LiITA: a Knowledge Base of Interoperable Resources for Italian

Eleonora Litta and Marco Passarotti and Paolo Brasolin
and Giovanni Moretti and Francesco Mambrini
CIRCSE Research Centre
Università Cattolica del Sacro Cuore
Largo Gemelli 1, 20123 Milan, Italy
eleonoramaria.litta|marco.passarotti|paolo.brasolin|
giovanni.moretti|francesco.mambrini@unicatt.it
and **Valerio Basile and Andrea Di Fabio and Cristina Bosco and Eliana Di Palma**
Università di Torino
Via Verdi, 8 - 10124 Torino
valerio.basile|andrea.difabio|cristina.bosco|eliana.dipalma@unito.it

Abstract

This paper describes the LiITA Knowledge Base of interoperable linguistic resources for Italian. By adhering to the Linked Open Data principles, LiITA ensures and facilitates interoperability between distributed resources. The paper outlines the lemma-centered architecture of the Knowledge Base and details its core component: the Lemma Bank, a collection of Italian lemmas designed to interlink distributed lexical and textual resources.

1 Introduction

In terms of the quantity of digital linguistic resources—both lexical and textual—Italian is among the most well-represented languages, and can be considered a highly resourced language. The CLARIN Virtual Language Observatory, a search engine powered by linguistic resource repositories,[1] currently lists over 8,000 resources dedicated to the Italian language. Among these, there is a substantial set of essential and widely used resources, including ItalWordNet v.2 (Roventini et al., 2016), ten treebanks available from the Universal Dependencies collection[2], historical corpora, like Midia[3] and TLIO-OVI[4], and reference corpora for both written (e.g., CORIS/CODIS (Favretti et al., 2002)) and spoken language (e.g., KIParla (Mauri et al., 2019)).

Unfortunately, the many resources for Italian display considerable variation in encoding methods, data formats, annotation criteria, and tag sets, often presenting information with different levels of granularity. These inconsistencies hinder seamless interaction between the (meta)data provided by different resources, limiting researchers' ability to fully exploit the empirical potential of linguistic data and diminishing the usability of the resources.

As a result, over the past decade, a dynamic scholarly community, centered around the recently concluded COST Action *Nexus Linguarum*[5], has been actively working to establish standardised practices for representing and publishing linguistic resources following the principles of the Linked Data paradigm, which underpins the Semantic Web (Berners-Lee et al., 2001). Several vocabularies for describing linguistic knowledge have emerged from this initiative and have been widely adopted in designing new resources and adapting existing ones. For Italian, some resources are now available as Linked Open Data, including the CompL-it lexicon[6], ItalWordNet v.2[7], and a collection of names from the PAROLE SIMPLE CLIPS (PSC) lexicon[8].

An exemplary application of the Linked Open Data (LOD) principles to the publication of interoperable linguistic resources is the LiLa (Linking Latin) Knowledge Base (KB), which focusses on resources for the Latin language. Building on LiLa as a reference model to achieve online interoperability between distributed linguistic resources—and leveraging its largely language-independent architecture—the LiITA (Linking Italian)[9] project is developing a KB of interoperable resources for Italian, published as Linked Data. This short paper presents the development of the core component of

[1] https://vlo.clarin.eu
[2] https://universaldependencies.org
[3] https://www.corpusmidia.unito.it/
[4] http://www.ovi.cnr.it/en/Il-Corpus-Testuale.html
[5] https://nexuslinguarum.eu
[6] http://hdl.handle.net/20.500.11752/ILC-1007
[7] http://hdl.handle.net/20.500.11752/ILC-66
[8] http://hdl.handle.net/20.500.11752/ILC-558
[9] http://www.liita.it/

the LiITA KB, the Lemma Bank, which is a collection of Italian lemmas published as LOD serving as the linkage point between word occurrences and their corresponding entries in the corpora and lexical resources to be interlinked in the KB.

2 The LiITA Knowledge Base

2.1 Architecture

The architecture of the LiITA Knowledge Base (KB) is inspired by the design of the LiLa KB for Latin[10], based on the key principle that most data and metadata within the resources to be integrated into the KB are fundamentally related to words. Lexical resources, such as dictionaries or lexicons that describe word properties, are organised as lexical entries, and textual resources, including corpora, treebanks, and digital libraries that provide textual content, are composed of word occurrences. In the LiLa LOD architecture, lexical entries and word occurrences from various distributed corpora are made interoperable by linking them to their corresponding lemmas within a collection of conventional citation forms (lemmas). This collection forms the central component of LiLa. LiITA adopts the same lemma-based pivot structure, enabling the integration of diverse resources and supporting federated searches across multiple linguistic datasets.

Similar to LiLa, conceptual interoperability (Ide and Pustejovsky, 2010) among the distributed resources linked within LiITA is achieved through the use of a knowledge description vocabulary based on ontologies widely adopted by the Linguistic LOD community, such as OntoLex[11] for lexical resources, NIF[12], ConLL-RDF (Chiarcos and Fäth, 2017) and Powla (Chiarcos, 2012) for corpus annotation, OLiA[13] for linguistic annotation, DCMT[14] and LIME[15] (Fiorelli et al., 2015) for metadata.

2.2 The Lexical Base of the Lemma Bank

The lemmas included in the initial release of the LiITA Lemma Bank were extracted from an online version of the Nuovo De Mauro dictionary[16], totaling approximately 145,000 entries. Of these, around 13,000 multi-word expressions were excluded as they were considered unnecessary. This decision was based on the fact that the first step in linking a resource is lemmatisation—and since lemmatisers typically work on single tokens, incorporating multi-word expressions into the lexical base would provide minimal practical benefit.

From the remaining 132,000 entries, a total of 129,442 records were generated. In the Lemma Bank, these are divided into 113,112 lemmas and 16,330 hypolemmas. Hypolemmas are inflected forms within the inflectional paradigm of a lemma that commonly appear in lexical resources as canonical citation forms in independent lexical entries. They are assigned a different part of speech (PoS) than their corresponding reference lemma. Common examples of hypolemmas include present and past participles, which are categorised as adjectives and linked to their corresponding verbal lemmas: e.g., *abbagliato* 'dazzled' and *abbagliante* 'dazzling' are linked to *abbagliare* 'to dazzle'. Another example includes adverbs derived from adjectives (the reference lemma) either through conversion (*lento* 'slow' > *lento* 'slowly') or regular suffixation (*lentamente* 'slowly'). Table 1 shows the distribution of hypolemmas across different categories.

Lemmas	Type
10,689	Past Participle
4,544	Adverbs
1,097	Present Participle

Table 1: Distribution of hypolemmas across different categories

Entries from the Nuovo De Mauro were analysed and separated so that each lemma is assigned a single PoS. Additionally, nouns are annotated with their gender, and verbs are categorised by their inflectional class. For instance, the entry *abate*[17] corresponds to two distinct lemmas in LiITA: one as a masculine noun ('abbot') and the other as a feminine noun ('a variety of pear').

The PoS tags used in the Nuovo De Mauro were automatically converted into the Universal PoS tag set (UPOS) (Petrov et al., 2012) to facilitate easier alignment with existing resources. Table 2 shows the distribution lemmas by PoS.

The Nuovo De Mauro PoS tag set was adopted with a number of in-house modifications. Because the original tagging conformed to traditional Italian grammar, certain categories required adjustment.

[10] https://lila-erc.eu/
[11] https://www.w3.org/2016/05/ontolex/
[12] https://persistence.uni-leipzig.org/nlp2rdf/
[13] https://acoli-repo.github.io/olia/
[14] https://www.dublincore.org
[15] https://art.uniroma2.it/lime/
[16] https://dizionario.internazionale.it/
[17] https://dizionario.internazionale.it/parola/abate

Lemmas	Part of Speech
72,073	Nouns
22,449	Adjectives
16,475	Verbs
981	Abbreviations
532	Adverbs
393	Interjections
361	Proper Nouns
136	Pronouns
123	Prepositions
100	Sub. Conjunctions
83	Determiners
67	Coord. Conjunctions
65	Numerals

Table 2: Distribution of lemmas across different parts of speech

Conjunctions, in particular, required specific attention, as De Mauro's scheme does not distinguish between subordinate and coordinate forms. Consequently, each conjunction in the dictionary was manually aligned with the corresponding UPOS tag. For the remaining PoS categories, mapping to the UPOS tag set was largely straightforward.

3 The Lemma Bank as Linked Open Data

The LiITA Lemma Bank employs the OntoLex-Lemon vocabulary (McCrae et al., 2017), one of the most widely adopted models for the representation and publication of lexical resources as LOD. To ensure consistency with the LiLa KB, the custom ontology developed for the Lemma Bank of the LiLa KB[18] was also integrated, thereby preserving a shared vocabulary across both collections.

Figure 1 illustrates the OntoLex-Lemon model, where Classes are depicted as rectangles, and relationships among classes are represented by arrows labeled with the corresponding Properties.

The main Class of OntoLex-Lemon is `ontolex:LexicalEntry`[19], a unit of lexicon analysis that gathers one or more forms (`ontolex:Form`[20]) and one or more lexical senses (`ontolex:LexicalSense`[21]), lexical concepts (`ontolex:LexicalConcept`[22]) or entities from ontologies.

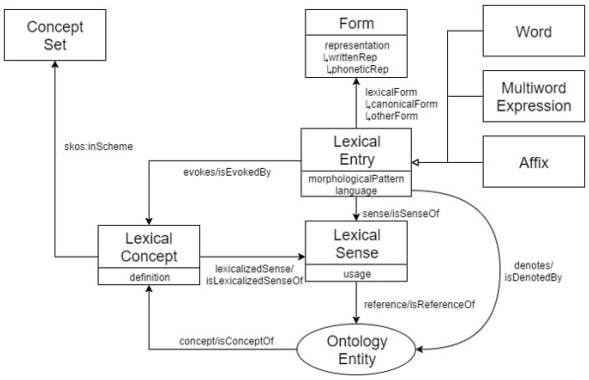

Figure 1: The OntoLex-Lemon model.

In the LiLa KB, lemmas are modelled using a custom ontology[23], which provides detailed morphological and linguistic features specific to Latin, including PoS, gender, and inflectional information, by leveraging the OLiA annotation model (Cimiano et al., 2020, 151-155). This ontology also defines Classes and Properties for the lemmatisation process, notably the Property `lila:hasLemma`[24], which links lemmas to tokens in a corpus.

Within the OntoLex-Lemon framework, forms may exhibit one or more graphical variants (written representations) represented through the Property `ontolex:writtenRep`[25], as well as one or more phonetic variants via the Property `ontolex:phoneticRep`[26]. Among these forms, the Property `ontolex:canonicalForm`[27] designates the conventionally chosen form of all inflected forms of a lexical entry. The Lemma Bank of LiLa—and, consequently, LiITA—consists of such forms, modelled as individuals of the Class `lila:Lemma`[28], which is a subclass of `ontolex:Form`.

Regarding morphological information, each lemma in the Lemma Bank is assigned a PoS tag via the Property `lila:hasPos`[29], following the UPOS tag set.

[18] http://lila-erc.eu/ontologies/lila/

[19] http://www.w3.org/ns/lemon/ontolex#LexicalEntry

[20] http://www.w3.org/ns/lemon/ontolex#Form

[21] http://www.w3.org/ns/lemon/ontolex#LexicalSense

[22] http://www.w3.org/ns/lemon/ontolex#LexicalConcept

[23] http://lila-erc.eu/ontologies/lila/

[24] http://lila-erc.eu/ontologies/lila/hasLemma

[25] http://www.w3.org/ns/lemon/ontolex#writtenRep

[26] http://www.w3.org/ns/lemon/ontolex#phoneticRep

[27] http://www.w3.org/ns/lemon/ontolex#canonicalForm

[28] http://lila-erc.eu/ontologies/lila/Lemma

[29] http://lila-erc.eu/ontologies/lila/hasPOS

3.1 Data Harmonisation

As noted in the previous section, the LiITA Lemma Bank is not a standalone lexical resource. Instead, it is a curated collection of canonical forms that (i) is designed to expand over time as new resources—including previously undocumented lemmas—are integrated, and (ii) serves as a basis for text lemmatisation and for indexing lexical entries within distributed resources published as LOD. Nevertheless, many linguistic resources employ distinct tag sets, standards, and annotation criteria, particularly for lemmatisation.

To accommodate the diverse lemmatisation criteria present in linguistic resources for Italian, the LiITA Lemma Bank implements two dedicated Properties. First, the symmetric Property `lila:lemmaVariant`[30] connects different forms within the same inflectional paradigm that may be used as lemmas, while preserving their assigned PoS. A typical example involves *pluralia tantum*, which can be lemmatised either in the plural or the singular form. Accordingly, the Lemma Bank model allows both the `lila:Lemma` *occhiali* (plural) and *occhiale* (singular) 'optical instrument/glasses' to coexist, linked via the Property `lila:lemmaVariant`. This Property is also applied to align with "simpler" verbal lemmas those citation forms of verbs that exhibit inflectional variations, including those containing reflexive pronouns (e.g., *lavarsi* 'to wash oneself', lemma variant of *lavare* 'to wash') or procomplementary clitics (e.g., *andarci* 'to go there', lemma variant of *andare* 'to go').

While `lila:lemmaVariant` connects different lemmas for the same word that share the same PoS, the Property `lila:hasHypolemma`[31] (and its inverse `lila:isHypolemma`[32]) links lemmas to hypolemmas, which differ in PoS from their corresponding lemma. These hypolemmas are modelled as instances of the Class `lila:Hypolemma`[33], a subclass of `lila:Lemma`.

Through this architecture, the Lemma Bank harmonises divergent lemmatisation practices across resources. For instance, resources that lemmatsze participles differently—some under the participial form and others under the base verbal form—can still be reconciled, thus ensuring interoperability among divergent lemmatisation criteria in corpora and lexical resources.

4 Conclusions and Future Work

In this paper, we have presented LiITA, a knowledge base of interoperable linguistic resources for Italian built in accordance with the principles of the Linked Open Data paradigm. At the core of LiITA lies the Lemma Bank, a centralised collection of Italian lemmas carefully curated to address divergent lemmatisation criteria found in existing linguistic resources. We have illustrated how this novel resource handles challenging cases, such as verbal participles and deadjectival adverbs, through explicit modelling choices that reconcile discrepancies and provide uniform access to lexical information.

By setting up a shared and interoperable framework, LiITA enables consistent and semantically transparent cross-resource integration. This approach not only brings clarity and consistency to the integrated resources, but also fosters reusability and long-term maintainability of linguistic assets across different communities and use cases.

The principles underlying LiITA make it a valuable infrastructure for a wide range of applications, from computational linguistics research to practical tasks in lexicography, corpus linguistics, and language technology. Its interoperable design, combined with Linked Data best practices, opens the possibility of creating richer knowledge graphs that go beyond isolated datasets, thus enabling advanced queries and data mining operations at scale.

Looking ahead, our near-term goals include linking an expanded set of lexical and textual resources for Italian, thereby enhancing LiITA's coverage and robustness. At present, two lexical resources are planned to be linked to LiITA: (i) a dictionary of the Parmigian dialect[34], featuring Italian lexical entries and their corresponding translations into the Parma-area dialect; and (ii) Compl-it, a Linked Open Data computational lexicon for Italian derived from a synthesis of extant linguistic resources.[35]

In terms of textual resources, all publicly available Italian treebanks in the Universal Dependencies repository are planned for linkage to LiITA.

[30] http://lila-erc.eu/ontologies/lila/lemmaVariant

[31] http://lila-erc.eu/ontologies/lila/hasHypolemma

[32] http://lila-erc.eu/ontologies/lila/isHypolemma

[33] http://lila-erc.eu/ontologies/lila/Hypolemma

[34] https://dialetto.comune.parma.it/vocabolarioparmigiano/avvio.htm

[35] https://iris.cnr.it/handle/20.500.14243/530422

Notable differences in tokenisation and lemmatisation among these treebanks[36] will represent an optimal test case for assessing the effectiveness of LiITA's harmonisation strategies.

Following the approach adopted for LiLa (Passarotti et al., 2024), LiITA will also develop an online service to facilitate linkage of raw texts in Italian through automatic tokenisation and lemmatisation. This service will rely on a newly trained model of the Stanza package for language analysis (Qi et al., 2020) which leverages all extant Italian treebanks as its training data[37]. In addition, LiITA will offer a user-friendly graphical interface to streamline advanced data interrogation across all interconnected resources, simplifying the construction of complex SPARQL queries.

As the Lemma Bank expands and more resources are integrated, LiITA will make a substantial contribution to the domain of Linguistic Linked Open Data. For example, as stated previously, LiITA does not currently address multiword expressions (MWEs). In corpora, MWEs are almost always lemmatised using the lemmas of their individual components and, in some cases, through a kind of super-lemma representing the MWE as a whole. However, MWEs are indeed present in certain lexical resources, and at present, these are not being captured in our work. The possibility of including MWEs—also supported by the OntoLex-Lemon model—in the lexical base will require careful consideration, precisely to accommodate the information provided by some of these lexical resources, and expand the potential of the LiITA Lemma Bank. Through these efforts, we aim to establish LiITA as a cornerstone of an Italian linguistic ecosystem—one that harmonises diverse data sources, stimulates collaborative research, and promotes high-quality linguistic insights for both humans and machines.

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. ESWC 2012*, volume 7295 of *Lecture Notes in Computer Science*, pages 225–239, Berlin, Heidelberg. Springer.

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

Philipp Cimiano, Christian Chiarcos, John P. McCrae, and Jorge Gracia. 2020. *Linguistic Linked Data: Representation, Generation and Applications*. Springer, Cham.

R Rossini Favretti, Fabio Tamburini, and Cristiana De Santis. 2002. 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*, pages 27–38.

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. ESWC 2015*, volume 9088 of *Lecture Notes in Computer Science*, pages 225–239, Cham. Springer.

Nancy Ide and James Pustejovsky. 2010. What does interoperability mean, anyway? toward an operational definition of interoperability for language technology. In *Proceedings of the Second International Conference on Global Interoperability for Language Resources. Hong Kong, China*.

Caterina Mauri, Silvia Ballarè, Eugenio Goria, Massimo Cerruti, Francesco Suriano, et al. 2019. Kiparla corpus: a new resource for spoken italian. In *CEUR WORKSHOP PROCEEDINGS*, pages 1–7. SunSITE Central Europe.

John P McCrae, Julia Bosque-Gil, Jorge Gracia, Paul Buitelaar, and Philipp Cimiano. 2017. The ontolex-lemon model: development and applications. In *Proceedings of eLex 2017 conference*, pages 19–21.

Marco Passarotti, Francesco Mambrini, and Giovanni Moretti. 2024. The services of the lila knowledge base of interoperable linguistic resources for latin. In *Proceedings of the 9th Workshop on Linked Data in Linguistics@ LREC-COLING 2024*, pages 75–83.

Slav Petrov, Dipanjan Das, and Ryan McDonald. 2012. A Universal Part-of-Speech Tagset. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2089–2096, Istanbul, Turkey. European Language Resources Association (ELRA).

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A Python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*.

[36] https://universaldependencies.org/treebanks/it-comparison.html

[37] The model can be found at https://github.com/LiITA-LOD/LiITA_NLP_Models

Adriana Roventini, Rita Marinelli, and Francesca Bertagna. 2016. ItalWordNet v.2. ILC-CNR for CLARIN-IT repository hosted at Institute for Computational Linguistics "A. Zampolli", National Research Council, in Pisa.

Acknowledgments

This contribution is funded by the European Union - Next Generation EU, Mission 4 Component 1 CUP J53D2301727OOO1. 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.