemmatisation standards. By modelling each lemma using the OntoLex-Lemon vocabulary and a shared ontology derived from LiLa, LiITA ensures that lexical entries and their associated textual occurrences can be connected across otherwise incompatible datasets. Its architecture not only allows for the integration of existing datasets but also accommodates the dynamic evolution of linguistic knowledge as new resources become available in the KB, in an ever-growing fashion.

As part of its ongoing development, LiITA is currently in the process of interlinking via its LB several key lexical and textual resources. These include the *Vocabolario della Lingua Parmigiana* glossary, a bilingual lexicon having Italian entries and the corresponding translations in Parmigiano,[7] and CompL-it[6], a computational lexicon for Italian already published as Linked Open Data. This paper describes the process of linking the computational lexicon CompL-it to LiITA and it is structured as follows: Section 2 contains a short description of the LiITA architecture, section 3 contains a description of the CompL-it resource and of how it is modelled in RDF; Section 4 describes the process of linking to the LiITA KB and how the LiITA LB has been enriched by the addition of new lemmas from CompL-it; Section 5 contains examples of the advantages given by the linking of the CompL-it resource to LiITA, including an example of a SPARQL queries performed on the current KB; Section 6 draws conclusions and outlines future perspectives and developments.

## 2. LiITA - Architecture

In the LiITA LB, lemmas are represented using a dedicated ontology,[8] inherited from LiLa, which was specifically developed to capture the morphological and linguistic characteristics of Latin. This ontology encodes features such as Part-of-Speech (PoS), gender, and inflectional properties, drawing on the OLiA annotation framework [7, 151–155] to ensure consistency and formal interoperability.

The ontology also defines the essential Classes and Properties required for modelling lemmatisation. Among these is the Property `lila:hasLemma`,[9] which associates lemmas with the tokens they annotate within a corpus.

Within the OntoLex-Lemon model [8], lexical forms can have one or more graphical variants, captured using the Property `ontolex:writtenRep` (http://www.w3.org/ns/lemon/ontolex#writtenRep), as well as phonetic realisations, specified by the Property `ontolex:phoneticRep` (http://www.w3.org/ns/lemon/ontolex#phoneticRep). The Property `ontolex:canonicalForm` (http://www.w3.org/ns/lemon/ontolex#canonicalForm) identifies the standard or representative form within an inflectional paradigm.

The LiITA LB is composed of such canonical forms, which are represented as instances of the Class `lila:Lemma`,[10] a subclass of `ontolex:Form` within the OntoLex-Lemon ontology. Moreover, the class `lila:Hypolemma`, a subclass of `lila:Lemma`, is used to represent citation forms that belong to a word's regular inflectional paradigm but receive a different PoS tag than the lemma. It is the case of participles such as *amato* 'loved', adjective, which is part of the inflectional paradigm of *amare*, 'to love', verb.

With respect to morphological annotation, each lemma in the LB is assigned a Part-of-Speech label using the Property `lila:hasPos`,[11] in accordance with the UPOS (Universal POS) tag set [9].

The LiITA LB is not made of lexical entries because it does not function as an autonomous lexical resource. Rather, it constitutes a curated repository of canonical forms that (i) is intended to grow progressively as new sources, including those containing previously unrecorded lemmas, are integrated, and (ii) serves as a foundation for both text lemmatisation and the indexing of lexical entries within distributed resources published as LOD.

However, linguistic resources often adopt heterogeneous tag sets, standards, and annotation schemes, particularly with respect to lemmatisation.

[6] https://dizionario.internazionale.it/
[7] https://github.com/LiITA-LOD/LocalVarieties/tree/main/Parmigiano
[8] http://lila-erc.eu/ontologies/lila/
[9] http://lila-erc.eu/ontologies/lila/hasLemma
[10] http://lila-erc.eu/ontologies/lila/Lemma
[11] http://lila-erc.eu/ontologies/lila/hasPOS

To accommodate this variation in lemmatisation approaches found across linguistic resources, the LiITA LB defines two specialised Properties. The first is the symmetric Property `lila:lemmaVariant`,[12] which links different forms within the same inflectional paradigm that may be used as lemmas, while maintaining their associated PoS. A common case involves *pluralia tantum*, which can appear as either singular or plural lemmas. For example, both the plural *occhiali* and the singular *occhiale* ('glasses/optical instrument') are represented as distinct `lila:Lemma` , connected via the `lila:lemmaVariant` Property.

In contrast, the Property `lila:hasHypolemma`,[13] along with its inverse relation `lila:isHypolemma`,[14] is used to relate a `lila:Lemma` to a `lila:Hypolemma`.

By means of this modelling framework, the LB provides a coherent structure capable of accommodating divergent lemmatisation practices. For example, some resources lemmatise participles under their participial form, while others prefer the base verbal form. Thanks to this flexible architecture, such differences can be reconciled, thereby promoting interoperability across corpora and lexical resources employing distinct lemmatisation conventions.

## 3. CompL-it

CompL-it is a computational lexicon for contemporary Italian, modelled according to the already cited OntoLex-Lemon model, the *de facto* standard for lexical resources and compliant with the principles of LOD. This resource was created by merging three different sources of data: M-GLF (MAGIC-Generated Lemmatized Forms), a list of lemmatised forms with morphological information generated by the MAGIC tool, a morphological analyser [10] [11]; a set of Italian language treebanks available through the UD repository (Italian Stanford Dependency Treebank, ISDT[15]; Venice Italian Treebank, VIT[16]; ParallelTut, ParTut[17]; ParlaMint-It[18]); the computational lexicon LexicO [12], which constitutes the base of the entire resource, from the point of view of the model.

LexicO represents the revised version of another important resource in the framework of Italian Lexicography, Parole-Simple-Clips [13], with which it shares the same model based on the theory of Generative Lexicon by James Pustejovsky [14], with four different layers of linguistic information (morphological, semantic, syntactic and phonological). The lemmas of the resources have been converted as Lexical Entries of the OntoLex-Lemon model and the forms as Lexical Forms; regarding the PoS and the morphological traits (e.g. gender, number), each of the three resources had a different vocabulary for describing them. Therefore, they were mapped and converted according to the LexInfo vocabulary, the main linguistic ontology for OntoLex-Lemon model.

The strength of CompL-it, however, is the semantic layer, partly converted from LexicO; it is worth noting that the senses in CompL-it (derived from LexicO, since there are no senses in either M-GLF or treebanks) are richly described through a vocabulary consisting of 137 relations, divided in eight classes. Where possible, some relations have been mapped to LexInfo[19], otherwise, custom object properties were created. The conversion of the data thus prepared, coming from the three sources into OntoLex-Lemon, was performed by an algorithm in two steps: i) conversion of the linguistic information according to the formalisation described in the core `ontolex` module of the model; ii) serialisation of the data into Turtle. The obtained lexicon was then loaded into Ontotext GraphDB[20], a semantic repository compliant with RDF and SPARQL[21].

The following is an example of an RDF OntoLex-Lemon representation of a CompL-it lexical entry in Turtle format.

```
:coniglio_entry a ontolex:Word;
lexinfo:partOfSpeech lexinfo:noun;
ontolex:canonicalForm coniglio_lemma;
ontolex:otherForm coniglio_form_1;
ontolex:sense coniglio_sense_1, coniglio_sense_2,
     coniglio_sense_3 .

coniglio_lemma a ontolex:Form;
lexinfo:gender lexinfo:masculine;
lexinfo:number lexinfo:singular;
ontolex:writtenRep "coniglio"@it, "rabbit"@en .

coniglio_form_1 a ontolex:Form;
lexinfo:gender lexinfo:masculine;
lexinfo:number lexinfo:plural;
ontolex:writtenRep "conigli"@it, "rabbits"@en .

coniglio_sense_1 a ontolex:LexicalSense;
skos:definition "mammifero della famiglia dei
     Leporidi, con pelame di vario colore, lunghe
     orecchie, occhi grandi e sporgenti e grossi
     incisivi"@it , "Mammal of the Leporidae family,
     with variously colored fur, long ears, large,
     protruding eyes and large incisors"@en;
lexinfo:hyponym mammifero_sense;
simple:polysemyAnimalFood coniglio_sense_3 .

coniglio_sense_2 a ontolex:LexicalSense;
```

[12] http://lila-erc.eu/ontologies/lila/lemmaVariant
[13] http://lila-erc.eu/ontologies/lila/hasHypolemma
[14] http://lila-erc.eu/ontologies/lila/isHypolemma
[15] https://github.com/UniversalDependencies/UD_Italian-ISDT
[16] https://github.com/UniversalDependencies/UD_Italian-VIT
[17] https://github.com/UniversalDependencies/UD_Italian-ParTUT
[18] https://github.com/UniversalDependencies/UD_Italian-ParlaMint
[19] https://lexinfo.net/
[20] https://www.ontotext.com/products/graphdb/
[21] https://www.w3.org/TR/sparql11-overview/

```
skos:definition "persona timida e molto paurosa"@it,
     "shy and very fearful person"@en;
lexinfo:hyponym persona_sense;
simple:metaphor coniglio_sense_1 .

coniglio_sense_3 a ontolex:LexicalSense;
skos:definition "carne dell'omonimo animale"@it,
     "meat of the animal"@en .
```

In this example, the lexical entry *coniglio* (rabbit) is linked to two word forms: one designated as the canonical form (lemma), and the other corresponding to the plural form *conigli* (rabbits). Both forms are annotated with the appropriate morphological features.

The lexical entry is also connected, via the `ontolex:sense` property, which links lexical entries to their semantic interpretations, to three lexical senses, each of which includes a definition expressed in natural language.

Furthermore, the first two senses are semantically enriched through relations that connect them to other lexical senses in the resource. For instance, *rabbit_sense_2* is modelled as a hyponym of *mammal_sense*.

CompL-it contains 101,795 lexical entries (comprising a total of 791,541 word forms), classified with 36 PoS categories and described with morphological traits; from a semantic standpoint, CompL-it describes 55,713 word senses connected to each other through 137 types of semantic relations, totaling 86,577 instances.

Table 1 shows a distribution of the 10 most numerous types of semantic relation instances:

| Semantic relation | # instances | an example |
|---|---|---|
| hyponym | 43,069 | medicina, scienza (medicine, science) |
| approximateSynonym | 5,666 | sciocco, stupido (foolish, stupid) |
| usedFor | 3,291 | matita, scrivere (pencil, to write) |
| partMeronym | 3,159 | giorno, settimana (day, week) |
| partHolonym | 3,159 | cinghiale, grugno (boar, snout) |
| createdBy | 2,857 | quadro, dipingere (painting, to paint) |
| ObjectOfTheActivity | 1,366 | bistecca, mangiare (steak, to eat) |
| memberMeronym | 1,318 | segretario, partito (secretary, party) |
| ResultingState | 1,063 | bruciare, bruciato (to burn, burnt) |
| memberHolonym | 979 | stormo, uccello (flock, bird) |
| other | 20,650 | - |
| total | 86,577 | |

**Table 1**
Distribution of semantic relations instances

# 4. Linking

Linking a lexical resource to the LiITA LB entails establishing a relationship between the lexical entries of the resource and the lemmas in the LB. Typically, this process begins with modeling the resource as a LOD resource, followed by creating the connections between the resource's entries and the LB lemmas. Modelling the link between CompL-it and LiITA was, however, relatively straightforward. One of the main advantages of integrating a resource that already adheres to LOD standards is that each CompL-it entry, already represented as an `ontolex:Word`, a subclass of `ontolex:LexicalEntry`, can be directly linked to LiITA via the `ontolex:canonicalForm` relation.

The linking process between CompL-it and LiITA begins necessarily with a mapping between the different PoS tags used in CompL-it, which are described using Lexinfo, and the UPOS tagset used in LiITA. Table 2 shows the PoS mapping between the two tagsets operated on the data before matching CompL-it entries with LiITA lemmas.

Subsequently a match between CompL-it lexical entries and lemmas in LiITA was performed on the lemma-PoS pair. Out of over 101k lexical entries in CompL-it, the matching process yielded the following results:

- **1:1 match**: 83,340 lexical entries (an exact match between a CompL-it lexical entry and a LiITA lemma + PoS combination)
- **1:N match**: 4,219 lexical entries (more than one potential lemma-POS pairs in LiITA corresponding to a single CompL-it lexical entry)
- **1:0 match**: 14,314 lexical entries (no corresponding lemma-POS pair found in LiITA)

The linking is operationalised using the `ontolex:canonicalForm` relation, which connects a CompL-it lexical entry to a corresponding lemma in LiITA. For example:

> `http://lexica/mylexicon#MUSmerendaNOUN`
> `ontolex:canonicalForm`
> `http://liita.it/data/id/lemma/1010136`
> (merenda)

Disambiguation of 1:N matches posed a significant challenge. At the time of this initial linking effort, CompL-it was the first external resource to be linked to the LiITA LB, meaning that no additional semantic cues, such as sense distinctions or contextual usage, were yet available in the lemma database. As a result, each lemma in LiITA was limited to grammatical information such as PoS, gender, or conjugation and reflexivity (for verbs). Although, as noted in Section 1, the lemmas were extracted from

| Lexinfo | UPOS |
|---|---|
| adjective | ADJ |
| adposition | ADP |
| adverb | ADV |
| article | DET |
| auxiliary | VERB |
| cardinalNumeral | NUM |
| commonNoun | NOUN |
| conjunction | SCONJ-ADV |
| coordinatingConjunction | CCONJ |
| definiteArticle | DET |
| demonstrativeDeterminer | DET |
| demonstrativePronoun | PRON |
| determiner | DET |
| exclamativeDeterminer | DET |
| exclamativePronoun | PRON |
| fusedPreposition | ADP |
| indefiniteArticle | DET |
| indefiniteDeterminer | DET |
| indefinitePronoun | PRON |
| interjection | INTJ |
| interrogativeAdverb | ADV |
| interrogativeDeterminer | DET |
| interrogativePronoun | PRON |
| noun | NOUN |
| numeral | NUM |
| numeralDeterminer | DET |
| numeralPronoun | PRON |
| particle | PART |
| personalPronoun | PRON |
| possessiveAdjective | ADJ |
| possessiveDeterminer | DET |
| possessivePronoun | PRON |
| pronoun | PRON |
| relativeDeterminer | DET |
| relativePronoun | PRON |
| subordinatingConjunction | SCONJ |
| verb | VERB |

**Table 2**
PoS mapping between LexInfo and UPOS tags

the *Nuovo De Mauro* Dictionary, no sense-level metadata was incorporated from the dictionary.

In the absence of semantic information, we adopted a pragmatic yet arbitrary strategy for disambiguation: where multiple LiITA lemmas shared the same form and PoS, we selected the lemma that appears first in the LB (by id). While this approach lacks empirical grounding, it provided a consistent criterion for initiating the alignment process.

In cases involving a 1:0 match, the correspondence with the string may be either complete—for instance, in the case of a previously unseen word—or partial, as when inflected forms of lemmas already present in the LB are encountered. The strategy for inclusion varies according to the characteristics of the lexical resource being linked.

The CompL-it resource contains a substantial number of words in plural form. Entries such as *pantaloni* ("trousers") and *mutande* ("underpants"), *braccia*, *ottavi*, which refers to the "round of 16" in a tournament setting, have been added to the LB. In such cases the new lemma has been linked to their singular variant in the LB with the Property `lila:lemmaVariant` as described in Section 2.

A few additional noteworthy inclusion strategies from the CompL-it resource that have been adopted are outlined below:

- **Truncated word forms**, such as *quest'*, *nessun'*, and *verun*, have been added as written representations of existing lemmas.
- **Adjectives and determiners** occurring in feminine or plural forms have been systematically linked to their corresponding singular masculine lemmas in LiITA.
- **Adverbial forms** that appear to be derived from adjectives, pronouns, or determiners (e.g., *quante*, *prese*) have been included in the resource as hypolemmas of their corresponding base entries. This modelling choice ensures compatibility with texts in which such adverbial forms are lemmatised under their base categories—namely, adjectives, pronouns, or determiners—thereby promoting consistency across heterogeneous lemmatisation practices.
- **Composite pronouns**, such as *glieli*, *glielo*, *gliene*, and others, have also been included in the LB, following the same rationale outlined above. This ensures alignment with sources in which these forms are treated as distinct lemmas (as opposed to split into e.g. *glielo*   *gli* + *lo*)
- **Orthographic errors** (e.g., *perchè*, with grave accent on the final *e*, instead of the correct *perché*) have been linked to the appropriate lemma, although their incorrect spellings have not been recorded as alternative written representations.

## 5. Querying CompL-it in LiITA

One of the key advantages of storing data in RDF is the ability to formulate federated SPARQL queries that retrieve information from datasets distributed across multiple endpoints. Examples of SPARQL queries performed on the LiITA Knowledge Base are continuously added to https://www.liita.it/?page_id=158. The integration of CompL-it into the LiITA Knowledge Base enables the exploitation of its rich semantic network and facilitates interoperability with other linked linguistic resources. For instance, it becomes possible to retrieve Italian lexical entries linked to CompL-it whose definitions begin with *uccello* (bird) and to display their corresponding translations in the Parmigiano Glossary, another resource linked to LiITA.[22] It is interesting to explore the added value that CompL-it contributes through its dense network of semantic relations. For instance, one of the example queries provided on the LiITA website retrieves lexical entries associated with color by filtering definitions that begin with the string *colore* (“colour”). While this method yields relevant results, a more semantically informed strategy involves querying for all hyponyms of the specific sense of the lemma *colore* defined as "qualità dei corpi per cui essi riflettono in vario modo la luce" (“property of bodies by which they reflect light in various ways”). Below is the SPARQL query text retrieving all the hyponyms of *colore*.

```
PREFIX lime: <http://www.w3.org/ns/lemon/lime#>
PREFIX vartrans: <http://www.w3.org/ns/lemon/
     vartrans#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-
     schema#>
PREFIX skos: <http://www.w3.org/2004/02/skos/core
     #>
PREFIX dct: <http://purl.org/dc/terms/>
PREFIX onto: <http://www.ontotext.com/>
PREFIX lexinfo: <http://www.lexinfo.net/ontology
     /3.0/lexinfo#>
PREFIX ontolex: <http://www.w3.org/ns/lemon/
     ontolex#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-
     syntax-ns#>

SELECT ?senseHyponym
       (GROUP_CONCAT(str(?_definition);SEPARATOR="
     ; esempio: ") AS ?definition)
       ?liitaLemma ?parmigianoLemma ?wr
WHERE {
  SERVICE <https://klab.ilc.cnr.it/graphdb-compl-
     it/> {
    ?word a ontolex:Word ;
          lexinfo:partOfSpeech [ rdfs:label ?pos ]
      ;
          ontolex:sense ?sense ;
          ontolex:canonicalForm [ ontolex:
     writtenRep ?lemma ] .
    ?sense lexinfo:hypernym ?senseHyponym .
    OPTIONAL { ?senseHyponym skos:definition ?
     _definition } .
    FILTER(str(?pos) = "noun") .
    FILTER(str(?lemma) = "colore") .
    ?wordHyponym ontolex:sense ?senseHyponym .
  }
  ?wordHyponym ontolex:canonicalForm ?liitaLemma .
  ?leItaLexiconPar ontolex:canonicalForm ?
     liitaLemma ;
                   ^lime:entry <http://liita.it/
     data/LexicalReources/DialettoParmigiano/
     Lexicon> .
  ?leItaLexiconPar vartrans:translatableAs ?
     leParLexiconPar .
  ?leParLexiconPar ontolex:canonicalForm ?
     parmigianoLemma .
  ?parmigianoLemma ontolex:writtenRep ?wr
}
GROUP BY ?senseHyponym ?liitaLemma ?
     parmigianoLemma ?wr
ORDER BY ASC(?wr)
```

The query interrogates the CompL-it repository hosted in GraphDB to extract lexical entries classified as nouns, whose written representation is *colore* and which are associated with a sense that has at least one hyponym. Additionally, it retrieves all the available definitions of such hyponyms. Subsequently, the query accesses the local LiITA graph to extract the Italian written representation of each hyponym, identify the corresponding lexical entry, verify its inclusion in the Parmigiano lexicon, and retrieve its translation along with the written representation in dialect. The final output includes the hyponymic senses, their definitions (if available), the Italian canonical forms, their written representations, and the corresponding lemma in the Parmigiano resource. A selection of the results is shown in Table 3, including the written representations of the Italian and corresponding Parmigiano lemmas.

| **italian** | **parm.** | **italian** | **parm.** |
|---|---|---|---|
| argento | argént | tabacco | pisighén |
| azzurro | azúr | piombo | piómb |
| grigio | bergnôl | mattone | quaderlètt |
| grigio | biz | mattone | quaderlón |
| grigio | bizón | mattone | quadrél |
| blu | blò | rame | ram |
| cenere | bornìza | pisello | reviót |
| bronzo | brónz | rosso | ròss |
| prugna | brùggna | ruggine | rùzzna |
| caramella | caraméla | topo | sorghén |
| carminio | carmzén | topo | sorgón |
| carota | caròtla | ciliegia | sréza |
| crema | crèmma | sabbia | sàbia |
| cremisi | crèmmez | cenere | sèndra |
| ferro | fér | topo | sòrrogh |
| giallo | gialdètt | tabacco | tabach |
| giallo | gialdón | topo | topén |
| giallo | giäld | verde | verdzén |
| grigio | griz | verde | verdén |
| limone | limón | verdone | verdón |
| muschio | musc’ | violetto | violètt |
| miele | méla | ciliegia | vìssola |
| nocciola | nisôla | giallo | zaldón |
| paglia | paja | oro | òr |

**Table 3**
An excerpt of the results from the query on hyponyms of *colore*, showing the correspondences between Italian and Parmigiano lemmas.

[22] https://liita.it/data/id/DialettoParmigiano/lemma/LemmaBank.html

This sense-centred approach results in approximately thirty additional lexical entries, as many of the corresponding definitions do not explicitly include the word *colore*, but are nonetheless semantically linked through hyponymy. This example highlights the potential of leveraging CompL-it's semantic network to formulate richer and more accurate queries.

## 6. Conclusions

The integration of CompL-it into the LiITA Knowledge Base marks a significant milestone in the development of interoperable linguistic resources for Italian. By linking over 100,000 lexical entries, many of which include rich semantic annotations, to LiITA's LB, this initiative enhances the interoperability and expressiveness of both resources. The linking process also prompted the creation of new lemma variants, refinement of linking strategies, and the accommodation of plural forms and multiword expressions, thereby contributing to the ongoing enrichment of the LB. This work demonstrates the feasibility and advantages of integrating heterogeneous linguistic resources using Linked Open Data principles and shared ontologies. The ability to execute cross-resource SPARQL queries further exemplifies the practical benefits of semantic interoperability. One of the next crucial steps will be the integration of Italian textual corpora into LiITA. This will allow not only for the validation of lemma-token alignment but also for exploring contextual usage patterns of lexical entries. Moreover, this will allow for the semantic richness of CompL-it to be exploited through designing and testing of more complex SPARQL queries. Lastly, one of the key challenges in achieving impact within the linguistic community, or more broadly, the humanities fields that engage with data, will be to evaluate and explore text-to-SPARQL systems using Large Language Models (LLMs). This can be done through Retrieval-Augmented Generation (RAG), where a set of SPARQL queries over the LIITA KB is provided, and various few-shot prompts are tested to equip the LLM with knowledge about the Classes and Properties used in the KB.

## Acknowledgments

This contribution is funded by the European Union - Next Generation EU, Mission 4 Component 1 CUP J53D23017270OO1. The PRIN 2022 PNRR project

**LiITA: Interlinking Linguistic Resources for Italian via Linked Data** is carried out jointly by the Università Cattolica del Sacro Cuore, Milano and the Università di Torino.

## References

[1] A. Roventini, R. Marinelli, F. Bertagna, ItalWordNet v.2, 2016. URL: http://hdl.handle.net/20.500.11752/ILC-62, ILC-CNR for CLARIN-IT repository hosted at Institute for Computational Linguistics "A. Zampolli", National Research Council, in Pisa.

[2] R. R. Favretti, F. Tamburini, C. De Santis, Coris/-codis: A corpus of written italian based on a defined and a dynamic model, A rainbow of corpora: Corpus linguistics and the languages of the world (2002) 27–38.

[3] C. Chiarcos, POWLA: Modeling linguistic corpora in OWL/DL, in: C. P. P. A. C. O. P. V. Simperl, E. (Ed.), The Semantic Web: Research and Applications. ESWC 2012, volume 7295 of *Lecture Notes in Computer Science*, Springer, Berlin, Heidelberg, 2012, pp. 225–239. doi:10.1007/978-3-642-30284-8_22.

[4] M. Passarotti, F. Mambrini, G. Franzini, F. M. Cecchini, E. Litta, G. Moretti, P. Ruffolo, R. Sprugnoli, Interlinking through lemmas. the lexical collection of the lila knowledge base of linguistic resources for latin, Studi e Saggi Linguistici 58 (2020) 177–212.

[5] E. M. G. Litta Modignani Picozzi, M. C. Passarotti, P. Brasolin, G. Moretti1, F. Mambrini, V. Basile, A. D. Fabio, C. Bosco, The Lemma Bank of the LiITA Knowledge Base of Interoperable Resources for Italian, ITA, 2024. URL: https://publicatt.unicatt.it/handle/10807/299843, accepted: 2024-12-04T14:12:09Z.

[6] F. Sciolette, A. Bellandi, E. Giovannetti, S. Marchi, CompL-it: a Computational Lexicon of Italian, AIDAinformazioni 42 (2024) 119–148. URL: https://doi.org/10.57574/596545646. doi:10.57574/596545646.

[7] P. Cimiano, C. Chiarcos, J. P. McCrae, J. Gracia, Linguistic Linked Data: Representation, Generation and Applications, Springer, Cham, 2020. URL: https://www.springer.com/gp/book/9783030302245. doi:10.1007/978-3-030-30225-2.

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

[9] S. Petrov, D. Das, R. McDonald, A Universal Part-of-Speech Tagset, in: N. C. C. Chair), K. Choukri, T. Declerck, M. U. Doğan, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, S. Piperidis (Eds.), Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12), European Language Resources Association (ELRA), Istanbul, Turkey, 2012, pp. 2089–

2096. URL: http://www.lrec-conf.org/proceedings/lrec2012/pdf/274_Paper.pdf.

[10] M. Battista, V. Pirrelli, Una piattaforma di morfologia computazionale per l'analisi e la generazione delle parole italiane, Technical Report, 1999.

[11] V. Pirrelli, M. Battista, The paradigmatic dimension of stem allomorphy in Italian verb inflection, Italian Journal of Linguistics 12 (2000) 307–380.

[12] F. Sciolette, E. Giovannetti, S. Marchi, LexicO: an Italian Computational Lexicon derived from Parole-Simple-Clips, Umanistica Digitale 7 (2023) 169–193. URL: https://umanisticadigitale.unibo.it/article/view/15176. doi:10.6092/issn.2532-8816/15176.

[13] AA.VV., PAROLE-SIMPLE-CLIPS, 2016. URL: http://hdl.handle.net/20.500.11752/ILC-88.

[14] J. Pustejovsky, The Generative Lexicon, The MIT Press, 1995. URL: https://direct.mit.edu/books/book/4726/The-Generative-Lexicon. doi:10.7551/mitpress/3225.001.0001.

## Declaration on Generative AI

During the preparation of this work, the author(s) used ChatGPT (OpenAI) and Gemini (Google) in order to: Paraphrase and reword, Improve writing style, and Grammar and spelling check. After using these tool(s)/service(s), the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.