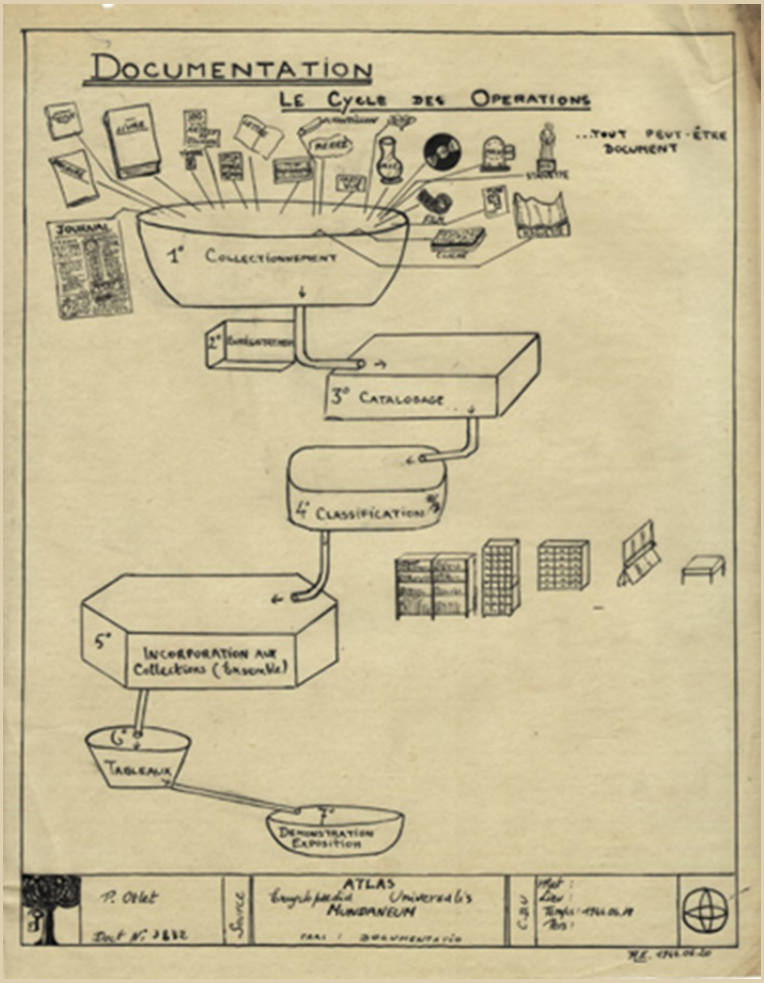
AIDAinformazioni

RIVISTA SEMESTRALE DI SCIENZE DELL'INFORMAZIONE

NUMERO 3-4
ANNO 42
LUGLIO-DICEMBRE 2024



CACUCCI EDITORE
BARI

AIDAinformazioni

Rivista semestrale di Scienze dell'Informazione

Fondata nel 1983 da Paolo Bisogno

Proprietario della rivista:
Università della Calabria

Direttore Scientifico:
Roberto Guarasci, *Università della Calabria*

Direttore Responsabile:
Fabrizia Flavia Sernia

Editrice: Cacucci Editore S.a.s.
Via D. Nicolai, 39 – 70122 Bari (BA)
www.cacucciedITore.it
e-mail: riviste@cacucciedITore.it
Telefono 080/5214220

AIDAinformazioni

Rivista semestrale di Scienze dell'Informazione

«AIDAinformazioni» è una rivista scientifica che pubblica articoli inerenti alle Scienze dell'Informazione, alla Documentazione, all'Archivistica, alla Gestione Documentale e all'Organizzazione della Conoscenza ma amplia i suoi confini in ulteriori campi di ricerca affini quali la Terminologia, la Linguistica Computazionale, la Statistica Testuale, ecc. È stata fondata nel 1983 quale rivista ufficiale dell'Associazione Italiana di Documentazione Avanzata e nel febbraio 2014 è stata acquisita dal Laboratorio di Documentazione dell'Università della Calabria. La rivista si propone di promuovere studi interdisciplinari oltre che la cooperazione e il dialogo tra profili professionali aventi competenze diverse, ma interdipendenti. I contributi pubblicati affrontano questioni teoriche, metodologie adottate e risultati ottenuti in attività di ricerca o progettuali, definizione di approcci metodologici originali e innovativi, analisi dello stato dell'arte, ecc.

«AIDAinformazioni» è riconosciuta dall'ANVUR come rivista di Classe A per l'Area 11 – Gruppo Scientifico Disciplinare 11/HIST-04 – Scienze del libro, del documento e storico-religiose e come rivista scientifica per le Aree 10 – Scienze dell'antichità, filologico-letterarie e storico-artistiche; 11 – Scienze storiche, filosofiche, pedagogiche e psicologiche; 12 – Scienze giuridiche; 14 – Scienze politiche e sociali. È anche annoverata dall'ARES (Agence d'évaluation de la recherche et de l'enseignement supérieur) tra le riviste scientifiche dell'ambito delle Scienze dell'Informazione e della Comunicazione. La rivista è, inoltre, indicizzata in: ACNP – Catalogo Italiano dei Periodici; BASE –Bielefeld Academic Search Engine; ERIH PLUS – European Reference Index for the Humanities and Social Sciences – EZB – Elektronische Zeitschriftenbibliothek – Universitätsbibliothek Regensburg; Gateway Bayern; KVK – Karlsruhe Virtual Catalog; The Library Catalog of Georgetown University; SBN – Italian union catalogue; Ulrich's; Union Catalog of Canada; LIBRIS – Union Catalogue of Swedish Libraries; Worldcat.

I contributi sono valutati seguendo il sistema del *double blind peer review*: gli articoli ricevuti sono inviati in forma anonima a due referee, selezionati sulla base della loro comprovata esperienza nei topics specifici del contributo in valutazione.

AIDAinformazioni

Anno 42

N. 3-4 – luglio-dicembre 2024

CACUCCI EDITORE

BARI

Via Nicolai, 39 – 70122 Bari – Tel. 080/5214220
http://www.cacucciedi tore.it e-mail: info@cacucci.it

Sommario

Contributi

Rubriche

Contributi

AIDAinformazioni
ISSN 1121–0095
ISBN 979-12-5965-456-4
DOI 10.57574/596545646
pag. 119-148 (luglio-dicembre 2024)

CompL-it: a Computational Lexicon of Italian

Flavia Sciolette, Andrea Bellandi, Emiliano Giovannetti, Simone Marchi*

Abstract: This paper describes CompL-it, a new open computational lexicon for contemporary Italian. The resource was constructed from three sources: an already available Italian lexicon, a lemmatized list of inflected forms obtained from a morphological analyser, and a set of treebanks. Integrating these resources required a standardisation process in accordance with the standards of the Linguistic Linked Open Data community, which was necessary for the subsequent conversion into the OntoLex-Lemon model. The resulting computational lexicon comprises approximately 100,000 lexical entries, 790,000 forms, 57,000 senses, and 86,000 semantic relations. The lexicon, thanks to its rich and articulated linguistic structure, can be used, as shown, to enhance information retrieval in the context of full-text search tasks.

Keywords: Computational Lexicon, Linguistic Resources, Linguistic Linked Open Data, OntoLex-Lemon, Information Retrieval.

1. Introduction

While a significant number of digital lexical resources are available for many languages (such as various multilingual WordNets) (Princeton University n.d.; MultiWordNet n.d.; Global WordNet Association n.d.), only a few integrate different layers of linguistic information, such as morphology, semantics, and syntax. This is not surprising, as the construction of a computational lexicon[1] that conveys linguistic information across different layers can be an extremely time-consuming task that requires advanced linguistic expertise.

* CNR-Istituto di Linguistica Computazionale (ILC) "A. Zampolli", Pisa, Italy. flavia.sciolette@ilc.cnr.it; andrea.bellandi@ilc.cnr.it; emiliano.giovannetti@ilc.cnr.it; simone.marchi@ilc.cnr.it.

[1] In this context, a computational lexicon can be defined as a resource that contains information about words, their meanings, and linguistic properties, designed to be used by computer systems for tasks like natural language processing (NLP), machine translation, or text analysis. It typically includes details such as word categories (e.g., noun, verb), syntactic information, and semantic relationships.

On the other hand, the need for resources of this kind is long-standing. Italian linguistics, for example, has always shown an interest for lexical data (Sabatini 2006), which has been encouraged by the increasing availability of many corpus-based resources. As documented in Chiari (2012) many projects involving corpora (monolingual, parallel, domain-specific) have flourished and both digitised traditional dictionaries and computational dictionaries have taken advantage of them, for example to calculate the frequency of words or to increase their lexical coverage (for example by adding neologisms). In terms of exploitation, a number of applications are meant to take advantage of lexical resources, such as sentiment analysis (Prakash and Aloysius 2021), and «semantic role labelling, verb sense disambiguation, and ontology mapping» (Brown et al. 2022, 2).

In the context of archival science and document management, the availability of linguistic resources to support the organisation and retrieval of information has been considered crucial for many years (Chen et al. 1995; Smith 1997; Thompson et al. 2011). In these fields, the development of computational lexicons can provide a fundamental contribution to the community, expanding the potential for knowledge analysis and management. It is believed that a resource capable of formalising a language's lexical and semantic structures in a complex way can improve the efficiency of archiving, classification, and information retrieval activities, within a document management paradigm increasingly supported by IT tools (Bamman and Crane 2010; Hmeidi et al. 2016; Passarotti and Mambrini 2021). The creation of increasingly efficient tools for automating archival and document practices can greatly simplify the management of large volumes of unstructured data, enhancing precision in indexing and retrieving information. Furthermore, a computational lexicon can serve as a key linguistic resource for building Knowledge Organization Systems (KOSs), such as ontologies and thesauri, crucial elements for knowledge organisation (Hodge 2000; Shiri 2015).

In this work, we illustrate CompL-it, an Italian computational lexicon built by leveraging existing resources, whose data have been thoroughly analysed, extracted, converted, and interconnected. CompL-it has been made freely available as Linguistic Linked Open Data (LLOD) on the CLARIN repository (CLARIN-IT n.d.a).

2. State of the art

In order to define the state of the art regarding computational lexical resources for the Italian language we first conducted a search on the Virtual Language Observatory (CLARIN VLO n.d.) (VLO) of the European infrastructure CLARIN (Common Language Resources and Technology Infrastructure). This database, which contains hundreds of thousands of references

to language resources and tools, was browsed using the available filters. In particular, a search was carried out by type of resource, selecting *lexicalResource*, and specifying the Italian language. In addition, from the obtained list we excluded automatically produced resources (i.e, not revised by hand), lists of idiomatic expressions, lists of terms with no linguistic information at all, multilingual named entities, sets of embeddings, parallel corpora, metadata, resources that only appear by virtue of some references to the Italian language and, finally, all the resources whose data are not open and freely available. Among this last type of resources, however, it is worth mentioning BabelNet (Navigli and Ponzetto 2012; BabelNet n.d.), Senso Comune (Vetere et al. 2011), and Italian FrameNet (Basili et al. 2017), particularly for the richness of data they offer.

The resources identified on CLARIN VLO that met the above criteria were 14, and can be classified into two categories: 11 lexical resources and 3 terminological resources.

The available lexical resource that, for this work, has been taken as the main reference (and used as one of the sources) is LexicO (Sciolette, Giovannetti, and Marchi 2023), a multi-layered computational lexicon developed at CNR-ILC and built from Parole-Simple-Clips (PSC) (Bel et al. 2000; Ruimy et al. 2002; ILC4CLARIN CNR 2016). More details on the nature of LexicO will be provided later in section 3.1.

Another very rich lexical resource, also developed at CNR-ILC, is ItalWordNet (Roventini et al. 2003), available as an SQL dump on the VLO in its second version (Roventini, Marinelli, and Bertagna 2016). The VLO also mentions MultiWordNet (Pianta, Bentivogli, and Girardi 2002; MultiWordNet n.d.), realised as an extension of Princeton's WordNet (Miller 1995; Princeton University n.d.), and which also includes data for the Italian language. Multilingual language resources include OmegaWiki (Meijssen 2014), an open and collaborative resource whose aim is «to describe all words of all languages with definitions in all languages» and includes lexical, terminological and ontological information (WikiMedia 2022).

In addition to the resources listed so far, which are structured as lexicons, other resources are available for Italian that convey individual layers of linguistic information. Building on the aforementioned PSC and ItalWordNet, a resource that provides semantic data called the Italian Sense Inventory was created. This resource was developed within the ELEXIS project (ELEXIS n.d.) to support Word Sense Disambiguation tasks.

On the VLO, there are also two resources developed by the same author that complement each other: Italian Function Words (Grella 2018a) and Italian Content Words (Grella 2018b). The former, as the name suggests, contains Italian function words and is designed to support tasks such as POS tagging and syntactic parsing. The second constitutes a morphological dictionary of

over 2 million inflected forms that, however, includes hundreds of thousands of forms that, although morphologically correct, are not represented in linguistic usage. The last two lexical resources we include in this review are Universal Derivations (Kyjánek et al. 2021) and Universal Segmentations (Žabokrtský et al. 2022), both multilingual, in which about 10 thousand Italian lemmas are linked to their respective segmentations, derived forms and compounds.

With regard to the three terminological resources identified in the VLO we first mention Geodomain WordNet, a collection of geographical terms linked to the English and Italian WordNets (Frontini, Del Gratta, and Monachini 2016). The other two resources, developed within the Pan-Latin Terminology Network (Realiter n.d.), are the Pan-Latin Lexicon of Collars and Sleeves in Fashion and Costume (Zanola et al. 2023) and the Pan-Latin Textile Fibres Vocabulary (Dankova, Zanola, and Calvi 2022).

Although not available on CLARIN VLO, Morph-it! (Zanchetta and Baroni 2005) is a freely accessible and rich morphological resource for Italian, consisting of 504,906 inflected forms and 34,968 lemmas. However, as the authors note (Morph-it! 2018), because it is derived from an Italian newspaper corpus, the resource has «many gaps in basic, every-day vocabulary».

Another interesting resource worth mentioning is SimpleLEX-IT, as it was built similarly to CompL-it, i.e., by combining together different existing resources (Mazzei 2016; SimpleLEX-IT n.d.). In particular, SIMPLELex-it was developed by integrating morphological data from the previously cited Morph-it!, the *Vocabolario di base della lingua italiana* by Tullio De Mauro (De Mauro 1980; 2016), two entries of the Italian Wikipedia concerning verbs (Wikipedia 2024a; 2024b) and, finally, data from the Italian Universal Dependencies (UD) treebanks (Universal Dependencies n.d.a).

In the context of Linked Open Data – or, more precisely, LLOD, understood as the reference community for the creation and sharing of resources according to LOD principles (Cimiano et al. 2020; LLOD n.d.) – the linguistic resources currently available for Italian include RDF datasets for the previously mentioned PSC (Del Gratta et al. 2015) and IWN (Bartolini 2016) resources. The LLOD landscape, however, offers resources for different languages, both contemporary and historical varieties; as an illustrative and non-exhaustive example in a constantly expanding field, it is worth mentioning Dbnary (Sérasset 2015), the multilingual resource based on Wiktionary, made available according to LLOD principles. For historical varieties, we cite LiLa – Linking Latin, a knowledge base for Latin that now includes several resources (Mambrini and Passarotti 2023), and the DigitAnt project for ancient language varieties in Italy (Mallia et al. 2024). On the terminological front, CHAMUÇA is noted, a resource for Portuguese loanwords in Asian languages (Khan et al. 2024), and initial studies for a resource related to terms in the Babylonian Talmud (Sciolette 2024), with the formalisation of contex-

ts through the OntoLex module FrAC (Frequency, Attestation and Corpus Information), which is currently under development (Chiarcos et al. 2022; Github n.d.a).

As a source of reference data for the construction of the CompL-it lexicon, and as already mentioned, we chose LexicO, one of the freely available lexical resources for the Italian language. The motivation for this choice is twofold, and is partly also evident from the data reported in section 4.3, where LexicO has been quantitatively compared to five other resources. First of all, LexicO (see section 3.1) is a multi-layered linguistic resource, in which information of various kinds (phonological, morphological, syntactic and semantic) is encoded: in this sense, it constitutes a *unicum* of its kind, given that all the others limit themselves to representing, essentially, lexemes linked to each other through semantic relations. Moreover, from a more quantitative point of view, LexicO with its dense network of relations constitutes an extremely rich resource of linguistic data.

However, LexicO also has limitations, both in terms of coverage in the number of lexical entries, in terms of specific content (such as inflected forms or missing lexical senses) and, finally, in terms of the data format in which it is currently represented.

The idea of building CompL-it arose precisely due to these limitations of LexicO: to ensure maximum lexical coverage, the linguistic data from LexicO was integrated with data from two additional sources. Furthermore, standards defined by the LLOD community were adopted for the model and representation format.

3. The sources

As stated in the previous section, LexicO was selected as the foundational resource for constructing CompL-it. Additionally, we considered two other sources: M-GLF (MAGIC-Generated Lemmatized Forms), a list of lemmatized forms with morphological information generated by the MAGIC tool (Battista and Pirrelli 1999; Pirrelli and Battista 2000), and a set of Italian language treebanks available through the UD repository (Universal Dependencies n.d.b). All these resources have been chosen both for the richness of the data they provide and because they have been manually constructed or validated.

These three resources are very different from each other in terms of formats, models and purposes, and therefore their integration required a process of standardisation, as described in Section 4.1. In the following sections, we describe the three resources together with some specific pre-processing interventions carried out prior to the data standardisation and conversion steps necessary to create CompL-it.

3.1. LexicO

LexicO is a computational lexicon of Italian, available on CLARIN as a relational database (CLARIN-IT n.d.b). This resource is derived from the above mentioned PSC, with which it shares the same model based on the theory of Generative Lexicon by James Pustejovsky (Pustejovsky 1995).

LexicO contains four layers of linguistic information: a morphological layer, which describes lemmas, parts of speech (POS), and inflectional rules; a semantic layer, which includes information about senses and their relationships; a syntactic layer, detailing the syntactic behaviour of units and their phrase structure; and finally, a phonological layer, which involves inflected forms generated from the inflectional rules in the morphological layer. Although each layer operates independently, there are connections between different units, such as between syntactic and semantic entries. Since its initial use in tasks such as full-text search (Giovannetti et al. 2022), it has become evident that there is a need to convert all the data into a format compliant with current standards.

Morphological units form the basis of lexical entries in LexicO. Each unit is associated with a POS value and a set of morphological rules used to generate grammatically correct forms. These forms are defined as a type of entry called phonological units.

Each association between a lemma and a form is described with a POS and a certain number of morphological traits, as shown in Table 1.

Lemma	Form	POS	MorphFeat
abbandonare	abbandonai	V	1|singular|indicative|past

Table 1: *abbandonare* (to abandon) with its form *abbandonai*, its POS (verb), and its morphological features (first person singular, indicative, past).

As already mentioned in the previous section, LexicO is directly derived from the PSC computational lexicon. This initial resource, while already quite comprehensive, contained redundant or duplicated data and some entries lacking in information: an emblematic example is the absence of the form *vado* (I go) of the verb *andare* (to go). Although these issues did not diminish the intrinsic value of the source, they required interventions to address the gaps where possible. All interventions are documented in Sciolette, Giovannetti, and Marchi (2023).

3.2. M-GLF

The second lexical source we used to build up CompL-it was M-GLF, a list of lemmatised forms generated by MAGIC, a morphological analyser for

Italian (Battista and Pirrelli 1999). The tool includes three modules: a lexicon compiler, the morphological analyser itself, and a morphological generator. We used this latter to generate the M-GLF list of forms (CLARIN-IT n.d.c) by starting from a list of morphological rules for lemmas, endings and idiosyncratic entries, contained in a morphological database.

An example of a M-GLF entry follows which is relative to a form of the Italian verb *abbaiare* (to bark):

```
[1]          MACRO:word[l_abbaiare,abbaiera',v_fin,3,!,s,-
fut,ind,!,!]
```

In this example, *l_abbaiare* is the lemma of the form *abbaierà*, which is the third person singular of the finite verb *abbaiare* in the indicative mood and future tense. These morphological traits are indicated in the line, separated by commas. Exclamation points represent *null* values for unspecified features, such as degree, which is only relevant for adjectives.

The MAGIC generation tool is based on rules that constitute an extremely rigorous model. First of all, the tool was unable to generate certain forms, such as the absolute superlative. Moreover, the generation of entries produced some inconsistencies. In particular, we found entries having multiple POS, such as *noun* and *adjective* (e.g., *svedese* can indicate both the noun for a resident of Sweden and the adjective for denoting the quality of being Swedish). In these cases, we decided to intervene by splitting the entries with double POS into distinct entries, each of which having its own POS with the correct morphological traits.

3.3. Treebanks

To further enrich the morphological layer of CompL-it, we also decided to consider lemmas, forms, and morphological information obtained from the available treebanks for Italian. The treebanks are collected in the UD repository, according to a common annotation scheme (Universal Dependencies n.d.c), used for resources in different languages.

We only included treebanks that have been manually revised: three based on balanced corpora of general-purpose texts (such as newspapers, legal documents, etc.) and one from a specific domain. We considered the following treebanks:

- ISDT (Italian Stanford Dependency Treebank) (Universal Dependencies n.d.d): this resource was obtained through a semi-automatic conversion process starting from MIDT (the Merged Italian Dependency Treebank). It is the result of merging pre-existing dependency-based resources, aimed at improving the interoperability of available data.

The schema was partially adapted to account for the specific features of the Italian language (Simi, Bosco, and Montemagni 2014);

- VIT (Universal Dependencies n.d.e): it is a conversion of VIT (Venice Italian Treebank), developed at the Laboratory of Computational Linguistics at Università Ca' Foscari in Venice. Originally a constituency-based treebank, VIT includes linguistic materials of various types, extracted from five text typologies and spoken dialogues. The data underwent conversion to the CoNLL-U format (Universal Dependencies n.d.f), along with several stages of data harmonisation;
- ParTUT (ParallelTut) (Universal Dependencies n.d.g): it is a conversion of a multilingual parallel treebank developed at the University of Turin consisting of a variety of text genres, including talks, legal texts and Wikipedia articles (Sanguinetti and Bosco 2015);
- ParlaMint-It (Universal Dependencies n.d.h): it is a collection of transcriptions of parliamentary sessions of the Italian Senate, annotated in Universal Dependencies. The corpus is part of a larger multilingual collection of parliamentary transcripts built during the ParlaMint project (CLARIN n.d.).

Although the selected treebanks were chosen precisely because they underwent manual revision, they are not entirely free of errors, including gaps and inconsistencies, particularly in morphological features, which can introduce noise.

For example, there are cases where a word appearing in the treebanks is annotated with fewer morphological features than the same word appearing in the other two resources. This is the case for the form *abilitati* (enabled, as in "enabled users"), described in the treebanks only through lemma and POS, while in LexicO and M-GLF, this word is also provided with number (plural) and gender (masculine) features. In all these cases, in CompL-it, the words from the resource richer in linguistic information have been added.

4. The nature of CompL-it

This section illustrates the resource CompL-it, by starting with the necessary standardisation process that had to be carried out, described in Section 4.1. The conversion in RDF format is briefly described in Section 4.2, while a quantitative analysis of the resource is carried out in Section 4.3. To ensure data interoperability we chose Ontolex-Lemon as the backbone model of Compl-It, as it is the *de facto* standard for representing lexical resources in the Linked Open Data community.

In order to proceed to the standardisation and conversion processes, it was necessary to carry out some pre-processing steps. Some of these interventions concerned all the resources and involved, in particular: i) the conversion of superscripts into accented letters and distinction of high and low accents; ii) the removal of proper nouns, such as named entities (e.g. *Petrarca*) and trade names (e.g. *Xerox*); iii) the exclusion of abbreviations (e.g. *Dott*. instead of *doctor*); iv) the exclusion of multiword expressions, which included nouns, adjectives, adverbs and prepositions (e.g. the expression *a ferro e fuoco*); v) the removal of unadapted loanwords (e.g. word processor).

Loanwords, multi words and proper nouns require different treatment and will therefore be the subject of future work.

4.1. Standardisation

The following paragraphs describe the interventions undertaken to make the models of the considered resources homogeneous in terms of morphology and semantic relations.

4.1.1 Morphology

As can be seen from Section 3, the models and reference vocabularies of LexicO, M-GLF and treebanks differ from each other, often profoundly. This divergence between linguistic information representation systems is also motivated by the different approach used to represent linguistic data.

In fact, M-GLF and LexicO can be included in the category of lexicographic resources, whereas the standard used for treebanks, based on the UD paradigm, pertains to the annotation of linguistic corpora.

In order to standardise the vocabularies, we decided to use LexInfo, an inventory of types, values and properties designed to describe linguistic data categories (LexInfo n.d.). LexInfo comprises morphological properties, such as gender, number, mood, grammatical categories (POS), and semantic relations, such as synonymy, hypernymy, and so on. This choice is justified primarily by the alignment of this vocabulary with the OntoLex-Lemon model (W3C 2016) used to represent CompL-it, as LexInfo serves as the reference linguistic ontology for resources created with this model. Additionally, LexInfo complies with other standards related to the OntoLex-Lemon model, including OLiA (Ontologies of Linguistic Annotation) (Chiarcos and Sukhareva 2015; OLiA n.d.), a repository of linguistic categories specific to annotated corpora. Ultimately, the selection of the LexInfo vocabulary was largely driven by the need to produce a lexical resource that is as interoperable as possible with other re-

sources based on the OntoLex-Lemon model, in accordance with the Linked Data paradigm.

This difference in the representation of linguistic information in the models of the three sources occurs mainly at the level of the association between POS and morphological traits. In fact, LexicO and M-GLF are based on a model and specific vocabularies of labels, which are characterised by a very fine-grained categorisation of POS. In the case of the treebanks, the UPOS has a coarser grain: for example, the combination of UPOS and trait "possessive" with the value "yes" in the treebanks has been mapped to a specific LexInfo POS. For example, the entry *mio* (mine) appears in treebanks with the following annotation: PRON for the UPOS, with the feature "Poss=Yes". In CompL-it, this entry has been described with POS "possessivePronoun", according to the LexInfo vocabulary. More in general, the vocabularies of LexicO, M-GLF and the treebanks were mapped into LexInfo, according to the following scenarios: i) direct mapping between POS, if available (as was often the case for LexicO and M-GLF); ii) conversion of POS and trait combinations present in the treebanks into a LexInfo POS; iii) conversion into OLiA[2] or proposal of an *ad hoc* label, if the trait was not present in LexInfo. The conversion tables have been made available on (Github n.d.b).

4.1.2 Semantics

In CompL-it, 137 types of semantic relations derived from LexicO have been included. These relations are categorised into eight classes, listed below.

- Four classes are related to the four *qualia* roles taken from the Generative Lexicon theory[3], namely:
  - Formal: the role that describes the entity conveyed by the sense in relation to other entities. An example of a relation associated with the formal role is hyponymy, e.g., *gatto-mammifero* (cat-mammal);
  - Agentive: the role that provides information about the origin of an entity. An example of a relation associated with the agentive role is *caused by*, e.g., *infezione-batterio* (infection-bacterium);
  - Telic: the role that specifies a function of an entity. An example of a relation associated with the telic role is *Object of activity*, linking an object to a certain event, such as *libro-leggere* (book-to read);

[2] OLiA has been used in the conversion of two traits in M-GLF for "Diminutive" and "Augmentative".

[3] For an overview of the theory and the relationship between qualia roles and relations, see (Sciolette, Giovannetti, and Marchi 2023). Following the terminology in the PSC documentation, entities refer to the concept expressed by the sense, conveyed by a specific entry. These entities can be connected to each other through semantic relations. Semantic relations are also classified according to qualia roles.

  - Constitutive: the role that describes the composition of an entity; an example of a relation associated with the constitutive role is meronymy, as *senatore-senato* (senator-senate).
- A derivational class, reserved for relations concerning senses that undergo a change in grammatical category, e.g., from adjective to noun, as *triste-tristezza* (sad-sadness).
- A class related to polysemy relations, as listed in Malmgren (1988). An example of a regular polysemy class is *Substance-Colour*, as seen in the sense of *turchese* (turquoise), which can refer to both the gemstone and the colour.
- Two classes not documented in the original PSC model, namely synonymy, e.g. *ciclone-uragano* (cyclone-hurricane), and metaphor, e.g. *leone* (lion) to relate the sense of a brave man to the sense of the animal.

These few examples convey the image of a system of relations aimed at defining meaning according to very fine-grained categories, as exemplified in the case of meronymy, which distinguishes senses related to parts of a set, components of a group, and, as a subclass, followers of a certain movement, as for example *Marxista* (Marxist).

At present, it has proved particularly complex to find exact correspondences between the semantic relations described in the reference ontologies (LexInfo and OLiA) for the OntoLex-Lemon model. In some cases, it was necessary to define relations from scratch.

The need to update lexical resources in Linked Data formats was also felt in the past and led to the creation of some resources conforming to previous versions of the OntoLex-Lemon model (Del Gratta et al. 2015; Villegas and Bel 2015). However, it was not possible to reuse these resources either because they did not include updates to the OntoLex-Lemon model, or because they did not include a mapping with other reference models, such as LexInfo.

To ensure maximum interoperability, where possible, relations formalised from scratch were linked to the corresponding reconstructed resources in previous versions of the OntoLex-Lemon model, with the seeAlso relation (W3C 2005).

For the construction of a vocabulary of CompL-it relations, the following were also considered: i) equivalences, where possible, with LexInfo, such as in the case of synonymy; ii) additional properties, not previously defined by other resources, but reconstructed through documentation and analysis of the resource.

The vocabulary definition phase also had a direct effect on the enrichment interventions of the resource. For example, the mapping with LexInfo made it possible to define an additional relation, hypernymy, as the inverse of hyponymy (which translates the *isA* relation present in the LexicO model); since hy-

pernymy is the inverse of hyponymy, even if the relation is not described in the source resource, it was still possible to infer a number of instances. This happened for all relations of which we could formalise additional properties from the study of the documentation (as in the case of *causes*, inverse of *causedBy*).

4.2. Conversion to Linked Data

Once the data from the three lexical sources had been standardised, they were converted into the Linked Data format. After introducing the reference model adopted, an example of converted data is provided.

4.2.1 The OntoLex-Lemon model

In the context of the representation and publication of lexical data as knowledge graphs and/or as Linguistic Linked Open Data, the OntoLex-Lemon model has become a *de facto* standard. This model was created with the aim of supporting the linguistic foundation of a given ontology by adding information on how ontological entities are lexicalised in different languages. However, OntoLex-Lemon can also be used as a lexicographic model to represent linguistic entities without any concept they denote being defined. OntoLex-Lemon is inspired by many other models, in particular the Lexical Markup Framework (LMF) (Francopoulo et al. 2006), LexInfo (Cimiano et al. 2011) — aligned with DatCatInfo (DatCatInfo n.d.) — and LIR (Linguistic Information Repository) (Montiel-Ponsoda et al. 2008).

Figure 1 represents the core of the model, called *ontolex*. The rectangles represent the classes of the model, the arrows with full heads represent the properties of the objects, and the arrows with empty heads represent the subclass relationships.

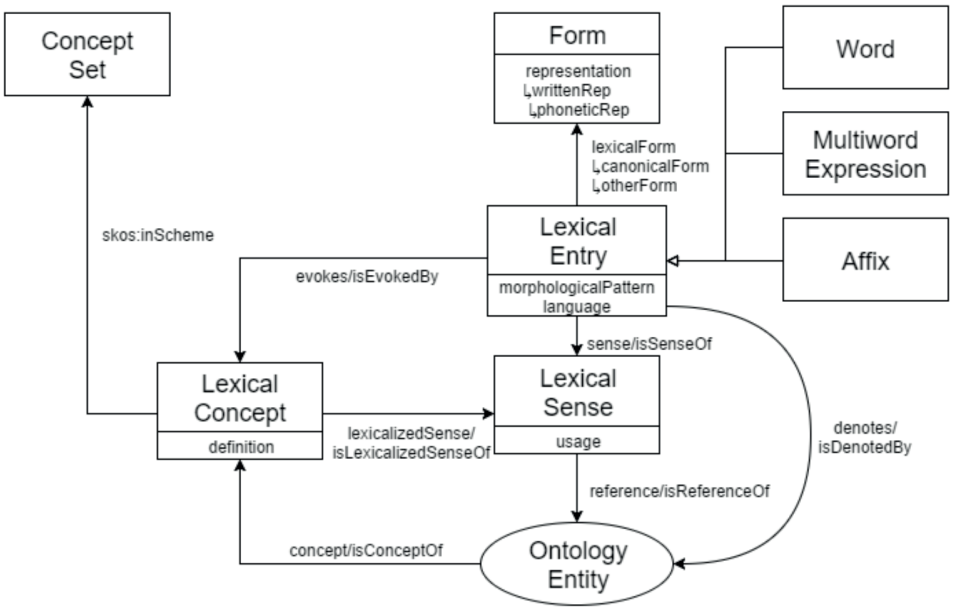


Figure 1: The core model of OntoLex-Lemon (ontolex). Picture taken from the W3C OntoLex Final Community Report at (W3C 2016).

ontolex is based on the definition of three fundamental classes: i) LexicalEntry, that «represents a unit of analysis of the lexicon that consists of a set of forms that are grammatically related and a set of base meanings that are associated with all of these forms»; ii) Form, that «represents one grammatical realisation of a lexical entry»; iii) LexicalSense, that «represents the lexical meaning of a lexical entry when interpreted as referring to the corresponding ontology element, if it is given». With reference to Figure 1, it is necessary to emphasise that *ontolex* also allows us to express the fact that a given lexical entry evokes a certain mental concept or refers to an entity with a formal interpretation defined in an ontology. Therefore, OntoLex-Lemon introduces a fourth element, the LexicalConcept class, which represents a mental abstraction, concept or unit of thought that can be lexicalised by a given collection of meanings.

The rest of the architecture of OntoLex-Lemon is divided into 4 modules, each representing a different linguistic aspect, namely: i) decomposition (decomp), i.e. the process of describing which elements constitute a multiword or compound; ii) the lexical and semantic relations between lexical entries and lexical senses respectively (vartrans); iii) the syntactic behaviour of lexical entries (synsem); iv) the description of the metadata of the lexical resource (lime). However, in this paper, our conversion work will mainly use *ontolex*, dealing with neither composition nor syntactic aspects in particular.

It is important to emphasise that the model abstracts from specific linguistic theories or category systems used to describe the properties of lexical entries and their syntactic behaviour. The re-use of existing category systems or linguistic ontologies is therefore strongly encouraged. In our case, as recommended by the community that developed the model and as described in section 4.1.1, the LexInfo model was used, which offers a rich vocabulary of linguistic categories and relationships for morphology, syntax and semantics.

4.2.2 Representing data in RDF OntoLex-Lemon

The conversion of the standardised data 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[4]. The obtained lexicon was then loaded into Ontotext GraphDB (Ontotext n.d.), a semantic repository compliant with RDF and SPARQL (W3C 2013). Below 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, co-
      niglio_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
```

[4] Turtle is a serialisation format for RDF data types (W3C 2014).

```
      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;
      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 associated with two forms, one of which is defined as the canonical form (the lemma) and the other suitable for representing the plural form *conigli* (rabbits), both of which are equipped with the appropriate morphological traits. The lexical entry is also associated, via the *ontolex:sense* relation, with three lexical senses, each of which has a natural language definition. Furthermore, the first two senses are also endowed with semantic relations that link them to other lexical senses. For example, *rabbit_sense_2* is defined as a hyponym of *mammal_sense*.

4.3. CompL-it in numbers

In this section, the CompL-it lexicon is described from a quantitative perspective, both by enumerating the entities and relations it comprises and by comparing it with lexicographic resources available for the Italian language of a similar nature.

From a morphological standpoint, the resource is composed of 101,795 lexical entries (comprising a total of 791,541 word forms), classified with 36 POS categories and described with morphological traits. Figure 2 depicts a Venn diagram representing the different dimensions, in terms of lexical entries, of the three source resources and their intersections. As observed, the most significant contribution of words comes from M-GLF. However, both LexicO and the treebanks contribute significantly with a total of 47,069 forms and 9,028 additional lexical entries.

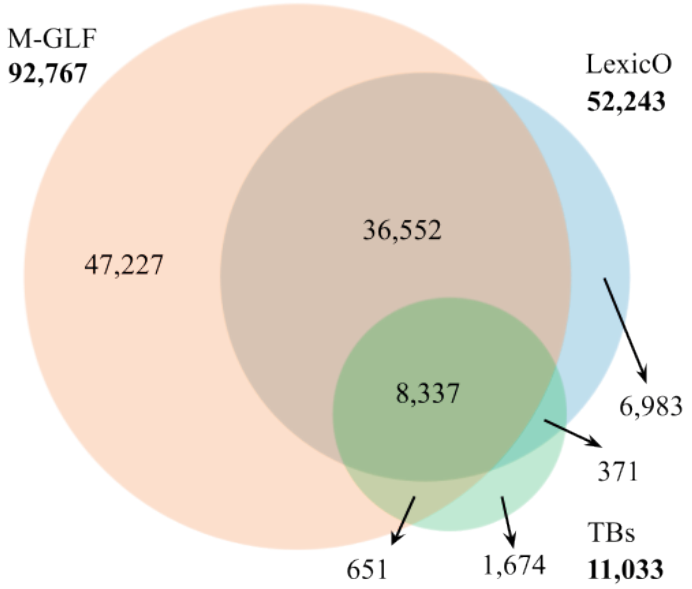


Figure 2: Representation, by the number of lexical entries, of the three source resources and their various intersections.

Going more into the specific linguistic content, Table 2 shows the distribution of word forms by POS: as expected, most are verbal forms (69% of the total). The *other* class encompasses the 32 POS not explicitly specified, such as, for example, number, conjunction, determiner, and preposition.

POS	LexicO	M-GLF	TBs	CompL-it
verb	345,307	526,635	10,741	545,342
noun	76,396	115,562	11,723	136,744
adj.	45,712	98,603	7,083	103,869
adv.	742	2,818	845	3223
other[5]	935	670	926	2,364
total	469,092	744,288	31,318	791,542

Table 2: Distribution of word forms by POS

[5] This category includes the following POS: adposition, article, auxiliary, cardinalNumeral, conjunction, coordinatingConjunction, definiteArticle, demonstrativeDeterminer, demonstrativePronoun, determiner, exclamativeDeterminer, exclamativePronoun, fusedPreposition, indefiniteArticle, indefiniteDeterminer, indefinitePronoun, interjection, interrogativeAdverb, interrogativeDeterminer, interrogativePronoun, numeral, numeralDeterminer, numeralPronoun, particle, personalPronoun, possessiveAdjective, possessiveDeterminer, possessivePronoun, pronoun, relativeDeterminer, relativePronoun, subordinatingConjunction.

As for the data related to the semantic layer, CompL-it describes 55,713 word senses connected to each other through 137 types of semantic relations, totalling 86,577 instances. Table 3 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	20,255	-
total	86,577	

Table 3: Distribution of semantic relations instances.

To provide an overview of the dimensions and richness of linguistic information conveyed by CompL-it, we finally present, in Table 4, a comparison with other lexical resources available for Italian[6].

[6] Data updated at the time of writing.

	entries	forms	senses/synsets	semantic relations instances	semantic relations types
LexicO	71,021	469,708	56,870 senses	89,340	137
IWN	48,416	-	49,350 synsets	138,385	83
MWN	41,491	-	32,673 synsets	45,593	14
OmegaWiki	30,258[7]	-	23,417 senses	66,005[8]	41
SIMPLELex-IT	7,022	26,560	-	-	-
Morph-it!	34,968	504,906	-	-	-
CompL-it	101,795	791,541	56,870 senses	86,577[9]	137

Table 4: Concise comparison of some of the main freely available resources containing lexical data for the Italian language.

5. CompL-it: access and use

The resource, in addition to being available for download, can be queried through a dedicated web interface (KLAB n.d.). This interface, shown in Figure 3, allows the user to select a series of precompiled SPARQL queries (visible on the left), modify one of them using the right panel, or formulate a new query from scratch.

As an example, the figure includes a precompiled query that allows for displaying all meanings of the verb *fare* (to do). If selected, the interface queries the resource and returns 7 senses of that verb, displaying their definitions and some examples. Using the corresponding SPARQL query shown in the right panel, it is possible to modify the label *fare* (highlighted in the figure) to insert another Italian verb, click the execute query button at the top right, and view the meanings of that verb in CompL-it.

In addition to its presentation as a linguistic resource in itself, as it is freely distributed, numerically rich and conforms to Linked Open Data standards, CompL-it can also be described in relation to the uses that can be made of it for information management and retrieval tasks.

For instance, a computational lexicon can be used in full-text search approaches in which queries can fully exploit the morphological information contained therein and the complex and articulated system of semantic rela-

[7] This number includes both dictionary entries and encyclopaedic data (such as named entities).

[8] Semantic relations are defined between "concepts" and not between senses of a specific language.

[9] There are fewer instances of semantic relations in CompL-it than in LexicO because proper nouns were not extracted from the latter resource (as specified at the beginning of section 4) and the associated semantic relations were excluded with them.

tions between the words of the lexicon (especially synonymy and hyponymy). Better results are naturally obtained if searches are carried out on linguistically pre-analysed texts (at least with POS tagging) to reduce ambiguity in the results. In the following section we provide an example of a full-text search supported by a computational lexicon. For more details, see (Giovannetti et al. 2022).

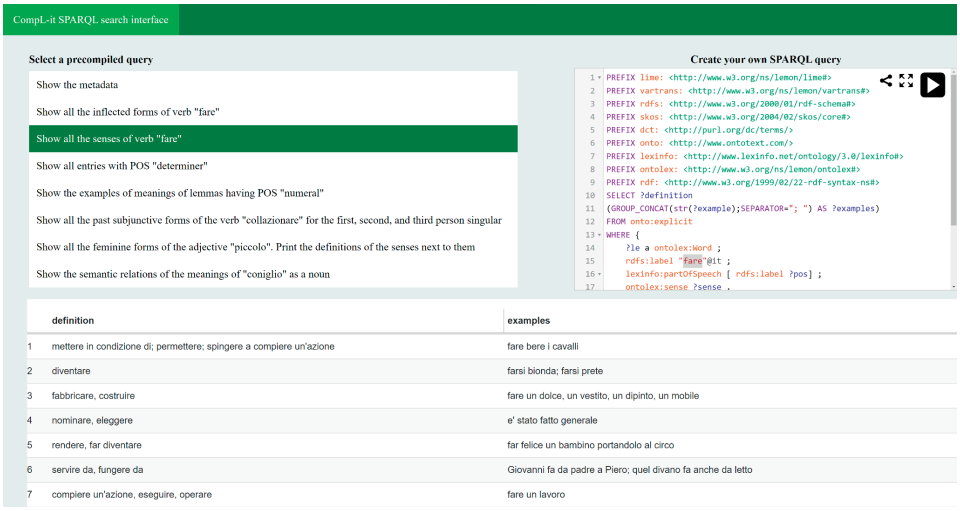


Figure 3: The query interface with the example of search for the senses of the Italian verb fare (to do).

Finally, we would also like to emphasise that representing data according to the LOD paradigm can bring some advantages, namely: (i) the federation mechanism with other datasets potentially allows the integration and improvement of queries results for more enriched searches, e.g., linking with etymological datasets (see Section 6); (ii) the addition of a semantic layer to the data through ontologies, allows the implicit knowledge in the dataset to be inferred and exploited in queries to the text, e.g., exploiting the transitivity of synonymy or hypernymy.

5.1. An example: lexicon-based search of the Babylonian Talmud

In this example, we show a query of the Italian translation of the Babylonian Talmud. This translation is being carried out by expert translators in the context of the Babylonian Talmud Translation Project (PTTB n.d.) using *Traduco*, a computer-aided translation tool developed at the CNR-ILC (Giovannetti et al. 2016). The search function of the tool allows the text to be accessed with complex queries, which can include information conveyed by the Italian lexicon related to morphology and semantics.

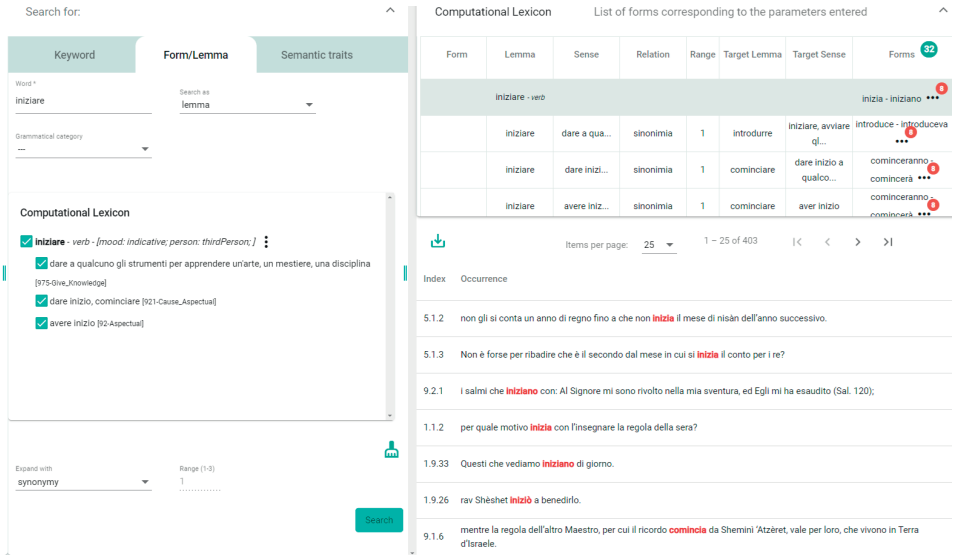


Figure 4: An example of query and relative results with the verb iniziare.

Considering the lemma *iniziare* (to start) it is possible to insert both morphological and semantic traits into the query (Fig. 4). By adding, for example, a restriction on the indicative mood and the third person, it is possible to extract all the contexts of the Talmud with the forms of the verb *iniziare* characterised by these traits. On the semantic side, we can also expand the query to include all lemmas having at least one sense as synonym to one of the (three) senses available in the lexicon for *iniziare*: the lexicon returns *introdurre* (to introduce) and *cominciare* (to begin). With the aforementioned morphological and semantic restrictions, the system is able to return Talmudic contexts containing, for example, the form *inizia* (singular, present indicative), but also *comincia*, *cominciano* (plural, present indicative), and *iniziò* (singular, past indicative).

6. Conclusions and perspectives

In this article, we presented CompL-it, a new computational lexicon for contemporary Italian. In the first part, we described the state of the art for this type of resource and outlined a landscape that, although rich, presents some challenges, such as the heterogeneity of data formats and linguistic models. Next, we described the three resources from which the lexicon was constructed: a computational lexicon (LexicO), a lemmatised form list obtained from a morphological analyser (M-GLF) and a set of treebanks. The three sources were based on different formats and models, which made it necessary

to standardise the data, also in accordance with the standards used by the Linguistic Linked Open Data community. Standardisation was followed by a conversion phase, which led to the final version of the lexicon in the form of Linguistic Linked Open Data according to the OntoLex-Lemon model.

We have also described the resource on a quantitative basis, also comparing CompL-it with some of the lexicographic resources available for Italian. The lexicon has been released as an open resource, is freely downloadable and can be consulted through a SPARQL interface. Finally, it was shown, through an example of a search on the Italian text of the Babylonian Talmud, how such a resource can be usefully exploited to provide linguistic-semantic access to textual corpora.

By its very nature, the editing of a lexicon can never be called a finished work. In the immediate future, CompL-it will first be further enriched in the semantic layer from the data that have not yet been extracted from LexicO (including templates and semantic traits). Subsequently, a merging methodology similar to the one adopted for morphology will also be applied to the semantic layer, in particular by considering available semantic resources such as ItalWordNet. A version of CompL-it will also be released, albeit limited to the morphological layer, conforming to the Universal Dependencies model. The resource will also be further extended to include cliticised forms, multiword forms and forms generated with suffixes, as in the case of *papà-papino* (dad, daddy). Finally, other linguistic layers will be considered, such as syntax (including data related to syntax-semantics interface) and phonetics, starting by leveraging on such data already available in LexicO.

Acknowledgement

Scientific publication produced thanks to the agreement between the National Research Council – Institute of Computational Linguistics and the PTTB S.c.a r.l. – Babylonian Talmud Translation Project.

References

BabelNet. n.d. “BabelNet | Il Più Grande Dizionario Enciclopedico e Rete Semantica Multilingue.” Accessed December 3, 2024. https://babelnet.org/.

Bamman, David, and Gregory Crane. 2010. “Computational Linguistics and Classical Lexicography.” In *Changing the Center of Gravity*, edited by Melissa Terras and Gregory Crane, 297-322. Gorgias Press. https://doi.org/10.31826/9781463219222-015.

Bartolini, Roberto. 2016. “IWN-LOD.” http://hdl.handle.net/20.500.11752/ILC-66.

Basili, Roberto, Silvia Brambilla, Danilo Croce, and Fabio Tamburini. 2017. "Developing a Large Scale FrameNet for Italian: The IFrameNet Experience." In *Proceedings of the Fourth Italian Conference on Computational Linguistics CLiC-It 2017*, edited by Roberto Basili, Malvina Nissim and Giorgio Satta, 59-64. Torino: Accademia University Press. https://doi.org/10.4000/books.aaccademia.2364.

Battista, Marco, and Vito Pirrelli. 1999. "Una Piattaforma di Morfologia Computazionale per l'analisi e la Generazione delle Parole Italiane." ILC-CNR Technical Report.

Bel, Nuria, Federica Busa, Nicoletta Calzolari, Elisabetta Gola, et al. 2000. "SIMPLE: A General Framework for the Development of Multilingual Lexicons." In *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC'00)*, edited by M. Gavrilidou, G. Carayannis, S. Markantonatou, S. Piperidis and G. Stainhauer. Athens, Greece: European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2000/pdf/61.pdf.

Brown, Susan Windisch, Julia Bonn, Ghazaleh Kazeminejad, Annie Zaenen, James Pustejovsky, and Martha Palmer. 2022. "Semantic Representations for NLP Using VerbNet and the Generative Lexicon." *Frontiers in Artificial Intelligence* 5 (April):821697. https://doi.org/10.3389/frai.2022.821697.

Chen, Hsinchun, Tak Yim, David Fye, and Bruce Schatz. 1995. "Automatic Thesaurus Generation for an Electronic Community System." *Journal of the American Society for Information Science* 46 (3): 175-93.

Chiarcos, Christian, and Maria Sukhareva. 2015. "OLiA – Ontologies of Linguistic Annotation." Edited by Sebastian Hellmann, Steven Moran, Martin Brümmer, and John P. McCrae. *Semantic Web* 6 (4): 379-86. https://doi.org/10.3233/SW-140167.

Chiarcos, Christian, Elena-Simona Apostol, Besim Kabashi, and Ciprian-Octavian Truică. 2022. "Modelling Frequency, Attestation, and Corpus-Based Information with OntoLex-FrAC." In *Proceedings of the 29th International Conference on Computational Linguistics*, edited by Nicoletta Calzolari, Chu-Ren Huang, Hansaem Kim, James Pustejovsky, Leo Wanner, Key-Sun Choi, Pum-Mo Ryu, et al., 4018-27. Gyeongju, Republic of Korea: International Committee on Computational Linguistics. https://aclanthology.org/2022.coling-1.353.

Chiari, Isabella. 2012. "Il Dato Empirico in Lessicografia: Dizionari Tradizionali e Collaborativi a Confronto." *Bollettino Di Italianistica* II (January): 94-125.

Cimiano, Philipp, Christian Chiarcos, John P. McCrae, and Jorge Gracia. 2020. *Linguistic Linked Data: Representation, Generation and Applications*. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-30225-2.

Cimiano, Philipp, 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. https://doi.org/10.1016/j.websem.2010.11.001.

CLARIN. n.d. "ParlaMint: Comparable and Interoperable Parliamentary Corpora | CLARIN ERIC." Accessed December 3, 2024. https://www.clarin.eu/parlamint.

CLARIN-IT. n.d.a. "CompL-It." Accessed December 3, 2024. https://dspace-clarin-it.ilc.cnr.it/repository/xmlui/handle/20.500.11752/ILC-1007.

CLARIN-IT. n.d.b. "LexicO." Accessed December 3, 2024. https://dspace-clarin-it.ilc.cnr.it/repository/xmlui/handle/20.500.11752/ILC-977.

CLARIN-IT. n.d.c. "MAGIC - Generated Lemmatized Forms." Accessed December 3, 2024. https://dspace-clarin-it.ilc.cnr.it/repository/xmlui/handle/20.500.11752/ILC-1002.

CLARIN VLO. n.d. "Virtual Language Observatory." Accessed September 30, 2024. https://www.clarin.eu/content/virtual-language-observatory-vlo.

Dankova, Klara, Maria Teresa Zanola, and Silvia Calvi. 2022. "Pan-Latin Textile Fibres Vocabulary." http://hdl.handle.net/20.500.11752/OPEN-975.

DatCatInfo. n.d. "Welcome to DatCatInfo." Accessed December 3, 2024. https://datcatinfo.net/.

De Mauro, Tullio. 1980. *Guida all'uso delle parole: parlare e scrivere semplice e preciso per capire e farsi capire*. Libri di base. Roma: Editori Riuniti.

De Mauro, Tullio, a cura di. 2016. *Il Nuovo Vocabolario Di Base Della Lingua Italiana*. December 23, 2016. https://www.dropbox.com/scl/fi/zg2y99xqik4k11nj19fgi/nuovovocabolariodibase.pdf?rlkey=s0uf8ggv11kf44ip6a2ldz16n&e=1&dl=0.

Del Gratta, Riccardo, Francesca Frontini, Anas Fahad Khan, and Monica Monachini. 2015. "Converting the PAROLE SIMPLE CLIPS Lexicon into RDF with Lemon." *Semantic Web* 6 (4): 387-92. https://doi.org/10.3233/SW-140168.

ELEXIS. n.d. "ELEXIS European Lexicographic Infrastructure." Accessed September 30, 2024. https://elex.is/.

Francopoulo, Gil, Monte George, Nicoletta Calzolari, et al. 2006. "Lexical Markup Framework (LMF)." In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, edited by Nicoletta Calzolari, Khalid Choukri, Aldo Gangemi, Bente Maegaard, Joseph Mariani, Jan Odijk, and Daniel Tapias. Genoa, Italy: European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2006/pdf/577_pdf.pdf.

Frontini, Francesca, Riccardo Del Gratta, and Monica Monachini. 2016. "Geodomain WordNet ITA ENG V 1.0." http://hdl.handle.net/20.500.11752/ILC-68.

Giovannetti, Emiliano, Davide Albanesi, Andrea Bellandi, and Giulia Benotto. 2016. "Traduco: A Collaborative Web-Based CAT Environment for the Interpretation and Translation of Texts." *Digital Scholarship in the Humanities* 32 (suppl_1): i47-62. https://doi.org/10.1093/llc/fqw054.

Giovannetti, Emiliano, Davide Albanesi, Andrea Bellandi, Simone Marchi, Mafalda Papini, and Flavia Sciolette. 2022. "The Role of a Computational Lexicon for Query Expansion in Full-Text Search." In *Proceedings of the Eighth Italian Conference on Computational Linguistics CliC-It 2021*, edited by Elisabetta Fersini, Marco Passarotti, and Viviana Patti, 162-68. Accademia University Press. https://doi.org/10.4000/books.aaccademia.10638.

Github. n.d.a. "The Ontolex Module for Frequency, Attestation and Corpus Information." Accessed December 3, 2024. https://github.com/acoli-repo/frac-addenda/blob/master/index.md.

Github. n.d.b. "CompL-It Mapping Tables." Accessed December 3, 2024. https://github.com/klab-ilc-cnr/Tables-for-mapping-of-Italian-lexicon-CompL-it.

Global WordNet Association. n.d. "Main Page." Accessed December 3, 2024. http://globalwordnet.org/.

Grella, Matteo. 2018a. "Italian Content Words V3." http://hdl.handle.net/11372/LRT-2894.

Grella, Matteo. 2018b. "Italian Function Words V3." http://hdl.handle.net/11372/LRT-2893.

Hmeidi, Ismail, Mahmoud Al-Ayyoub, Nizar A. Mahyoub, and Mohammed A. Shehab. 2016. "A Lexicon Based Approach for Classifying Arabic Multi-Labeled Text." *International Journal of Web Information Systems* 12 (4): 504-32. https://doi.org/10.1108/IJWIS-01-2016-0002.

Hodge, Gail. 2000. *Systems of Knowledge Organization for Digital Libraries: Beyond Traditional Authority Files*. Washington, DC: Digital Library Federation, Council on Library and Information Resources.

ILC4CLARIN CNR. 2016. "PAROLE-SIMPLE-CLIPS." http://hdl.handle.net/20.500.11752/ILC-88.

Khan, Fahad, Ana Salgado, Isuri Anuradha, et al. 2024. "CHAMUÇA: Towards a Linked Data Language Resource of Portuguese Borrowings in Asian Languages." In *Proceedings of the 9th Workshop on Linked Data in Linguistics @ LREC-COLING 2024*, edited by Christian Chiarcos, Katerina Gkirtzou, Maxim Ionov, Fahad Khan, John P. McCrae, Elena Montiel Ponsoda, and Patricia Martín Chozas, 44-48. Torino, Italia: ELRA and ICCL. https://aclanthology.org/2024.ldl-1.6.

KLAB. n.d. "CompL-It SPARQL Search Interface." Accessed December 3, 2024. https://klab.ilc.cnr.it/CompL-it-SPARQL-interface/.

Kyjánek, Lukáš, Zdeněk Žabokrtský, Jonáš Vidra, and Magda Ševčíková. 2021. "Universal Derivations v1.1." http://hdl.handle.net/11234/1-3247.

LexInfo. n.d. "About the Ontology." Accessed September 30, 2024. https://lexinfo.net/.

LLOD. n.d. "Linguistic Linked Open Data." Accessed December 3, 2024. https://linguistic-lod.org/.

Mallia, Michele, Michela Bandini, Andrea Bellandi, et al. 2024. "DigItAnt: A Platform for Creating, Linking and Exploiting LOD Lexica with Heterogeneous Resources." In *Proceedings of the 9th Workshop on Linked Data in Linguistics @ LREC-COLING 2024*, edited by Christian Chiarcos, Katerina Gkirtzou, Maxim Ionov, Fahad Khan, John P. McCrae, Elena Montiel Ponsoda, and Patricia Martín Chozas, 55-65. Torino, Italia: ELRA and ICCL. https://aclanthology.org/2024.ldl-1.8.

Malmgren, Sven-Göran. 1988. "On Regular Polysemy in Swedish." In *Studies in Computer-Aided Lexicology*, 179-200. Data Linguistica 18. Stockholm: Almqvist & Wiksell.

Mambrini, Francesco, and Marco Carlo Passarotti. 2023. "The LiLa Lemma Bank: A Knowledge Base of Latin Canonical Forms." *Journal of Open Humanities Data* 9 (November):1-5. https://doi.org/10.5334/johd.145.

Mazzei, Alessandro. 2016. "Building a Computational Lexicon by Using SQL." In *Proceedings of the Third Italian Conference on Computational Linguistics CLiC-It 2016*, 200-04. Napoli: Accademia University Press. https://doi.org/10.4000/books.aaccademia.1808.

Meijssen, Gerard. 2014. "OmegaWiki." http://hdl.handle.net/11372/LRT-853.

Miller, George A. 1995. "WordNet: A Lexical Database for English." *Communications of the ACM* 38 (11): 39-41. https://doi.org/10.1145/219717.219748.

Montiel-Ponsoda, Elena, Wim Peters, Mauricio Espinoza, Asunción Gómez-Pérez, and Margherita Sini. 2008. "Multilingual and Localization Support for Ontologies." Technical Report 2.4.2. http://neon-project.org/deliverables/WP2/NeOn_2008_D242.pdf.

Morph-it! 2018. "Resources:Morph-It." Last Modified May 03. https://docs.sslmit.unibo.it/doku.php?id=resources:morph-it.

MultiWordNet. n.d. "NLP Research Group - MultiWordNet." Accessed December 3, 2024. https://nlplab.fbk.eu/tools-and-resources/lexical-resources-and-corpora/multiwordnet.

Navigli, Roberto, and Simone Paolo Ponzetto. 2012. "BabelNet: The Automatic Construction, Evaluation and Application of a Wide-Coverage Multilingual Semantic Network." *Artificial Intelligence* 193 (December): 217-50. https://doi.org/10.1016/j.artint.2012.07.001.

OLiA. n.d. "Ontologies of Linguistic Annotation (OLiA) | Olia." Accessed December 3, 2024. https://acoli-repo.github.io/olia/.

Ontotext. n.d. "Ontotext GraphDB." Accessed December 3, 2024. https://www.ontotext.com/products/graphdb/.

Passarotti, Marco Carlo, and Francesco Mambrini. 2021. "Linking Latin: Interoperable Lexical Resources in the LiLa Project." In *Building New Resources for Historical Linguistics*, edited by Erica Biagetti, Chiara Zanchi and Silvia Luraghi, 103-24. https://doi.org/10.5281/zenodo.5994271.

Pianta, Emanuele, Luisa Bentivogli, and Christian Girardi. 2002. "MultiWordNet: Developing an Aligned Multilingual Database." In *Proceedings of the First International Conference on Global WordNet*.

Pirrelli, Vito, and Marco Battista. 2000. "The Paradigmatic Dimension of Stem Allomorphy in Italian Verb Inflection." *Italian Journal of Linguistics* 12 (2): 307-80.

Prakash, T. Nikil, and Amalanathan Aloysius. 2021. "Textual Sentiment Analysis Using Lexicon Based Approaches." *Annals of the Romanian Society for Cell Biology* 25 (4): 9878–85.

Princeton University. n.d. "WordNet." Accessed December 3, 2024. https://wordnet.princeton.edu/.

PTTB. n.d. "Progetto Traduzione Talmud Babilonese." Accessed December 3, 2024. https://www.talmud.it/it/.

Pustejovsky, James. 1995. *The Generative Lexicon*. The MIT Press. https://doi.org/10.7551/mitpress/3225.001.0001.

Realiter. n.d. "Home | Realiter." Accessed December 3, 2024. https://www.realiter.net/.

Roventini, Adriana, Antonietta Alonge, Francesca Bertagna, et al. 2003. "'ItalWordNet': Building a Large Semantic Database for the Automatic Treatment of Italian." *Linguistica computazionale: XVIII/XIX, 1998/1999*, 745-91. https://doi.org/10.1400/18178.

Roventini, Adriana, Rita Marinelli, and Francesca Bertagna. 2016. "ItalWordNet v.2." http://hdl.handle.net/20.500.11752/ILC-62.

Ruimy, Nilda, Monica Monachini, Raffaella Distante, et al. 2002. "CLIPS, a Multi-Level Italian Computational Lexicon: A Glimpse to Data." In *Proceedings of the Third International Conference on Language Resources and Evaluation, LREC 2002, May 29-31, 2002, Las Palmas, Canary Islands, Spain*. European Language Resources Association. http://www.lrec-conf.org/proceedings/lrec2002/sumarios/197.htm.

Sabatini, Francesco. 2006. "La Storia dell'Italiano nella Prospettiva della Corpus Linguistics." In *Proceedings of the 12th EURALEX International Congress*, edited by Cristina Onesti, Elisa Corino and Carla Marello, 31-37. Torino: Edizioni dell'Orso.

Sanguinetti, Manuela, and Cristina Bosco. 2015. "PartTUT: The Turin University Parallel Treebank." In *Harmonization and Development of Resources and Tools for Italian Natural Language Processing within the PARLI Project*, edited by Roberto Basili, Cristina Bosco, Rodolfo Delmonte, Alessandro Moschitti, and Maria Simi, 589: 51-69. Studies in Computational Intelligence. Springer. https://doi.org/10.1007/978-3-319-14206-7_3.

Sciolette, Flavia, Emiliano Giovannetti, and Simone Marchi. 2023. "LexicO: An Italian Computational Lexicon Derived from Parole-Simple-Clips." *Umanistica Digitale* 7 (15): 169-93. https://doi.org/10.6092/issn.2532-8816/15176.

Sciolette, Flavia. 2024. "Modeling Linking between Text and Lexicon with OntoLex-Lemon: A Case Study of Computational Terminology for the Babylonian Talmud." In *Proceedings of the 9th Workshop on Linked Data in Linguistics @ LREC-COLING 2024*, edited by Christian Chiarcos, Katerina Gkirtzou, Maxim Ionov, Fahad Khan, John P. McCrae, Elena Montiel Ponsoda, and Patricia Martín Chozas, 103-7. Torino, Italia: ELRA and ICCL. https://aclanthology.org/2024.ldl-1.13.

Sérasset, Gilles. 2015. "DBnary: Wiktionary as a Lemon-Based Multilingual Lexical Resource in RDF." *Semantic Web* 6 (4): 355-61. https://doi.org/10.3233/SW-140147.

Shiri, Ali. 2015. "Semantic Access and Exploration in Cultural Heritage Digital Libraries." In *Cultural Heritage Information: Access and Management*, edited by Ian Ruthven and Gobinda G. Chowdhury, 177-96. Facet Publishing.

Simi, Maria, Cristina Bosco, and Simonetta Montemagni. 2014. "Less Is More? Towards a Reduced Inventory of Categories for Training a Parser for the Italian Stanford Dependencies." In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, edited by Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, 83-90. Reykjavik, Iceland: European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2014/pdf/818_Paper.pdf.

SimpleLEX-IT. n.d. "Alexmazzei/SimpleLEX-IT: SimpleLEX-IT Is the Computational Lexicon Employed into SimpleNLG-IT." Accessed December 3, 2024. https://github.com/alexmazzei/SimpleLEX-IT.

Smith, Jocelyn C. 1997. "The Use of Lexicons in Information Retrieval in Legal Databases." In *Proceedings of the Sixth International Conference on Artificial Intelligence and Law - ICAIL '97*, 29-38. Melbourne, Australia: ACM Press. https://doi.org/10.1145/261618.261625.

Thompson, Paul, John McNaught, Simonetta Montemagni, et al. 2011. "The BioLexicon: A Large-Scale Terminological Resource for Biomedical Text Mining." *BMC Bioinformatics* 12 (1): 397. https://doi.org/10.1186/1471-2105-12-397.

Universal Dependencies. n.d.a. "Introduction." Accessed December 3, 2024. https://universaldependencies.org/it/overview/introduction.html.

Universal Dependencies. n.d.b. "Universal Dependencies." Accessed December 3, 2024. https://universaldependencies.org/.

Universal Dependencies. n.d.c. "UD Guidelines." Accessed December 3, 2024. https://universaldependencies.org/guidelines.html.

Universal Dependencies. n.d.d. "UD_Italian-ISDT." Accessed December 3, 2024. https://github.com/UniversalDependencies/UD_Italian-ISDT/tree/master.

Universal Dependencies. n.d.e. "UD_Italian-VIT." Accessed December 3, 2024. https://github.com/UniversalDependencies/UD_Italian-VIT/.

Universal Dependencies. n.d.f. "CoNLL-U Format." Accessed December 3, 2024. https://universaldependencies.org/format.html.

Universal Dependencies. n.d.g. "UD_Italian-ParTUT." Accessed December 3, 2024. https://github.com/UniversalDependencies/UD_Italian-ParTUT/.

Universal Dependencies. n.d.h. "UD_Italian-ParlaMint." Accessed December 3, 2024. https://github.com/UniversalDependencies/UD_Italian-ParlaMint/.

Vetere, Guido, Alessandro Oltramari, Isabella Chiari, Elisabetta Jezek, Laure Vieu, and Fabio Massimo Zanzotto. 2011. "Senso Comune, an Open Knowledge Base of Italian Language." *Traitement Automatique Des Langues* 52 (3): 217-43.

Villegas, Marta, and Núria Bel. 2015. "PAROLE/SIMPLE 'Lemon' Ontology and Lexicons." Edited by Sebastian Hellmann, Steven Moran, Martin Brümmer, and John McCrae. *Semantic Web* 6 (4): 363-69. https://doi.org/10.3233/SW-140148.

W3C. 2005. "UsingSeeAlso - W3C Wiki." Last Modified January 13. https://www.w3.org/wiki/UsingSeeAlso.

W3C. 2013. "SPARQL 1.1 Overview." https://www.w3.org/TR/sparql11-overview/.

W3C. 2014. "RDF 1.1 Turtle." https://www.w3.org/TR/turtle/.

W3C. 2016. "Lexicon Model for Ontologies: Community Report, 10 May 2016." https://www.w3.org/2016/05/ontolex/.

WikiMedia. 2022. "OmegaWiki - Meta." Last Modified December 03. https://meta.wikimedia.org/wiki/OmegaWiki.

Wikipedia. 2024a. "Verbi Incoativi - Wikipedia." Last Modified January 26. https://it.wikipedia.org/wiki/Verbi_incoativi.

Wikipedia. 2024b. "Verbi Irregolari Italiani - Wikipedia." Last Modified November 6. https://it.wikipedia.org/wiki/Verbi_irregolari_italiani.

Žabokrtský, Zdeněk, Nyati Bafna, Jan Bodnár, et al. 2022. "Universal Segmentations 1.0 (UniSegments 1.0)." http://hdl.handle.net/11234/1-4629.

Zanchetta, Eros, and Marco Baroni. 2005. "Morph-It! A Free Corpus-Based Morphological Resource for the Italian Language." In *Proceedings of Corpus Linguistics Conference Series 2005 (ISSN 1747-9398)*, 1: 1-12. University of Birmingham.

Zanola, Maria Teresa, Klara Dankova, Claudio Grimaldi, and Anna Serpente. 2023. "Pan-Latin Lexicon of Collars and Sleeves in Fashion and Costume." http://hdl.handle.net/20.500.11752/OPEN-987.

AIDAinformazioni

Rivista semestrale di Scienze dell'Informazione

Anno 42

N. 3-4 – luglio-dicembre 2024

Contributi

Rubriche

mundaneum

In copertina
Disegno di Paul Otlet, Collections Mundaneum, centre d'Archives, Mons (Belgique).

ISBN 979-12-5965-456-4

ISSN 1121-0095