Towards a New Computational Lexicon for Italian: building the morphological layer by harmonizing and merging existing resources

Flavia Sciolette[1,*,†], Simone Marchi[1,†] and Emiliano Giovannetti[1,†]

[1]Istituto di Linguistica Computazionale "Antonio Zampolli" (CNR-ILC), Area della Ricerca del CNR di Pisa, Via G. Moruzzi, 1, 56124 Pisa, Italy

Abstract

The present work illustrates the first steps towards the construction of a new computational lexicon for the Italian language. Following an analysis of existing lexical resources, it was decided to use LexicO as the reference base. In this first phase a resource of nearly 800,000 inflected forms was produced, accompanied by lemmas and morphological traits, obtained by integrating the available data in LexicO with those coming from two support sources: the tool MAGIC and a selection of Italian treebanks.

Keywords

computational lexicon, lexical resources, morphology, morphological harmonization

1. Introduction

A significant number of digital lexical resources are available for many languages. In CLARIN Virtual Language Observatory (VLO)[1], a search for "lexicalResource" of Italian provides 52 results. Two resources appear in several versions and updates: Parole-Simple-Clips (PSC)[2] [1], a multilayered lexicon, and ItalWordNet[3]. The most part of the results includes monolingual and multilingual domain terminologies. Amongst the notable resources are worth mentioning Italian Function Words (IFWs)[4] and Italian Content Words (ICWs)[5], two lists in JSON Lines format developed for supporting POS tagging and syntactic parsing of Italian. In fact, a number of NLP tasks can take advantage of lexical resources, for example sentiment analysis [2] but also "semantic role labeling, verb sense disambiguation and ontology mapping" [3].

However, ICWs includes hundreds of thousands of forms generated automatically and not manually revised, which, despite being morphologically correct, have very low, if not zero, usage frequency.

Although not listed in Clarin's VLO, we also mention SIMPLELex-it, since built similarly to our lexicon by combining various existing resources [4].

Even in lexical resources which have been manually developed and revised, however, the linguistic coverage of entries can pose problems, both in terms of lexical coverage and content of entries. Hence, integrating information from different sources, as we did in this work, can be effective in filling the gaps, though it can present several challenges in terms of harmonization of distinct formats and models.

We here describe the first steps towards the construction of a new computational lexicon for the Italian language that we called CompL-it. We started from the enrichment of an existing resource, LexicO[6] [5], a computational lexicon which, in turn, was derived from the already cited PSC lexicon.

In particular, this first phase was focused on the expansion of the morphological layer, carried out through the integration of two other resources: a list of lemmatized forms generated by the morphological analyzer MAGIC[7] [6, 7], and a set of Italian treebanks. The obtained resource, constituted of nearly 800 thousand forms, was made available in a CoNLL-like format as a tabular separated values (or TSV)[8].

This core of forms, lemmas, and morphological traits will populate the morphological layer of the computational lexicon CompL-it under construction, which will later be released in the form of Linguistic Linked Open

CLiC-it 2023: 9th Italian Conference on Computational Linguistics, Nov 30 — Dec 02, 2023, Venice, Italy

*Corresponding author.

†These authors contributed equally.

flavia.sciolette@ilc.cnr.it (F. Sciolette); simone.marchi@ilc.cnr.it (S. Marchi); emiliano.giovannetti@ilc.cnr.it (E. Giovannetti)

https://klab.ilc.cnr.it/ (F. Sciolette); https://klab.ilc.cnr.it/ (S. Marchi); https://klab.ilc.cnr.it/ (E. Giovannetti)

0000-0002-7998-9768 (F. Sciolette); 0000-0003-4320-6466 (S. Marchi); 0000-0002-0716-1160 (E. Giovannetti)

[1]https://vlo.clarin.eu/ [25/07/2023]

[2]https://dspace-clarin-it.ilc.cnr.it/repository/xmlui/handle/20.500.11752/ILC-88

[3]https://dspace-clarin-it.ilc.cnr.it/repository/xmlui/handle/20.500.11752/ILC-62

[4]https://lindat.mff.cuni.cz/repository/xmlui/handle/11372/LRT-2893

[5]https://lindat.mff.cuni.cz/repository/xmlui/handle/11372/LRT-2894

[6]https://dspace-clarin-it.ilc.cnr.it/repository/xmlui/handle/20.500.11752/ILC-977

[7]https://dspace-clarin-it.ilc.cnr.it/repository/xmlui/handle/20.500.11752/ILC-1002

[8]https://github.com/klab-ilc-cnr/CompL-it_morphological_layer

Data (LLOD) (see Section 5). For this reason, it was chosen not to update the relational database of LexicO, but to use the CoNLL-like format as a temporary data representation format.

2. The sources

The sources we considered for building the morphological layer of CompL-it differ from each other for model, vocabularies, and aims; for this reason, it was first necessary to carry out a harmonization process to make the resources comparable to each other, as shown in Section 3.

Regarding the choice of sources, we opted to include only resources for which manual revision was documented. In this sense, we chose not to delve into the data at this initial stage. Corrective actions, aimed at preventing the generation and propagation of errors, focused on issues that could be resolved through automatic processes and were independent of data evaluation, such as redundancy or the comparison of entries to assess information richness (see 3.3).

2.1. LexicO

LexicO is available on CLARIN as a relational database, and shares the same linguistic model of PSC, which is based on the theory of Generative Lexicon by James Pustejovsky [8]. LexicO contains four layers of linguistic information: morphology, syntax, semantics, and phonology.

2.2. MAGIC

MAGIC is a morphological analyzer which includes three modules: a lexicon compiler for Italian, the morphological analyzer itself, and the morphological generator.

With an *ad hoc* script, we extracted all forms generated by the morphological analyzer. The generated output consists of a series of linguistic objects called "words", for each of which lemmas, morphosyntactic types, and features are specified. This resource was made available on CLARIN as "MAGIC - Generated Lemmatized Forms" (M-GLF).

2.3. Universal Dependencies treebanks

Treebanks are collected and listed in the Universal Dependencies (UD) repository[9]. We excluded non-manually revised treebanks from the selection. Additionally, we excluded treebanks aimed at representing specific case studies that could introduce sparsely attested forms into the lexicon or introduce excessive "noise". We considered the following treebanks: i) ISDT [9]; ii) VIT - Venice Italian Treebank [10]; iii) TUT [11]; iv) ParlaMint-It, based on ParlaMint-It corpus[10] [12].

3. The building of the morphological layer

3.1. Harmonization

Morphological data are represented in the considered resources in different ways. The vocabulary labels of each resource was mapped into LexInfo[11], the data category ontology for OntoLex-Lemon model[12], *de facto* standard for representing lexical resources in the Semantic Web. In the case of M-GLF and LexicO, it involved the direct conversion of their custom tagsets - specific for Italian - into the nomenclature of LexInfo.

The LexicO and M-GLF vocabularies also follow a different theoretical approach compared to the UD used in treebanks. In the first two cases, the vocabulary is designed for lexical resources, and the POS tags are fine-grained, often finding a direct counterpart in LexInfo, as LexInfo serves as an ontology for this type of resource. In the case of UD, used for corpus annotation, word descriptions are assigned a "universal" POS tag, further specified by features defined in the Universal Features vocabulary.

The cases addressed in the mapping can be classified into three types: i) perfect correspondence; in these cases, the value was directly converted into the LexInfo vocabulary; ii) correspondence of POS in combination with another value; in these cases, the mapping associated a LexInfo label with a combination of POS and a morphological feature, as seen in the case of demonstratives in UD; iii) correspondence not present in LexInfo; in this case, a new class was formalized and linked to OLIA[13]. The tables for mapping has been made available on GitHub[14].

3.2. Conversion to CoNLL-like format

Once the vocabularies were harmonized, each resource was converted into a file in CoNLL-like format.

The choice of this format is primarily due to two reasons: i) UD treebanks are already in tabular format (CoNLL is a TSV); ii) the information from LexicO and M-GLF does not have a specific output format (the former is stored in a relational database, while the latter is in a textual format that does not adhere to any standard) and can be easily transformed into a TSV format.

[9] https://github.com/UniversalDependencies

[10] https://www.clarin.eu/parlamint

[11] https://github.com/ontolex/lexinfo

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

[13] https://github.com/acoli-repo/olia

[14] https://github.com/klab-ilc-cnr/Tables-for-mapping-of-Italian-Lexicon-ComplIt

In a first phase, from each of the aforementioned resources, a list of forms with lemmatization and morphological traits was extracted. Subsequently, the obtained lists were converted in distinct CoNLL-like files, using *ad hoc* developed Perl scripts. In this phase, the tagsets were also converted according to the mappings in LexInfo mentioned in the previous section.

3.3. Merging

The merging process of the three resources represented by the CoNLL-like files was divided into two phases: i) initially, two resources were compared and combined in a partial merge; ii) subsequently, the third resource was added to the comparison to obtain the final output. The algorithm compared two entries at a time. If two entries were equal in terms of form, POS, and lemma, then their morphological features were compared. If the features of the first entry constituted a subset of those belonging to the second entry, the latter was considered for the final output, being richer in linguistic information. The algorithm, developed in Java, was made available on GitHub[15].

4. Evaluation

The resulting output consists of 790,758 forms associated with 102,000 lemmas and the relative traits.

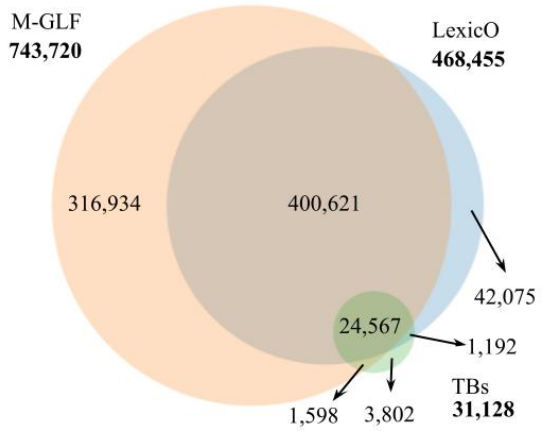

Figure 1: Venn diagram representing the size of the three resources (in terms of forms) and their intersections

Figure 1 shows, under the labels with the name of the resource, the total number of forms contained in that specific resource; the diagram illustrates the sizes of the intersections and of the areas that represent the forms which are specific of a resource.

[15]https://github.com/klab-ilc-cnr/compareAndMergeLexicons

Similarly, Figure 2 shows the distribution of lemmas per resource.

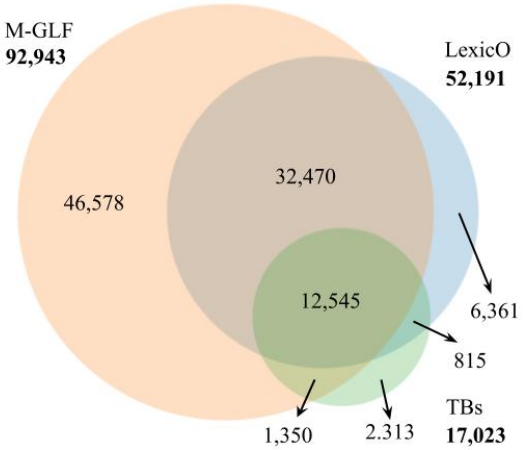

Figure 2: Venn diagram representing the size of the three resources (in terms of lemmas) and their intersections

In CompL-it, each form is associated with the following data: lemma, part of speech (POS), and morphological features specific to the considered POS. Table 1 shows an example of form for the lemma "gatto" (cat).

Table 1
Example of data associated to a form

form	lemma	pos	feats
gatte	gatto	noun	feminine | plural

Despite the evident larger size of M-GLF compared to LexicO, it is important to specify that the choice to use this latter as the reference base was mainly qualitative, in particular for its multilevel structure[16], which will be exploited for enriching CompL-it in subsequent works (Section 5).

In Table 2, the number of forms per POS in LexicO is compared to the final CompL-it resource, along with the respective percentage increase.

It is worth noting the significant increase in values, particularly for adjectives and adverbs, which have a lower coverage in LexicO[17].

To conclude this section, we provide in Table 3 a quantitative comparison between CompL-it and some of the lexical resources mentioned in the introduction, specifically PSC, ItalWordNet, and SIMPLELex-IT. We excluded

[16]For further details on the structure of entries in LexicO, please refer to [5]

[17]These POSs were already poorly covered in PSC, from which LexicO has been derived: in the final phase of the last project on the development PSC, the coding of adjectives and adverbs was still under construction [13].

Table 2
Percentage increase in the numbers of inflected forms, by POS, compared to those already available in LexicO.

POS	LexicO	CompL-it	increase
verb	345,109	545,104	+58%
noun	75,933	136,163	+79%
adj.	45,716	103,881	+127%
adv.	746	3,222	+332%
other	951	2,419	+151%
total	468,455	790,758	+69%

Table 3
Comparison of CompL-it and three other lexical resources in terms of numbers of lemmas and forms.

source	lemmas	forms
PSC	72,001	469,746
ItalWordNet	48,416	-
SIMPLELex-IT	7,022	26,500
CompL-it	102,000	790,758

ICWs from the comparison due to the mentioned issue of overgenerating forms (which doesn't align well with the need to represent lexically precise data) and IFWs, as it contains many multiword entries that we have chosen to exclude from our lexicon at this time.

5. Conclusions and future works

In this article, we documented a first step towards building a new computational lexicon of Italian. A set of approximately 800 thousand lemmatized forms with morphological features was created, through the integration of existing resources. In the next phases, this lexical core will be converted as a LLOD based on the OntoLex-lemon model, making the resulting lexicon more easily shareable, interoperable, and compliant with Semantic Web standards. Additionally, new linguistic layers will be added, starting from semantics, by using the information already available in LexicO and by integrating data from WordNets for Italian.

Acknowledgments

This work was conducted in the context of the TALMUD project and the scientific cooperation between S.c.a r.l. PTTB and CNR-ILC.

References

[1] A. Lenci, N. Bel, F. Busa, N. Calzolari, E. Gola, M. Monachini, A. Ogonowski, I. Peters, W. Peters, N. Ruimy, M. Villegas, A. Zampolli, SIMPLE: A General Framework for the Development of Multilingual Lexicons, International Journal of Lexicography 13 (2000) 249–263. doi:10.1093/ijl/13.4.249.

[2] T. N. Prakash, A. Aloysius, Textual Sentiment Analysis Using Lexicon Based Approaches, Annals of the Romanian Society for Cell Biology (2021) 9878–85. URL: http://annalsofrscb.ro/index.php/journal/article/view/3734.

[3] S. Brown, J. Windisch, G. Kazeminejad, A. Zaenen, J. Pustejovsky, M. Palmer, Semantic Representations for NLP Using VerbNet and the Generative Lexicon, Frontiers in Artificial Intelligence 5 (2022). doi:10.3389/frai.2022.821697.

[4] A. Mazzei, Building a computational lexicon by using SQL, in: Proceedings of the Third Italian Conference on Computational Linguistics CLiC-it 2016: 5-6 December 2016, Napoli, 2016. doi:10.4000/books.aaccademia.1808.

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

[6] M. Battista, V. Pirrelli, Una Piattaforma di Morfologia Computazionale per l'Analisi e la Generazione delle Parole Italiane, Technical Report, ILC-CNR Technical Report, 1999.

[7] V. Pirrelli, M. Battista, The Paradigmatic Dimension of Stem Allomorphy in Italian Verb Inflection, Rivista di Linguistica 12 (2000) 307–379.

[8] J. Pustejovsky, The Generative Lexicon, MIT Press, Cambridge, MA, 1995.

[9] M. Simi, C. Bosco, S. Montemagni, 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), 2014, pp. 83–90.

[10] R. Delmonte, A. Bristot, S. Tonelli, VIT - Venice Italian Treebank: Syntactic and Quantitative Features, in: Proceedings of the Sixth International Workshop on Treebanks and Linguistic Theories, volume 1, 2007, pp. 43–54.

[11] M. Sanguinetti, C. Bosco, PartTUT: The Turin University Parallel Treebank, in: Harmonization and development of resources and tools for Italian Natural Language Processing within the PARLI project, LNCS, Springer Verlag, 2014.

[12] T. Agnoloni, R. Bartolini, F. Frontini, S. Montemagni, C. Marchetti, V. Quochi, M. Ruisi, G. Venturi,

Making Italian Parliamentary Records Machine-Actionable: the Construction of the ParlaMint-IT Corpus, in: Proceedings of the Workshop ParlaCLARIN III within the 13th Language Resources and Evaluation Conference, 2022, pp. 117–124. URL: https://aclanthology.org/2022.parlaclarin-1.17/.

[13] N. Ruimy, M. Monachini, R. Distante, E. Guazzini, S. Molino, M. Ulivieri, N. Calzolari, A. Zampolli, Clips, a multi-level Italian computational lexicon: A glimpse to data, in: Proceedings of the Third International Conference on Language Resources and Evaluation (LREC02), 2002.