Umanistica Digitale – ISSN: 2532-8816 – n. 15, 2023
F. Sciolette, S. Marchi, E. Giovannetti– *LexicO: an Italian Computational Lexicon*
DOI: http://doi.org/10.6092/issn.2532-8816/15176

# LexicO: an Italian Computational Lexicon derived from Parole-Simple-Clips

Flavia Sciolette

Istituto di Linguistica Computazionale "A. Zampolli", CNR, Pisa, Italia
flavia.sciolette@ilc.cnr.it

Simone Marchi

Istituto di Linguistica Computazionale "A. Zampolli", CNR, Pisa, Italia
simone.marchi@ilc.cnr.it

Emiliano Giovannetti

Istituto di Linguistica Computazionale "A. Zampolli", CNR, Pisa, Italia
emiliano.giovannetti@ilc.cnr.it

## Abstract

Parole-Simple-Clips (PSC) is a computational lexicon of the Italian language, developed from 1996 to 2003 by the Institute of Computational Linguistics of the Italian National Research Council (ILC-CNR) in the context of national and European projects. The PSC resource is strongly structured, rich of data, and, for its features, may provide an edge if used in the support of text retrieval related tasks, such as full-text search. However, the lexicon still appears incomplete and presents some redundant, erroneous and missing data. This paper documents the first steps undertaken for the creation of LexicO, an Italian computational lexicon built upon PSC starting from an in depth analysis of its four linguistic layers (semantic, syntactic, morphological, and phonological) in which it is structured. As a result of this work, LexicO has been released and made freely available.

**Keyword:** Computational Lexicon; Parole-Simple-Clips; Linguistic Resources; Full-text Search; LexicO

*Parole-Simple-Clips (PSC) è un lessico computazionale per l'Italiano, sviluppato tra il 1996 e il 2003 presso l'Istituto di Linguistica Computazionale del Consiglio Nazionale delle Ricerche (ILC-CNR) nell'ambito di progetti nazionali ed Europei. PSC è una risorsa ricca, fortemente strutturata e, per le sue caratteristiche, può fornire un vantaggio in task di text retrieval come la ricerca full-text. PSC appare tuttavia incompleta in alcune sue sezioni e presenta dati ridondanti, erronei o mancanti.*

*Questo contributo descrive i primi passi compiuti per la creazione di LexicO, un lessico computazionale italiano costruito sulla base di PSC a partire da un'analisi approfondita dei suoi quattro livelli linguistici (semantico, sintattico, morfologico e fonologico). A seguito di questi interventi, LexicO è stato rilasciato e reso liberamente disponibile.*

**Parole chiave:** Lessico Computazionale; Parole-Simple-Clips; Risorse Linguistiche; Ricerca sul testo; LexicO

## Introduction

Despite the availability of several linguistic models and projects focussing on the systematisation of linguistic data (see Related works), just a few computational lexical resources seems to be available specifically for the Italian language[1]. This is understandable, if we consider how the construction of a large-scale lexical dataset, especially if structured by following a full-fledged linguistic model, can be extremely time consuming and require high-level linguistic skills. The resources most employed in different computational tasks are mainly the multilingual Wordnets; to the best of our knowledge, other lexicons appear under-used[2] for computational processes, in comparison with them. Due for their complexity, however, they could be profitably exploited to provide support in different tasks related to text and linguistic processing, such as information retrieval, word sense disambiguation, named entity recognition, full text search and, more in general, any task that may take advantage from the availability of structured linguistic information.

This work documents the analysis of a portion of a fundamental linguistic resource of Italian: the computational lexicon known as "PAROLE-SIMPLE-CLIPS" (PSC). The critical issues found during the analysis have led to the creation of LexicO, a new Italian computational lexicon conceived to gather data coming from PSC as they will be analysed and revised over time. PSC was built from 1996 to 2003 at the Institute of Computational Linguistics of the Italian National Research Council (ILC-CNR) in the context of national and European projects [25].

Thanks to its depth, richness, and the rigorous linguistic model it is built upon, PSC is a mine of information in itself, but it may also be productively exploited for many tasks, both of scholarly

1 In the repository of Italian node of CLARIN (Common Language Resources and Technology Infrastructure) - the research infrastructure ensuring availability and accessibility of language resources and tools - the search for "lexicalConceptualResource" as "type" and "Italian" as "language" provides 12 results. At the time of writing, the main detected resources are Parole-Simple-Clips (PSC) and IWN (ItalWordNet), both described in Related works; the other results include i) five resources derived from the first two; ii) five domain terminologies (https://dspace-clarin-it.ilc.cnr.it/repository/xmlui/). Expanding the search to all the Virtual Language Observatory, we can find other two resources (see Related works) of further public 20 in the repository. These 20 resources include multilingual domain terminologies and four previous stages of other resources (https://vlo.clarin.eu/?4)

2 For the Italian linguistic domain, an original use of resource for a computational task is represented in [28] as quantitative methods to detect polysemy, with the employ of PSC. The uses of domain terminologies for recognition and automatic extraction of terms are more represented (for example in the Automatic Misogyny Identification, task within Evalita 2018, see [8] for an overview of the methods and use of lexicons).In every case, this kind of studies appears less-represented in the main repositories (Wos, Scopus, Zenodo, Google Scholar).

and computational nature, such as, as in the example described in The use of a computational lexicon in an application of full-text search: the case of querying of the babylonian Talmud, for advanced full-text search. In particular, we here describe the analysis process and the consequent discovery of a set of redundant,[3] erroneous and missing data in PSC. This work appeared necessary after the experiments of full-text search described in [11] (see for example the case of missing synonyms reported in The use of a computational lexicon in an application of full-text search: the case of querying of the babylonian Talmud). The paper is organised as follows: the Related works is dedicated to related works about lexical resources for Italian with references to works involving the application of linguistic resources in information retrieval contexts; the The use of a computational lexicon in an application of full-text search: the case of querying of the babylonian Talmud introduces the case-study with which we started to experiment the use of the resource; the The PSC resource describes the underlying model and the current data stored in PSC; the Discovery and processing of redundant, erroneous and missing data shows the approach and outcome of the discovery and processing of redundant, erroneous, and missing data; the Conclusions and perspectives introduces the next steps planned for further updates of LexicO.

## Related works

Italian linguistics has always shown an interest for lexical data [27], an inclination that has been progressively encouraged by the increasing availability of computational technologies and the development of many "corpus-based" resources. Many projects involving corpora (monolingual, parallel, domain-specific) have flourished and both digitised traditional dictionaries and computational dictionaries have taken advantage of them (e.g. to calculate the frequency of lemmas, the adherence of the use, or the coverage) [5], as well as for applications in different tasks, like "machine translation, computer-aided translation, human-aided machine translation" [6], full-text search, topic modelling or information retrieval, as discussed in the following.

Besides PSC, just a few digital resources including lexical information are available for Italian. In the following we list those that met the following three criteria: i) free availability; ii) presence of structured linguistic information, also specifically for Italian (thus excluding terminologies of specific domains), and iii) structuring in a machine-readable format.

The first set of resources we took into account are those based on WordNet [7]. An important linguistic resource for Italian is ItalWordNet[4] (IWN) [24], firstly developed for EuroWordNet[5] and successively extended.

Successively, we cite MultiWordNet,[6] developed by the Bruno Kessler Center in Information and Communication Technology (FBK-ICT Irst Center); the alignment of the synsets belonging to the different languages populating the resource (including Italian) is strictly based on the

3 The term "redundant", in this work, indicates any kind of data in excess that needs to be removed.

4 http://www.ilc.cnr.it/iwndb/iwndb_php/

5 http://projects.illc.uva.nl/EuroWordNet/

6 http://multiwordnet.fbk.eu/english/home.php

semantic relations available in Princeton WordNet and the enrichment of the resource has been carried out with the help of semi-automatic techniques [21].

Amongst the lexicalized ontologies integrating WordNet, we here mention Babelnet,[7] a multilingual resource that includes Italian [20]. Babelnet is characterised by a huge amount of data, created by linking Wikipedia with Princeton WordNet and different many other resources: similarly to WordNet, Babelnet is organised into sets of synonyms, called "Babel synsets". Amongst the derived resource from Babelnet, we cite SyntagNet, a recent example of lexical-semantic combination database of syntagmatic relations, developed by Sapienza University of Rome in the context of ERC-Project MOUSSE[8]. This resource contains manually disambiguated 78.000 noun-verb and noun-noun lexical combinations of five languages: English, German, French, Spanish, and Italian. The 88.019 semantic combinations are linked with 20.626 WordNet synsets for 14.004 lemmas.

Another resource for Italian is Senso Comune,[9] a lexical knowledge base developed by University of Sapienza and CNR. The "core" of the resource is composed of 2.000 lexical units which have been manually associated with the concepts belonging to a foundational ontology of reference. This core of entries is freely available, while the complete resource is protected by copyright.

Though built more as support tools for NLP than as lexical resources, we also mention Italian Content Words [12] and Italian Function Words [13]. Both resources are conceived to help in pos-tagging and parsing, and - at the time of writing - they are distributed as UniMorph-annotated dictionaries[10] [18]. Regarding resources based on FrameNet,[11] it is worth mentioning the project of an Italian FrameNet, still under development [2], with 7.000 validated lexical units.

The growing attention to the creation of lexical resources is also reflected in the development of standards for the encoding of computational lexicons. At present, the *de facto* standard is OntoLex-Lemon[12], a model for lexical resources developed in the framework of Linked Open Data (LOD).

The effective use of lexical resources is basically limited to WordNet, though for different tasks. In the field of Information Retrieval, we cite SIRWWO (Semantic Information Retrieval using Wikipedia, WordNet, and domain Ontologies), a recent approach that combines multiple knowledge sources [14]; in [15] the authors describe the use of WordNet for text categorization of multi-label documents. With a specific focus on the task of Word Sense Disambiguation, we highlight the work of NLP Lab of Sapienza [3].

All the resources introduced in this section present different features, especially considering the model they adopt to describe linguistic information. Furthermore, some of them (such as Senso Comune) are not available in their entirety and thus they cannot be fully accessed and assessed. For these reasons, it is very difficult to do an in-depth and effective comparison among them.

7 https://babelnet.org/
8 http://syntagnet.org/
9 http://www.sensocomune.it/ [no longer working, 7/11/2022]
10 Universal Morphology (UniMorph) is a project that provides a universal schema for the annotation of morphology (https://unimorph.github.io/).
11 https://framenet.icsi.berkeley.edu/fndrupal/
12 https://www.w3.org/2016/05/ontolex/

However, to better substantiate our choice of exploiting PSC with respect to the other considered resources, in PSC compared to other resources for Italian language we will compare them on the basis of some selected portions of their data. Furthermore, as it will be shown in the The use of a computational lexicon in an application of full-text search: the case of querying of the babylonian Talmud, PSC presents features that can make its use particularly effective in text querying contexts. At the same time, however, this resource has some problems that need to be corrected before its employment, as highlighted in the experimentation described in the The use of a computational lexicon in an application of full-text search: the case of querying of the babylonian Talmud.

## The use of a computational lexicon in an application of full-text search: the case of querying of the babylonian Talmud

In this section, we introduce a case study where PSC has been exploited to query the text of the Italian translation of the babylonian Talmud, the primary source of Jewish religious law and theology. The translation is being carried out by domain-expert translators working in the context of the Babylonian Talmud Translation Project[13] using Traduco, a computer-assisted translation tool developed at ILC-CNR [10]. At the time of this writing, seven treatises of the Talmud have been published, and more than 1 million of Hebrew/Aramaic words have been translated. It is therefore a corpus of considerable dimensions covering a wide range of topics, the querying of which, for example by keyword, can prove to be extremely difficult. As a matter of fact, translators and editors of the Talmud often need to examine and compare different parts of the text in order to publish a coherent and homogeneous translation.

Traduco already includes two search functions, a standard keyword based search and a search based on the presence of manually inserted semantic annotation of different kinds, among which talmudic concepts, linguistic expressions, names of Rabbis, etc.

However, scholars working at the translation of the Talmud would often need to query the text combining different kinds of linguistic information - for example to find words belonging to a specific semantic field and having specific morphological constraints - a task that a simple keyword-based search cannot easily fulfil. For this reason, we decided to develop an advanced full-text search interface to the Talmud integrating a query expansion technique exploiting the PSC as a support lexicon. This interface adds to Traduco two new querying modalities: i) a search by form or lemma with constraints on morphological traits and the possibility of extending the search with semantic relations; ii) a search based on semantic traits.

The first kind of search is shown in Figure 1. A user can insert a word and indicate if it has to be interpreted as an inflected form or a lemma: in this second case, the system searches for all the relative forms.

13 https://www.talmud.it/

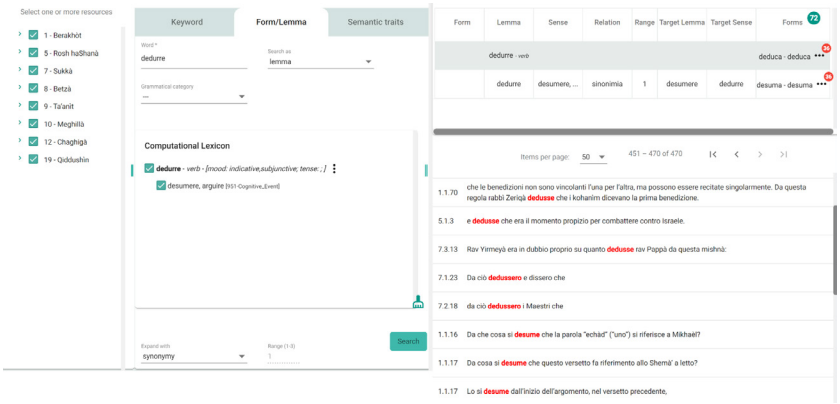


Figure 1: The form/lemma section of the interface for full-text search supported by PSC; the talmudic treatises, constituting the corpus, are shown on the left side, the query interface is placed at the centre and the results of the search are shown in the rightmost part.

The underlying system accesses the lexicon to retrieve all the lemmas, the relative inflected forms, and the lexical senses[14] corresponding to the indicated word. Queries can be enriched with filters on morphological features (e.g. on pos, gender, number, tense and mood for verbs) and extended with semantic relations (at the moment limited to hypernymy, hyponymy, and synonymy) to be applied to the selected senses. The right side of the interface shows, on top, the forms retrieved from the lexicon and used to expand the query, while, below, the list of occurrences of the searched words in their context. In the following example, as reported in Figure 1, we describe a case of propagation of morphological features to related senses. We consider the case of word *dedurre*, with the sense of "to presume": the richness of linguistic information in PSC appears particularly productive when the query involves verbal entries, since the search can be focussed on forms with specific moods, tenses, etc. With the insertion of the morphological constraints "indicative" and "subjunctive", a user can search for all the forms with these finite moods. In addition, with the expansion of the search to synonyms, the forms of *desumere* (a synonym of *dedurre* present in PSC) are added to the query.

During the exploitation of the resource, it has been observed how the results of some queries were incorrect, for example when searches were extended with synonymy. If a user searches for synonyms of word *desumere*, the results do not include the occurrences of *dedurre*. As a matter of fact, as shown in Figure 2, PSC includes a synonymic relation between *dedurre* and *desumere* (the solid arrow) but does not include the symmetric relation holding between *desumere* and *dedurre* (the dotted arrow). This is an example of missing data that motivated the update of the resource, as detailed in Discovery and processing of redundant, erroneous and missing data.

14 In this work, we use the term "sense" to define the meaning described in the resource as Semantic Units (SemUs, cfr. The PSC resource), according to the documentation of PSC: "Following the terminology of the GENELEX Model, word senses are encoded as Semantic Units or SemUs" [17]. From a theoretical point of view, in the documentation, the concept of "meaning" is associated with the Qualia Structure; the meaning of a word is the result of an interaction among different components; these components of meaning can coexist in the same concept (e.g. "apple" is an object from a formal point of view and it is a food, from a telic perspective).

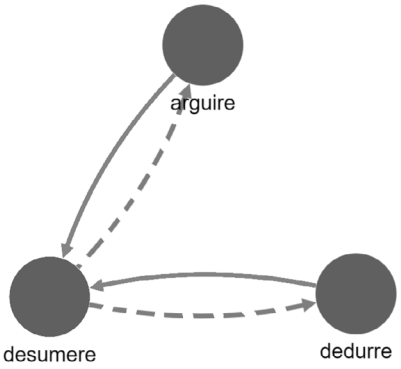


Figure 2: An example of missing data in PSC: the shown sense of verb "dedurre" is linked to the sense of "desumere" through a relation of synonymy, but the symmetric relation is missing (the dotted arrow); a similar case of asymmetry is present between verbs "desumere" and "arguire".

The second kind of search, involving semantic traits, is shown in Figure 3. In this example, the user selects the template[15] "Vegetal Entity", which is classified as a subclass of "Living Entity". Once the template is chosen, the system retrieves from the lexicon all the relative senses and shows them in a window to the right of the template hierarchy. It is then possible to select all the available 639 senses or just some of them. Finally, the user can run the search: the system composes the expanded query and retrieves 2.500 textual segments of the corpus containing words (both as lemmas and inflected forms) with senses referring to the semantic field of "Vegetal Entity" such as *frutti* (fruits), *pino* (pine), *alberi* (trees), and so on. The user may then further filter the result of the query by selecting just a "leaf" of the subtree under "Vegetal Entity", for example the template "Fruit".

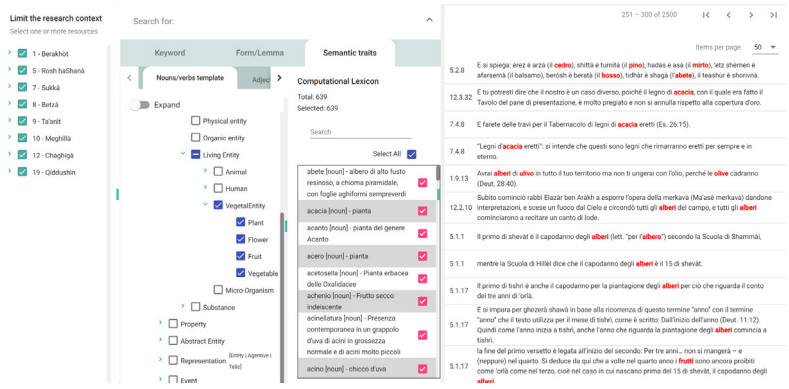


Figure 3: The semantic traits section of the interface, conceived for a more explorative search of the corpus. In the example, the user searched for occurrences of words the senses of which refer to the template "Vegetal Entity".

15 See The underlying linguistic model.

In this case, the system retrieves 63 senses from the lexicon for 398 occurrences on the corpus, such as *cedro* (citron), *dattero* (date), and *melagrana* (pomegranate). The results of this query do not give only purely quantitative data, but also allow qualitative considerations to be made: in this case we can note how the results contain just fruits widespread in the Mediterranean basin. More information on the search tool can be found in the already cited [11].

## The PSC resource

The Parole-Simple-Clips resource, as the name suggests, was developed in distinct projects: LE-PAROLE ("Preparatory Action for Linguistic Resources Organisation for Language Engineering") for the morphological and syntactic layers; LE-SIMPLE ("Semantic Information for Multifunctional Plurilingual Lexica") for the core of the semantic lexicon; CLIPS ("Corpora e Lessici dell'Italiano Parlato e Scritto"[16]) for the enrichment of the syntactic layers and the development of the phonological layer [19];[4]. Further works involved the analysis, enrichment and processing of PSC, among which we cite: i) the enrichment of the semantic representation with relational information between events and participants [26], ii) the conversion of the resource in a MySQL database[17], iii) the first phases of the conversion in a Linked Open Data (LOD)-compliant format [9];[16], iv) the mapping and the creation of a Sense Inventory [22]. In this long span of time, the resource has grown from a quantitative and qualitative point of view, and appears as one of the richest linguistic resources freely available for Italian. At the time we writing, PSC is available on CLARIN repository as a MySQL database, in which are represented all the four levels of linguistic information. There are no (both user and programming) interfaces. Despite the considerable investment made in the context of the European projects where its development has been carried out, this resource, at present, appears difficult to use and needs some work of renovation. The following sections present a preliminary analysis of model and data included in the database. As we will see, the entries include partially incomplete and redundant, missing, and erroneous data. These considerations justify subsequent interventions on the resource, for its better exploitation.

### *The underlying linguistic model*

The lexical model of PSC is the common architecture adopted for the development of the multilevel standardised lexicon of 12 European languages (Catalan, Danish, Dutch, Finnish, French, German, Greek, English, Portuguese, Spanish, Swedish, and Italian).

The theoretical framework of the model is the Generative Lexicon by James Pustejovsky [23], based on the Qualia ("roles"), constituents of the lexical information as "generative devices". A lexeme is expressed according to a compositional approach, although the meaning cannot be considered as a mere sum of a set of certain properties. The Qualia Structure (QS) includes four roles: formal, constitutive, telic, and agentive. The formal role provides the information

16 The speech corpus CLIPS (http://www.clips.unina.it/it/index.jsp) was developed by the CIRASS - Centro interdipartimentale di ricerca per l'Analisi e la Sintesi dei Segnali [1].

17 The dump is available on CLARIN: https://dspace-clarin-it.ilc.cnr.it/repository/xmlui/handle/20.500.11752/ILC-88

that distinguishes an individual within a larger set; the constitutive role expresses the relations concerning the composition of an entity or event; the telic role describes the functions or the scopes of an entity; finally, the agentive role formalises the origin of an entity.

The original QS was expanded in the "Extended Qualia Structure"[18] (EQS). In the EQS, subtypes are associated with the four roles of the original QS (e.g. for the telic role: object_of_the_activity, used_for, used_as, used_by), corresponding to the relations between two semantic units. Every relation is expressed by a prototypical or optional value. The Qualia relations can be combined to describe *semantic types* of different degrees of complexity. This system of semantic types is defined as the SIMPLE Ontology [17], composed of 157 concepts (as semantic types are also called in that context), the top-level of which have been modelled on the same concepts of EuroWordNet. The concepts of the ontology are divided into mono-dimensional types, organised through hypernymy relations (e.g. EARTH_ANIMAL is a sub-type of ANIMAL, that is a subtype of LIVING_ENTITY, and so on) and unified types (Multi-dimensional), represented using a combination of hypernymy relations and different dimensions of meaning (e.g. CHANGE_OF_LOCATION is a unified type, with embedded properties of his super-type CHANGE, but having also agentive information).

The ontology can be expressed with different levels of granularity: it includes a Core Ontology, the highest-level part with the most shared common concepts among the languages, and a Recommended Ontology, with specific types that represent the lower nodes in the hierarchy and which provide a more granular organisation of the word-senses [17].

An important element used to describe the semantic layer of PSC is the *template*, a predefined schema conceived to provide information to a given semantic type and, consequently, to a sense with which the template is associated. In this way, every word sense associated with a specific type is automatically equipped with a set of well-structured information, relative to different semantic features. The template can be considered as the interface between the lexicon and the ontology; the adoption of common templates is a solution to guarantee the correctness and the uniformity of the data among different lexicons.[19]

The PSC resource contains four layers of linguistic information: semantics, morphology, syntax, and phonology. These layers are separated but the single units are connected to each other through different relations: e.g. a semantic unit is connected to one or more syntactical units in order to link semantic properties (for example, the arguments) with syntactic features (for example the elements of a syntactic construction. See The available data).

### *The available data*

The structure of the relational database currently representing PSC reflects the multilayered nature of the model; we here provide a brief description of the schema and of the available data. The database is a MySQL conversion of the original MS Access database. A schematization of the main tables populating the database and their associations is shown in Figure 4.

18 http://www.ilc.cnr.it/AZ_bibliography/Z176.PDF

19 An example of a template can be found in [17].

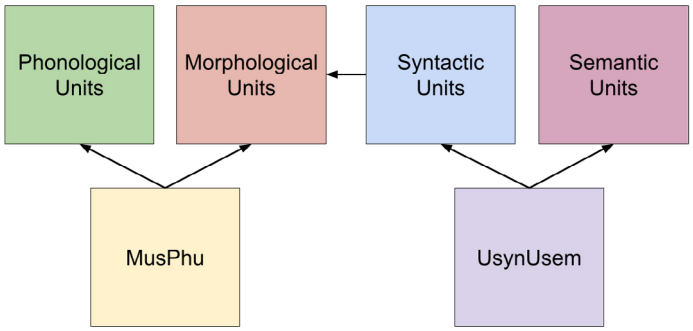


Figure 4: The main tables constituting the database that have been analysed in this work for the discovery of redundant, erroneous and missing data

We first introduce the phonological layer, composed of 387.125 phonological units (PhoUs). Each entry[20] is accompanied by its phonetic representation (described with "phono" and "sampa" attributes in the database) with hyphenation (syllables).

In particular, PhoUs represent the phonological description of words and they are connected with Morphological Units (MorUs), which constitute the lemmas of the lexicon. MorUs can be classified on the basis of their part-of-speech (pos), as it is illustrated in Table 1. The morphological layer is made up of the following information: the lemma, the pos, and a set of inflectional rules, described as "ginp" ("Combination of Morphological Features").[21] These features contain removal/addition rules for the derivation of words, and their morphological features (e.g. gender, number).

20 In this paper, we use the term "entry" to indicate a distinct unit, encoded in one of the available linguistic layers (phonological, morphological, syntactic and semantic) of the resource; an association is the representation of a relationship between two entries. More specifically, in this paper, we describe the associations between a PhoU and a MorU, between a SynU and a SemU, and between two SemUs (representing semantic relations).

21 http://www.tagmatica.fr/doc/parole_IT.pdf, 2.

| | **MorUs** | | | | | | |
|---|---|---|---|---|---|---|---|
| pos | N | A | V | ADV | NP | OTHER | Tot |
| # | 48.735 | 11.830 | 6.521 | 3.311 | 1.069 | 535 | 72.001 |
| % | 67.69% | 16.43% | 9.06 % | 4.60% | 1.48% | 0.74% | 100% |

Table 1: The distribution of the available 72001 MorUs (lemmas) wrt pos; OTHER gathers the POSs relative to the different pronouns and determinatives.

Every MorU is associated with a certain number of PhoUs, derived by the rules of the corresponding ginp, with pos and a morphological feature (morphFeat) that represents the description of inflected forms.

In Table 2, the MorU "chiesa" (church) is connected with two PhoUs: "chiesa" with morphFeat "FS" (feminine singular) and "chiese" with morphFeat "FP" (feminine plural).

| **MorU** | **PhoU** | **pos** | **morphFeat** |
|---|---|---|---|
| chiesa | chiesa | N | FS |
| chiesa | chiese | N | FP |

Table 2: The word-forms of chiesa (church).

Syntactical units (SynUs) are the entries populating the syntactic layer. SynUs are associated with a MorU as corresponding lemma and with a short description of the syntactic behaviour (e.g. the construction of transitive with an auxiliary verb). Optionally, a SynU can be described with a correlated description (e.g. the intransitive structure) and a frameset, a representation of the connection between the short description and the correlated description.[22] Table 3 illustrates, in percentage, the presence of the different features describing SynUs.

22 The descriptions of SynUs are represented in the resource according to a specific and concise annotation. For example, a value for a description is "t-xa": "t" stands for "transitive", while the "x" introduces the lexical information of the entry, in this case, the value "a" with the meaning of the auxiliary "avere" (to have). http://www.tagmatica.fr/doc/parole_IT.pdf, 6.

| | **SynUs** | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Attr. | pos | MorUs | SemUs | comment | example | description | descriptionL | framsetL |
| # | 65.565 | 65.565 | 51.070 | 16.737 | 42.796 | 65.565 | 1.481 | 1.478 |
| % | 100% | 100% | 77.89% | 25.53% | 65.27% | 100% | 0.22% | 0.22% |

Table 3: Presence of the different features describing SynUs (e.g. just 65.27% of SynUs have an example)

The semantic layer of PSC is expressed with semantic units (SemUs) and represents the richest among the linguistic layers of the resource from the point of view of encoded information. SemUs are codified with pos, lemma and definition. In addition, each SemU is associated with one template, a set of semantic traits and it's related to one or more other SemUs through semantic relations. For example, the SemU of word "piadina" (a kind of flatbread) is associated with the template "Artifact_Food". The relation IsA connects "piadina" as a hyponym of the semantic unit "pane" (bread). The description of the sense, expressed with a template, is further enriched by semantic traits; in the case of SemU "piadina", they are Concrete-Entity, Artifact (Agentive), Food (Telic), and the type of domain (Food). Regarding the argumental structure, while the SynU describes the syntactic behaviour of the word it represents, the type of arguments is associated with the relative SemU. For example, the Arg0 (ARG0offendere#1) of the verb "offendere" (to insult) is codified as "Role_ProtoAgent" (semantic role), while "ARG1offendere#1" is the Arg1 of the same verb, with the semantic role "Role_ProtoPatient".

In PSC, predicates are represented by entities associated to SemUs and to which it is possible to provide an argumentative structure, for example, a sense of word "fumatore" (smoker) is associated to the predicate "PREDfumare#1" (to smoke).

Predicates are associated with relative arguments and the appropriate semantic role; for example, "PREDabile#1" (skilled) is associated to the arguments "ARG0abile#1" and "ARG1abile#1". These arguments are connected with the semantic roles; in the case of "ARG0abile#1", the relative semantic role is "Role_ProtoAgent" while "ARG1abile#1" is associated to "Role_SOA_ARG" (State Of Affairs Argument). The correspondence between arguments of the argumental structure - the semantic frame - and the syntactic structure is defined according to values expressed with the attribute idCorresp. Table 4 shows the coverage of the different features characterising SemUs.

| | **SemUs** | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Attr. | definition | example | comment | SynU | traits | template | predicate[23] | relations |
| # | 47.937 | 24.865 | 16.513 | 57.100 | 54.437 | 54.437 | 16.317 | 48.198 |
| % | 83.85% | 43.49% | 28.88% | 99.87% | 95.22% | 95.22% | 28.54% | 84.30% |

Table 4: Presence of features describing SemUs, shown in percentage numbers

23 Predicates are encoded for verbs, nouns, and adjectives.

### *PSC compared to other resources for Italian language*

As anticipated, a comparison between PSC and the other lexical resources introduced in Related works appears difficult, especially since they are grounded on different linguistic models. However, they share some data that may be used, at least, to provide some hints about their different features and coverage.

We start from a purely quantitative point of view in the analysis of the following resources: PSC, ItalWordNet (IWN), MultiWordNet (MWN), BabelNet, and Italian Function/Content Words (IF/CW). We have decided to exclude FrameNet, Senso Comune, and BabelNet from this comparison for different reasons. Firstly, FrameNet is still under construction while Senso Comune is freely available just for a small "core" portion. BabelNet deserves a separate mention; this resource currently represents the most important knowledge graph, with highest number of senses (9.260.587[24]), synsets (5.923.356), and istances of relations available for Italian. However, as a knowledge graph and encyclopaedic dictionary, BabelNet consists of linking amongst different resources, one of them considered in our comparison (IWN), and includes named entities that do not pertain strictly to a lexicon of a language. Given its importance in the field of semantic resources and for its application in tasks as the word sense disambiguation, it was necessary to mention in the state of the art (see Related works), but it is not possible to compare BabelNet with the other considered resources.

We summarise the chosen data in the Table 5.

| | **Lemmas** | **Senses** | **Forms** | **Syntactic Units** | **Synsets** | **Relation instances** | **Number of relations** |
|---|---|---|---|---|---|---|---|
| PSC | 72.001 | 57.172 | 469.746 | 64.565 | | 83.415 | 138 |
| IWN | 48.416[25] | 68.241 | | | 49.350 | 138.385 | 83 |
| MWN | 41.491 | 57.934 | | | 32.673 | 45.593 | 14 |
| IF/CW | 90.192 | | 2.345.631[26] | | | | |

Table 5: Presence of the different features describing SynUs (e.g. just 65.27% of SynUs have an example)

While all the resources mentioned take into account a single layer of information and they do not always permit connections among the entries, PSC presents a complex inner structure,

24 Distribution: 9.219.129 nouns, 20.146 verbs, 16.662 adjectives, 4.650 adverbs, 2.651.799 concepts, and 3.271.557 named entities.

25 Distribution of pos: 3.918 proper nouns, 29.527 nouns, 8.015 verbs, 5.808 adjectives, and 1.090 adverbs.

26 The resource was probably created with an algorithmic generation of entries using morphological rules, which may have led to an hypergeneration of forms, for example for adverbs. Furthermore, there are duplicated lemmas in both resources; furthermore, the group of Function Words includes multiword expressions, thus making it further complex to estimate the number of headwords.

based on four levels of linguistic descriptions: morphology, phonology, semantics and syntax. Furthermore, it presents a varied set of relations among the entries that permits a fine-grained description of senses.

We believe that the uses of a multilevel linguistic resource have not yet been sufficiently explored in the context of NLP processes. In this sense, PSC can constitute a useful tool for further development and improvement in the context of human language technologies.

Furthermore PSC has the advantage of being conceived according to a consistent model for all levels of information and with lexicons of other languages; this feature makes it particularly suitable for its conversion into a compliant LOD format as it is already conceived to favor interoperability among different resources.

## Discovery and processing of redundant, erroneous and missing data

This section describes the process of discovery of PSC's redundant, erroneous, and missing data and which included PhoUs, MorUs, SemUs, SynUs, and the relative associations. As a result of this process, the revised data were conveyed in LexicO, which has been made available for download on CLARIN-IT repository.[27]

The first kind of data we analysed was related to the presence of redundant entries, some of which had to be removed while others represented potential redundant entries that will require a human verification. To isolate and point out candidate redundant entries some *ad hoc* algorithms and queries were developed to browse the various tables of the database; depending on the nature of the entry being examined (semantic unit, syntactic unit, morphological unit, or phonetic unit), we were able to discover those entries and mark them appropriately to be manually examined ex post on LexicO or not to be considered in case of sure redundant data.

To support this task the database representing the lexicon has been extended with a set of support tables, each of which consisting of three fields: i) the id of the candidate redundant entry, ii) the id of the redounded entry, and iii) a number representing a status, used to indicate the type of redundancy. For the sake of simplicity, from now on just two statuses will be considered for redundant entries: "sure redundant" (hereinafter referred to as redundants) and "to be verified" (that include possible redundant and erroneous or incomplete data).

A finer-grained classification of redundancy has been introduced and made available through an *ad hoc* developed web interface[28] where also the algorithms used to find out redundant, erroneous and missing data are shown and the results of which can be evaluated.[29]

Table 6 summarises the number of entries and associations (see Note 20) before and after the analysis (respectively, columns "PSC" and "LexicO"), and including the number of affected entities, summarising the extent of the intervention and containing the values relative to all inserted, not considered and modified entries of a certain kind in LexicO; finally, the column

27 http://hdl.handle.net/20.500.11752/ILC-977

28 https://klab.ilc.cnr.it/PSC-critical-entries/

29 The code of the evaluation interface is available at: https://github.com/klab-ilc-cnr/PSC-critical-entries

"to be verified" concerns all the data that cannot be modified with an automatic procedure, but require an additional check by lexicographers.

| | **PSC** | **LexicO** | **affected**[30] | **to be verified** |
|---|---|---|---|---|
| Semantic Units | 57.172 | 56.870 | 41 | 365 |
| Syntactic Units | 64.565 | 64.561 | 1.172 | 222 |
| Morphological Units | 72.001 | 71.021 | 980 | 0 |
| Phonological Units | 387.125 | 387.036 | 89 | 0 |
| Associations MorUs - PhoUs | 469.746 | 469.708 | 221 | 0 |
| Associations SynUs - SemUs | 59.089 | 59.048 | 729 | 0 |
| Semantic rels (synonymy, holonymy, meronymy) | 8.766 | 14.667 | 5.927 | 5.922 |

Table 6: The number of entries, associations and semantic relation instances in PSC and in LexicO (i.e. before and after the analysis), the affected (inserted, modified, deleted) entities and those marked to be manually verified; relations instances are, at the moment, limited to synonymy, holonymy and meronymy.

Each of the developed algorithms makes a preselection of candidate redundant entries by looking for equivalent values on a first set of features. Then, additional data are taken into account: they are added to the comparison and used by the algorithm to narrow the search by introducing more restrictions.

### *Redundant phonological units*

PhoUs populate the phonological layer of the model. The adopted procedural approach allows to discover and remove redundant PhoUs by isolating those having the same naming and phonetic representations, described by the "phono" and "sampa" attributes. The algorithm discovered a total of 89 redundant PhoUs, which, consequently, have not been exported to the new lexicon. In some cases, the removal of a PhoU involved also the update of the relative association (when present) with the MorU, as shown in the examples below.

The process of discovering candidate redundant entries distinguishes, for Phonological units, four cases. In the first case, 11 PhoUs were found as duplicate of other entries, since they had same naming, phono, morphFeat and they were linked to the same MorU; in the second case, the removal of the restriction on the equivalence in the morphFeat brought to 6 candidate redundants (to be manually reviewed). The third type of redundancy is relative to 9 "isolated" entries, i.e. PhoUs missing an association with any MorU. Finally, the algorithm found 63 PhoUs having the same naming and phono of other entries but that were referred to by different MorUs, indicating another kind of redundancy.

We first introduce an example belonging to the first case of redundancy, and relative to the two PhoUs representing the pronunciation of "liquefà" (liquefy, liquefies), which share the same phonetic information. The algorithm identified a case of redundancy and carried on the discovery process by taking into account their association with the MorU and the morphological features.

30 The number of affected entities does not include the other updated data referring to them, such as arguments, predicates, templates, traits, linked to SemUs.

As is evident from Table 7, associations 1 and 3 represent equal values in correspondence of MorU, pos, and morphological features. For this reason, PhoU "liquefa2" and its association with MorU "liquefare" were not transferred into LexicO.

| | **MorU** | **pos** | **morphFeat** | **PhoU** |
|---|---|---|---|---|
| 1 | liquefare | V | S2MP | liquefa |
| 2 | liquefare | V | S3IP | liquefa |
| 3 | liquefare | V | S2MP | liquefa2 |

Table 7: The association between the two PhoUs "liquefa" and MorU "liquefare" with the relative morphological features: "S2MP" stands for "singular second-person imperative present" while "S3IP" stands for "singular third-person indicative present".

Another type of case involved two phonological entries relative to "nocciòli" (hazelnut trees). One of them ("noccioli2") appeared "isolated" since it was not associated with any MorU; the algorithm proceeded in the same way as for entry "liquefa2". The last example is about the entry "lèggi" ([you] read), which was present as two distinct PhoUs. These two entries (identified by "leggi2" and "leggi3") appeared associated with the same MorU "leggere", but with distinct morphological features. In this case, the algorithm removed the redundant PhoU "leggi3" and modified the two forms of "leggere" in order to refer to "leggi2". In any case, if the two considered PhoUs had been associated with distinct MorUs, the algorithm would have acted in the same way, since two distinct PhoUs representing the same word pronunciation are not allowed to be present at the same time in the resource.

### *Redundant semantic units*

SemUs constituting the PSC resource are by far the richest in information among the four available types of entries. For this reason, a simple query to the database was not sufficient to identify candidate redundant SemUs. The algorithm for the discovery of redundant SemUs can be found in the "Redundant Entries" section of the web interface. The interface also shows the list of the discovered entries. Basically, the algorithm,[31] as illustrated in Figure 5, considers a SemU as a candidate redundant if it has a set of base characteristics of another SemU, namely, same naming, pos, semantic traits, template, predicates, and semantic relations.

31 The code implementing the algorithm, written in PERL, is available at https://github.com/klab-ilc-cnr/LexicO-scripts.

```
foreach (pair of entries SemU_a and SemU_b ∈ USEM)

 if (SemU_a &  SemU_b have same naming & same pos & same semantic traits &
   same templates & same predicates & same semantic relations) then
   if (SemU_a and SemU_b have same SynU of reference) then
     if (SemU_a and SemU_b have same definition & same example & same comment) then
       mark the entry with higher id with status := 15
     if (SemU_a and SemU_b have same definition & same example) then
       mark the entry with higher id with status := 14
     if (SemU_a and SemU_b have same comment & same example) then
       mark the entry with higher id with status := 13
     if (SemU_a and SemU_b have same example)
       mark the entry with higher id with status := 12
     if (SemU_a and SemU_b have same comment & same definition) then
       mark the entry with higher id with status := 11
     if (SemU_a and SemU_b have same definition) then
       mark the entry with higher id with status := 10
     if (SemU_a and SemU_b have same comment) then
       mark the entry with higher id with status := 9
     else
       mark the entry with higher id with status := 8
   else
     if (SemU_a and SemU_b have same definition & same example & same comment) then
       mark the entry with higher id with status := 7
     if (SemU_a and SemU_b have same definition & same example) then
       mark the entry with higher id with status := 6
     if (SemU_a and SemU_b have same comment & same example) then
       mark the entry with higher id with status := 5
     if (SemU_a and SemU_b have same example)
       mark the entry with higher id with status := 4
     if (SemU_a and SemU_b have same comment & same definition) then
       mark the entry with higher id with status := 3
     if (SemU_a and SemU_b have same definition) then
       mark the entry with higher id with status := 2
     if (SemU_a and SemU_b have same comment) then
       mark the entry with higher id with status := 1
     else
       mark the entry with higher id with status := 0;
```

Figure 5: The pseudocode of the algorithm for the discovery of redundant SemUs. Cases marked with status “15” are relative to sure redundant entries, which have not been included in LexicO. All the other cases have been included but will have to be manually checked by lexicographers.

A total of 41 SemUs have been discovered as (sure) redundant, si nce they shared the same base characteristics plus identical example, definition, comment and associated SynU of another SemU.

This step-by-step example documents the finding of a redundant SemU. In this case, the discovered entry is a sure duplicate of another one and, consequently, it has not been exported to LexicO. However, there are cases where all the attributes of two entries correspond except for some: each case, as aforementioned, have been marked appropriately as a candidate redundant to be manually verified.

The PSC database contained four SemUs (i.e. senses) relative to the word “abate” (abbot) and having the same naming and pos. It standed out, on the basis of the presence of repeated values in the definitions, that these SemUs could hide some redundant information.

The next step involved the comparison of the semantic relations attributed to each SemU: as depicted in Table 8, senses 2 and 4 of “abate” had identical semantic relations.

| | source | definition | relation | target |
|---|---|---|---|---|
| 1 | abate | titolo onorifico di ecclesiastici ... | isa | uomo |
| | | | memberof | chiesa |
| 2 | abate | superiore di una abbazia | isa | uomo |
| | | | memberof | abbazia |
| 3 | abate | superiore di una abbazia | NULL | NULL |
| 4 | abate | superiore di una abbazia | isa | uomo |
| | | | memberof | abbazia |

Table 8: The semantic relations attributed to each SemU of the word "abate"

In the following step the algorithm found out that the two detected SemUs also shared the same semantic traits, templates and predicates.

The last four characteristics considered by the algorithm were: i) the SynUs associated with the analysed SemUs, ii) the definition, iii) the example, and iv) the comment. If two SemUs share all these data they can be considered as identical, and one of them must be removed accordingly. This was the case of the example here described: SemUs 2 and 4 were related to the same SynU (i.e. "SYNUabateN") and shared identical example, definition, and comment: as a consequence, one of them was considered redundant and, thus, it was not exported to LexicO.

### *Redundant syntactic units*

To consider a SynU as a candidate redundant entry we required it to have, at least, the same naming, pos, description, descriptionL, and framesetL equal to those of another SynU. In addition, as we did for SemUs, on the basis of the sharing of other information the algorithm considers the entries as candidate redundants to be subsequently verified. Four SynUs sharing all data (i.e. in addition to the aforementioned base set of features, also comment, example, associated MorU and SemU) of another SynU, have been removed as redundant.

The first example of redundant SynU discovery is relative to the entry "centesimo" (hundredth) with pos "number", represented by two distinct SynUs in the resource. The two SynUs shared the same data, including the comment and SemU of reference: on the basis of the criteria followed by the algorithm the entry "centesimo2", as a sure redundant entry, was not exported from PSC to LexicO.

Another example is relative to the entry "querceto" (oak forest). In this case, the two analysed SynUs, namely "querceto" and "querceto2", differed on the associated SemUs: the algorithm thus assigned to "querceto2" a status of candidate redundant to be verified, meaning that this entry should be double-checked by a lexicographer.

### *Redundant morphological units*

The last case of automatic discovery of possible redundant entries is relative to MorUs. An in depth analysis of these elements and their relations revealed two kinds of redundancy. The first case concerns MorUs with the same naming, pos, ginp, and PhoUs of reference; the algorithm

found a total of 36 entries of this kind.

An example is relative to the entry "maltese", which appeared, as a noun, in three distinct MorUs. The algorithm focussed on the two of them which shared the same ginp, then it looked for the PhoUs they were associated to. The two MorUs referred to the same PhoUs: for this reason one of them was excluded from LexicO.

The second case is relative to a critical case of "disconnection" between SynUs and PhoUs, in which two MorUs share all data but one of them has no ginp. An example of disconnection, related to the entry "longevità" (longevity), is shown in Figure 6: the MorU with ginp was linked to a PhoU, but it was not referenced by any SynU; on the other hand, the MorU without ginp was linked to a SynU, but to no PhoUs.

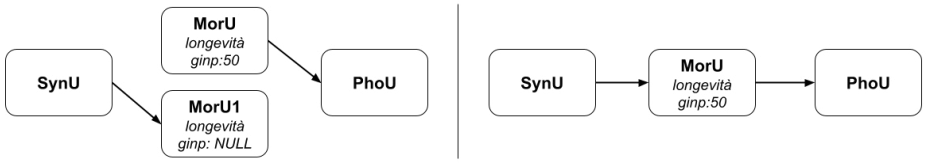


Figure 6: An example of "disconnection" between the syntactic and phonological layer caused by the presence of redundant morphological information related to lemma "longevità"; the figure represents the data before and after the update.

In this case, the algorithm proceeded with the removal of the MorU without ginp and with the update of the SynU to refer to the MorU with ginp.

### *Redundant and missing associations*

In addition to the introduced cases of redundant entries, a thorough analysis of the resource showed also the presence of i) redundant associations between MorUs and PhoUs (denoting inflected forms), ii) between SemUs and SynUs, and iii) redundant and missing associations between SemUs (semantic relations). Regarding the first and second type of associations, 24 pairs MorU/PhoU and 710 pairs SemU/SynU were recognized as redundant and, thus, not transferred into LexicO. As regards the third type of associations, the algorithm detected 10 redundant relations that, therefore, were not included in LexicO. The management of erroneous and missing relations related to synonymy, meronymy, and holonymy will be detailed in the following two subsections.

### Synonymy

The analysis of synonymy showed two distinct situations, which we faced and solved in the following order:

i) *reflexive synonyms*: 9 cases of redundant synonymic relations were found that matched the pattern "A is a synonym of A" (e.g. "USem75081buttare" synonymous of itself); these relation instances have been removed;

ii) *asymmetric synonymy*: 1645 relation instances were found as missing, since in the face of a

pattern of "A is a synonym of B" the resource lacked the inverse relation "B is a synonym of A"; these missing instances were added to LexicO.

We also found some cases of erroneous synonymic data, such as between senses of words "solito" (usual) and "insolito" (unusual). However, the fixing of these cases will need the manual intervention of a lexicographer.

### Meronymy and holonymy

Meronymy and holonymy are represented, in PSC, with four different relations; in fact, meronymy is distinguished in two values, as "Is a part of", to indicate a part of a set (e.g. "leaf" as part of a "crown of a tree") and "Is a member of", to indicate the individual in a group (e.g. "senator" as part of the "senate"). In the same way, the resource describes two kinds of holonymy: "Has as part", as the inverse relation of "Is a part of" and "Has a member", as inverse of "Is a member of". The analysis of these relations presented situations similar to synonymy. Given the pattern "A is the meronym of B", the resource does not always present the inverse relation, represented by the pattern "B is a holonym of A".

The column "PSC" of Table 9 contains the number of synonymic, meronymic and holonymic relations. Regarding meronymy, 1699 instances of relation "Is a part of" and 531 instances of "Is a member of" lacked the corresponding holonymic associations (for a total of 2230 missing associations). To bring an example, "riccio" (bur) was indicated as a meronym of "castagna" (chestnut), but this latter did not appear as a holonym of "riccio".

In the same way, 1365 instances of "Has as part" and 678 instances of "Has a member" did not present the corresponding meronymic relation (for a total of 2043 missing associations). For example, "monastero" (monastery) was present as a holonym of "monaco" (monk), but the inverse meronymic relation was not included in PSC. All these missing instances have been added to LexicO: the column "LexicO" of Table 9 indicates the final number of relations, while the rightmost column shows the number of affected instances.[32]

32 The DB table describing semantic relations also includes, for each association, a reference to the template of the source SemU and the indication of a "semantic weight" with two possible values: prototypical and essential. For the addition of missing inverse relations into LexicO the intervention of a lexicographer is required to define the semantic weight, and thus the relative field was, during the migration into the new resource, filled with a "NULL" value: the attribution of the correct values will be part of future versions of the resource. Furthermore, 43 semantic relations did not present the information of the source SemU's template: for the above mentioned reason, these relations have been moved in LexicO with a "NULL" value for the template, that will be added in future updates.

| | | **PSC** | **LexicO** | **affected** |
|---|---|---|---|---|
| synonymy | | 4062 | 5697 | 1655 |
| meronymy | Is a part of | 1803 | 3167 | 1364 |
| | Is a member of | 642 | 1318 | 679 |
| holonymy | Has a part | 1468 | 3167 | 1700 |
| | Has a member | 789 | 1318 | 529 |

Table 9: The number of synonymic and meronymic relations in PSC and in LexicO, i.e. before and after the analysis and fixing process, and the number of affected associations.

## Conclusions and perspectives

This paper describes the analysis of a portion of the Italian computational lexicon PSC and the consequent fixing of a set of data. From these revised data we started the construction of a new computational lexicon, grounded on PSC, and called LexicO. After an introduction to related works, some examples of use of PSC in a context of full-text search, a description of its underlying model and a quantification of current data, the paper has focussed on the analytical steps taken to unearth redundancies, erroneous and missing data, with the strategies adopted to fix them. While a number of duplicate and missing entries has been already detected and processed accordingly, another set of potential redundant entries has been marked for further manual revisions.

To assess the analysis carried out on PSC an *ad hoc* web interface has been developed through which it is possible to execute pre-compiled queries and analyse both the coverage and all the entries affected by the described algorithms.

On the basis of the conducted analysis it has been possible to make some considerations on current strengths and weaknesses of PSC. First of all, this resource represents the richest and most structured freely available computational lexicon for the Italian language: its multilayered nature, the huge number of relations forming a strongly interconnected network among all the different levels, and the already available linguistic data, make PSC a unique resource that is worth updating, enriching, and sharing.

Regarding issues, though we considered just a few of the available semantic relations, this kind of data appears as one of the most problematic: as shown in 9, for example, we had to double the number of meronymic relations in LexicO by adding those that were missing. On the other hand, concerning entries, just a few semantic and phonological units were found as clearly redundant, while morphological and syntactic units would need more attention.

Focussing on a more specific content and considering data with which each entry is described, the analysis highlighted some interesting facts. For example, 16.15% of SemUs do not have a definition (which may cause interpretation difficulties to users browsing the lexicon) while

15.7% of them have no semantic relation (which, in turn, may pose problems to NLP algorithms trying to take advantage from the semantic layer of the resource). We plan to continue the development of LexicO by starting from the following lines of intervention.

Firstly, all the candidate redundant entries will have to be manually examined, and, if needed, corrected, by lexicographers; moreover, we will experiment further algorithmic strategies to discover other critical data. Furthermore, the lexicon, currently stored in a relational database, will be converted into an RDF graph database, which may better represent the network-like nature of the resource and also ease and speed up its querying. At the same time, the lexicon will be made adherent to the specifications of OntoLex-Lemon model, which is natively compliant with the principles of Linguistic Linked Open Data (LLOD); we expect that in order to preserve the peculiarities and specific phenomena of the linguistic model inherited from PSC the OntoLex-Lemon model will have to be extended. This new resource will be made available on CLARIN repository as well. We also plan to link the new lexicon with other resources, starting from BabelNet. Finally, to share the resource in the community of lexicographers, and enable them to collaboratively work in the correction and enrichment of the lexicon, we intend to arrange and start up a dedicated project involving the most authoritative institutions in the field of Italian lexicography.

The availability of a renewed and updated Italian computational lexicon, especially in an easily shareable LLOD compliant format, could bring considerable benefits in terms of impact in all digital contexts involving the linguistic and semantic based access to texts and in the support of NLP tasks, such as word sense disambiguation. More in general, despite its current redundancies and gaps, especially if further fixed and enriched, LexicO may represent a computational lexicon of reference for all communities working with resources for the Italian language.

## Data Availability Statement

All data involved in this work are openly available at the following locations: i) the PSC resource in ILC4CLARIN repository at http://hdl.handle.net/20.500.11752/ILC-88, ii) the LexicO resource in ILC4CLARIN repository at http://hdl.handle.net/20.500.11752/ILC-977, iii) the web interface for the evaluation of PSC's critical entries on KLAB group website at https://klab.ilc.cnr.it/PSC-critical-entries, iv) the code of the interface in KLAB group GitHub repository at https://github.com/klab-ilc-cnr/PSC-critical-entries and, v) the scripts and the queries implementing the algorithms in KLAB group GitHub repository at https://github.com/klab-ilc-cnr/LexicO-scripts.

## Funding

This work was supported by the TALMUD project and carried out within the scientific collaboration between S.c.a r.l. PTTB and CNR-ILC (09/11/2017).

## References

[1]. Leoni, Federico Albano. 2007. "Un frammento di storia recente della ricerca (linguistica) italiana. Il corpus CLIPS." *Bollettino d'Italianistica*, n.s., IV, 2: 122-130. https://doi.org/10.7367/71826.

[2]. Basili, Roberto, Brambilla, Silvia, Croce, Danilo, and Tamburini, Fabio. 2017. "Developing a Large Scale FrameNet for Italian: the IFrameNet Experience." In *Proceedings of the Fourth Italian Conference on Computational Linguistics (CLiC-it 2017)*, AAccademia University Press. https://doi.org/10.4000/books.aaccademia.2364.

[3]. Bevilacqua, Michele, and Navigli, Roberto. 2020. "Breaking Through the 80% Glass Ceiling: Raising the State of the Art in Word Sense Disambiguation by Incorporating Knowledge Graph Information." In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2854–2864. https://doi.org/10.18653/v1/2020.acl-main.255.

[4]. Calzolari, Nicoletta. 2003. "Risorse Linguistiche per la lingua italiana scritta." in *Conferenza TIPI-Tecnologie Informatiche nella Promozione della Lingua Italiana*, Roma.

[5]. Chiari, Isabella. 2012a. "Il dato empirico in lessicografia: dizionari tradizionali e collaborativi a confronto." *Bollettino di italianistica*, 2: 94-125. https://doi.org/10.7367/72622.

[6]. Chiari, Isabella. 2012b. "Linguistic resources and machine translation trends for the Italian language: overview and perspectives" In *Language Translation Automation Conference (LTAC)*, edited by Cannavina, Valeria and Fellet, Anna, The Big Wave, 105-123.

[7]. Fellbaum, Christiane. 1998. *WordNet: An Electronic Lexical Database*. Cambridge (Mass.):MIT Press. https://doi.org/10.7551/mitpress/7287.001.0001.

[8]. Fersini, Elisabetta, Anzovino, and Rosso. 2018. "Overview of the task on automatic misogyny identification at ibereval. In Proceedings of the Third Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018)". CEUR Workshop Proceedings. Seville: CEUR-WS. org.

[9]. Del Gratta, Riccardo, Frontini, Francesca, Khan, Fahad, and Monachini, Monica. 2015. "Converting the PAROLE SIMPLE CLIPS lexicon into RDF with lemon." *Semantic Web*, 6, 4: 387–392. https://doi.org/10.3233/SW-140168.

[10]. Giovannetti, Emiliano, Albanesi, Davide, Bellandi, Andrea, and Benotto, Giulia. 2017. "Traduco: A collaborative web-based CAT environment for the interpretation and translation of texts." *Digital Scholarship in the Humanities*, 32, suppl_1: 47-62. https://doi.org/10.1093/llc/fqw054.

[11]. Giovannetti, Emiliano, Albanesi, Davide, Bellandi, Andrea, Marchi, Simone, Papini, Mafalda, and Sciolette, Flavia. 2022. "The role of a computational lexicon for query expansion in full-text search." In *Proceedings of CLiC-it 2021: Italian Conference on Computational Linguistics*, CEUR workshop proceedings. https://doi.org/10.4000/books.aaccademia.10638.

[12]. Grella, Matteo. 2018. Italian Content Words v3, LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University. http://hdl.handle.net/11372/LRT-2894.

[13]. Grella, Matteo, 2018, Italian Function Words v3, LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathe-

matics and Physics, Charles University, http://hdl.handle.net/11372/LRT-2893.

[14]. Jiang, Yuncheng. 2020. "Semantically-enhanced information retrieval using multiple knowledge sources." *Cluster Computing* 23, 4: 2925–2944. https://doi.org/10.1007/s10586-020-03057-7.

[15]. Jindal, Rajni and Shweta, Taneja. 2018. "A Novel Method for Efficient Multi-Label Text Categorization of research articles". In *International Conference on Computing, Power and Communication Technologies (GUCON)*, 333-336. https://doi.org/10.1109/GUCON.2018.8674985.

[16]. Khan, Fahad, Bellandi, Andrea, Frontini, Francesca, and Monica Monachini. 2018. "One Language to rule them all: Modelling Morphological Patterns in a Large Scale Italian Lexicon with SWRL." In *Proceedings of the 11th International Conference on Language Resources and Evaluation - LREC2018*, Miyazaki, Japan. https://aclanthology.org/L18-1694.

[17]. Lenci, Alessandro, Bel, Nuria, Busa, Federica, Calzolari, Nicoletta, Gola, Elisabetta, Monachini, Monica, Ogonowski, Antoine, Peters, Ivonne, Peters, Wim, Ruimy, Nilda, Villegas, Marta, and Zampolli, Antonio. 2000. "SIMPLE: A General Framework for the Development of Multilingual Lexicons". *International Journal of Lexicography* 13, 4: 249–263. https://doi.org/10.1093/ijl/13.4.249.

[18]. McCarthy, Arya D., Kirov, Christo, Grella, Matteo, Nidhi, Amrit, Xia, Patrick, Gorman, Kyle Vylomova, Ekaterina, Mielke, Sabrina J., Nicolai, Garrett, Silfverberg, Miikka, Arkhangelskiy, Timofey, Krizhanovsky, Nataly, Krizhanovsky, Andrew, Klyachko, Elena, Sorokin, Alexey, Mansfield, John, Ernštreits, Valts, Pinter, Yuval, L. Jacobs, Cassandra, et al. 2020. "UniMorph 3.0: Universal Morphology." In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, Marseille, 3922–3931. https://doi.org/10.3929/ethz-b-000462327.

[19]. Monachini, Monica, Calzolari, Federico, Mammini Michele, Rossi Sergio, and Ulivieri, Marisa. 2004. "Unifying Lexicons in view of a Phonological and Morphological Lexical DB" In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC04)*, Lisbona, European Language Resources Association (ELRA), 1107-1110.

[20]. Navigli, Roberto, and Pozzetto, Simone Paolo. 2012. "BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network." *Artificial Intelligence*, 193: 217-250. https://doi.org/10.1016/j.artint.2012.07.001.

[21]. Pianta, Emanuele, Bentivogli, Luisa, and Girardi, Christian. 2002. "MultiWordNet: developing an aligned multilingual database." In *Proceedings of the First International Conference on Global WordNet*, 293-302.

[22]. Poli, Francesca. 2021. *Italian Sense Inventory.* ILC-CNR for CLARIN-IT repository hosted at Institute for Computational Linguistics "A. Zampolli", National Research Council. http://hdl.handle.net/20.500.11752/OPEN-5

[23]. Pustejovsky, James. 1995. *The Generative Lexicon*. Cambridge, MA: MIT Press. https://doi.org/10.7551/mitpress/3225.001.0001.

[24]. Roventini, Adriana, Alonge, Antonietta, Bertagna, Francesca, Calzolari, Nicoletta, Cancila, Jean, Girardi, Magnini, Bernardo, Marinelli, Rita, Speranza, Manuela, and Zampolli, Antonio. 2003. "ItalWordNet: building a large semantic database for the automatic

treatment of Italian." In *Computational Linguistics in Pisa*. Edited by Zampolli Antonio, Calzolari, Nicoletta L. Cignoni. *Linguistica Computazionale, Special Issue,* 745-791.

[25]. Ruimy, Nilda, Monachini, Monica, Distante, Raffaella, Guazzini, Elisabetta, Molino, Stefano, Ulivieri Marisa, Calzolari, Nicoletta, and Zampolli, Antonio. 2002. "Clips, a multi-level italian computational lexicon: A glimpse to data." In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC02)*.

[26]. Ruimy, Nilda. 2010. "Simple_PLUS: a network of lexical semantic relations." *Procesamiento del Lenguaje Natural*, 44: 99-106. https://www.redalyc.org/pdf/5157/515751744018.pdf.

[27]. Sabatini, Francesco. 2011. "La storia dell'italiano nella prospettiva della corpus linguistics." in *I dialetti, il latino, modelli teorici, la Crusca, l'Europa. Saggi dal 1968 al 2009*, edited by; Coletti, V., Coluccia, R., D'Achille, P., De Blasi, N., & Proietti,, Napoli: Liguori editore, II: 223-232.

[28]. Vieu, Laure, Jezek, Elisabetta and van de Cruys, Tim. 2015. "Quantitative methods for identifying systematic polysemy classes." In *6th Conference on Quantitative Investigations in Theoretical Linguistics* (QITL 2015), Tübingen, 1-5.