The CIRCSE Collection of Linguistic Resources in CLARIN-IT

Marco Passarotti
CIRCSE Research Center
Università Cattolica del Sacro Cuore
Milan, Italy
marco.passarotti@unicatt.it

Rachele Sprugnoli
CIRCSE Research Center
Università Cattolica del Sacro Cuore
Milan, Italy
rachele.sprugnoli@unicatt.it

Abstract

In this paper, we present the collection of the linguistic resources for Latin made available by the CIRCSE Research Center in the CLARIN-IT repository. After an introduction about the history and the main research lines of the Center, the paper provides details both the lexical and the textual resources that were built across more than a decade at the CIRCSE and that are now accessible in CLARIN-IT.

1 The CIRCSE Research Center

The CIRCSE Research Center of the Università Cattolica del Sacro Cuore (Milan, Italy)[1] was founded in 2009 by Marco Passarotti and Savina Raynaud, to keep the legacy of a former Research Group (GIRCSE), which was started by the pioneer of linguistic computing father Roberto Busa at the end of the '70s, in strict connection with his course of Computational Linguistics at Università Cattolica (Bolognesi, 1999).

Following in the footsteps of father Busa, whose main contribution was the Index Thomisticus (IT) corpus collecting the opera omnia of Thomas Aquinas (Busa, 1974–1980), the research topics addressed at CIRCSE focus on building and disseminating linguistic resources and Natural Language Processing (NLP) tools for ancient languages, especially for Latin. Since its beginning, the core project of the Research Center was the Index Thomisticus Treebank, which aims to enhance the texts of the IT with syntactic annotation (Passarotti, 2019). From 2015 to 2017, the CIRCSE hosted a Marie Sklodowska-Curie IF grant[2] focused on Latin derivational morphology. Since 2018, the CIRCSE hosts the *LiLa: Linking Latin* ERC-Consolidator Grant.[3] The objective of LiLa is to build a Knowledge Base of interlinked linguistic resources for Latin according to the principles of the Linked Data paradigm (Passarotti et al., 2019), thus combining computational linguistics, semantic web and classical studies in an interdisciplinary perspective.

Reflecting its main research line, the CIRCSE contributes to CLARIN-IT by sharing the lexical and textual resources for Latin developed across more than a decade at the Center. Given the interdisciplinary nature of its resources, which provide (meta)data in an ancient language built and distributed according to state-of-the-art methods and formats in computational linguistics, the CIRCSE plays a strategic role in the CLARIN-IT context, considering that the Infrastructure aims to impact the entire research community that benefits from easily accessing linguistic data. Such community is large and diverse, including not only NLP scholars and computational linguists, but also digital and traditional humanists, who are interested in finding and accessing the CIRCSE resources in a wide, common repository such as CLARIN-IT.

[1] Centro Interdisciplinare di Ricerche per la Computerizzazione dei Segni dell'Espressione. https://centridiricerca.unicatt.it/circse_index.html.

[2] Agreement No 658332-WFL Word Formation Latin.

[3] https://lila-erc.eu/. This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program – Grant Agreement No. 769994.

2 Linguistic Resources

This section presents details the linguistic resources of the CIRCSE Research Center that were made available through the ILC4CLARIN data center[4] in a dedicated collection[5] and the CLARIN Virtual Language Observatory. [6] Table 1 summarizes the type, size and format of the resources.

- LiLa Lemma Bank (Passarotti et al., 2019): a large collection of Latin lemmas, serving as the backbone to achieve interoperability between the resources in the LiLa Knowledge Base, by linking all those entries in lexical resources and tokens in corpora that point to the same lemma. Each lemma is described using a set of grammatical and morphological information, such as the Part-of-Speech and the inflection type; different written representations of the same lemma can be reported (e.g. *metrum*, *metron*, *metrom* for the lemma meaning "a measure").
- Index Thomisticus Treebank (IT-TB) (Passarotti, 2019): syntactically annotated portion of the IT corpus. The IT-TB includes the analytical (i.e. surface syntactic) dependency annotation of the entire "Summa contra Gentiles" (4 books), as well as of the concordances of lemma *forma* ("form") from "Scriptum super libros sententiarum magistri Petri Lombardi" (entire) and from "Summa Theologiae" (partial). The annotation guidelines are inspired by those of the analytical layer of the Prague Dependency Treebank.[7] The resource features also more than 2,000 dependency trees for as many sentences from "Summa contra Gentiles" annotated at the tectogrammatical (i.e. underlying syntactic) layer following the corresponding annotation guidelines of the Prague Dependency Treebank.[8]
- Latin Vallex v.1 (Passarotti et al., 2016): valency lexicon for Latin. The first version was built in close connection with the semantic/pragmatic annotation of the Index Thomisticus Treebank and the Latin Dependency Treebank. Data are stored in a single XML file, whose structure is the same of that for the valency lexicon for Czech PDT-VALLEX.[9]
- LatinAffectus (Sprugnoli et al., 2020a; Sprugnoli et al., 2020c): prior polarity lexicon of Latin lemmas developed by merging a Gold Standard and a Silver Standard. The Gold Standard was created by two experts of Latin language and culture following a multi-stage process and an extensive reconciliation phase. It features a five-way classification: 1 (fully positive), 0.5 (somewhat positive), 0 (neutral), -0.5 (somewhat negative), -1 (fully negative). The Silver Standard was built by deriving new entries starting from those in the Gold Standard through synonym, antonym and derivational relations.
- Index Graecorum Vocabulorum in Linguam Latinam (IGVLL) (Franzini et al., 2020): manually-corrected OCR of Günther Alexander Saalfeld's list of Latin loans from Ancient Greek (1874). It contains the Latin loanword (occasionally accompanied by variants), the Ancient Greek source lemma(s) (many lemmas include graphical, morphological and dialectal variants), and the link between the Ancient Greek lemma and its corresponding canonical forms in a machine-readable version of the Greek-English Liddell-Scott Jones lexicon (Blackwell, 2018).
- Word Formation Latin (WFL) (Litta et al., 2016): derivational morphology resource where lemmas are analyzed into their formative components, and relationships between them are established on the basis of Word Formation Rules (WFRs).
- EvaLatin 2020 Data (Sprugnoli et al., 2020b): training and gold test data released in EvaLatin 2020. The two shared tasks proposed in the campaign, i. e. Lemmatization and Part-of-Speech tagging,

[4] https://ilc4clarin.ilc.cnr.it/
[5] https://dspace-clarin-it.ilc.cnr.it/repository/xmlui/handle/000-c0-111/525
[6] https://vlo.clarin.eu/search?0&fq=collection:CIRCSE
[7] http://static.perseus.tufts.edu/docs/guidelines.pdf
[8] https://ufal.mff.cuni.cz/pdt2.0/doc/manuals/en/t-layer/pdf/t-man-en.pdf
[9] http://ufal.mff.cuni.cz/pdt2.0/doc/data-formats/pdt-vallex/pdt-vallex-struct.html

NAME	TYPE	SIZE	FORMAT
LiLa Lemma Bank	lexicon	196,853 lemmas	RDF-Turtle
IT-TB	annotated corpus	450,000 nodes 25,000 sentences	XML
Latin Vallex v1	lexicon	1,426 lemmas 3,650 frames	XML
LatinAffectus	lexicon	2,437 lemmas	CSV
IGVLL	lexicon	1,763 lemmas	CSV
WFL	lexicon	34,951 relations 773 rules	SQL
EvaLatin 2020	annotated corpus	341,419 tokens 16 files	CoNLL-U
EDLIL	lexicon	1,874 lemmas	RDF-Turtle

Table 1: Information about the CIRCSE resources in CLARIN-IT.

were aimed at fostering research in the field of language technologies for Classical languages. The shared dataset consists of texts taken from the Perseus Digital Library, processed with UDPipe models (Straka and Straková, 2017) and then manually corrected by Latin experts. The training set includes only prose texts by Classical authors. The test set, alongside prose by the same authors represented in the training set, also includes poems and texts of the Medieval period.

- The Etymological Dictionary of Latin and the other Italic Languages (EDLIL) (Mambrini and Passarotti, 2020): collection of Proto-Italic and Proto-Indo-European reconstructed forms taken from the most recent "Etymological Dictionary of Latin and the other Italic Languages" (de Vaan, 2008), modeled following the Linked Data paradigm and released in RDF-Turtle format.

3 Examples of Use

By combining the resources described above, it is possible for the users of the CLARIN infrastructure to gather various types of linguistic information about Latin lemmas and their context of use in corpora. For example, the LiLa Lemma Bank reports that the lemma *dignus* ("worthy") is a first class adjective with a positive degree, having a deadjectival adverb (*digne* "worthily") and belonging to the same word formation family of other 25 lemmas such as the verb *digno* ("to deem worthy"). WFL describes how other lemmas derive from *dignus* through derivation or compounding: for example, *indignitas* ("unworthiness") derives from *dignus* by adding the prefix *in*(negation)- and the suffix *-tas/tat*. Semantic roles, called functors following the Functional Generative Description framework (Sgall et al., 1986) and expressing the types of relations between a word and its complements, are reported in Latin Vallex v.1: more specifically, *dignus* has one frame entry and one complement having the role of Patient that can be linguistically realized in four different ways. For example, in *[...] dono dignum esset* ("worthy [of] a gift"; Sallust, *Bellum Catilinae*, 54) the complement is realized with an ablative noun (*dono*). *Dignus* is also an entry in LatinAffectus with a fully positive polarity (+1) and it is a lemma appearing 177 times in the EvaLatin 2020 dataset and 36 times in the IT-TB. One of the occurrences of *dignus* in the IT-TB is annotated at both the analytical and the tectogrammatical layer. The occurrence in question is the following: *ipsa substantia angeli [...] est* ***dignior*** *rebus sensibilibus [...]* "the substance of an angel [...] is nobler than sensible things [...]" ("Summa contra Gentiles", book 1, chapter 3, number 5).[10] While the analytical dependency subtree of this portion of the sentence shows that the node for *dignior* is the nominal predicate of the copula verb *sum*, whose subject is *substantia*, the tectogrammatical tree includes the

[10]Translation taken from: `https://isidore.co/aquinas/ContraGentiles1.htm#3`.

semantic roles played by the content words of the sentence, reporting for instance that *substantia* is the Actor of the clause.[11]

By using the resources in the CIRCSE collection, it is also possible to perform etymological inquires. For example, the user can retrieve the Proto-Italic and Proto-Indo-European reconstructed forms that explain the history of the lemma *classis* ("class/army") from the EDLIL, that is **klāssi-* and **klh$_1$-d$^{(h)}$-ti-* respectively. In addition, the same lemma is described as a loanword of the Ancient Greek noun κλῆσις in the IGVLL.

References

Bolognesi, G. 1999. La «linguistica computazionale» nell'Università Cattolica del S. Cuore e l'origine del termine informatica. *Aevum*, 73:913–920.

Busa, R. 1974–1980. *Index Thomisticus*. Frommann-Holzboog, Stuttgart-Bad Cannstatt.

de Vaan, M. 2008. Etymological Dictionary of Latin: and the other Italic Languages. Brill, Amsterdam https://brill.com/view/title/12612.

Franzini, G., Zampedri, F., Passarotti, M., Mambrini, F., and Moretti, G. 2020. Græcissâre: Ancient Greek Loanwords in the LiLa Knowledge Base of Linguistic Resources for Latin. In Dell'Orletta, F., Monti, J., and Tamburini, F. (eds.), *Proceedings of the Seventh Italian Conference on Computational Linguistics*. Accademia university press, Collana dell'Associazione Italiana di Linguistica Computazionale.

Litta, E., Passarotti, M., and Culy, C. 2016. Formatio formosa est. Building a Word Formation Lexicon for Latin. In Corazza, A., Montemagni, S., and Semeraro, G. (eds.), *Proceedings of the Third Italian Conference on Computational Linguistics*. Accademia university press, Collana dell'Associazione Italiana di Linguistica Computazionale, vol. 2, Napoli, 185–189.

Mambrini, F. and Passarotti, M. 2020. Representing etymology in the LiLa knowledge base of linguistic resources for Latin. In *Proceedings of the 2020 Globalex Workshop on Linked Lexicography*. European Language Resources Association, 20–28.

Passarotti, M. 2019. The Project of the Index Thomisticus Treebank. In Berti M. (ed.), *Digital Classical Philology. Ancient Greek and Latin in the Digital Revolution*. De Gruyter GmbH, Berlin-Boston, 299–319.

Passarotti, M., Gonzalez Saavedra, B., and Onambele, C. 2016. Latin Vallex. A Treebank-based Semantic Valency Lexicon for Latin. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation*. European Language Resources Association (ELRA), Portorož, Slovenia, 2599–2606.

Passarotti, M., Mambrini, F., Franzini, G., Cecchini, F.M., Litta, E., Moretti, G., Ruffolo, P., and Sprugnoli, R. 2019. Interlinking through Lemmas. The Lexical Collection of the LiLa Knowledge Base of Linguistic Resources for Latin. *Studi e Saggi Linguistici*, LVIII(1): 177–212.

Sgall, P., Hajicová, E., and Panevová, J. 1986. *The meaning of the sentence in its semantic and pragmatic aspects*. Reidel, Dordrecht.

Sprugnoli, R., Mambrini, F., Moretti, G., and Passarotti, M. 2020a. Towards the Modeling of Polarity in a Latin Knowledge Base. In *Proceedings of the Third Workshop on Humanities in the Semantic Web*. CEUR Workshop Proceedings, Heraklion, Greece, 59–70.

Sprugnoli, R., Passarotti, M., Cecchini, F.M., and Pellegrini, M. 2020b. Overview of the EvaLatin 2020 Evaluation Campaign. In Sprugnoli, R. and Passarotti, M. (eds.), *Proceedings of LT4HALA 2020 Workshop - 1st Workshop on Language Technologies for Historical and Ancient Languages*. European Language Resources Association (ELRA), Paris, 105–110.

Sprugnoli, R., Passarotti, M., Corbetta, D., and Peverelli, A. 2020c. Odi et Amo. Creating, Evaluating and Extending Sentiment Lexicons for Latin.. In *Proceedings of the 12th Language Resources and Evaluation Conference*. European Language Resources Association (ELRA), Paris, 3078–3086.

Straka, M. and Straková, J. 2017. Tokenizing, PoS tagging, lemmatizing and parsing UD 2.0 with UDPipe. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*. Association for Computational Linguistics, 88–99.

[11] The Actor role is semantically quite underspecified in the annotation guidelines of the Prague Dependency Treebank, which define it as "the human or non-human originator of the event, the bearer of the event or a quality/property, the experiencer or possessor" (`https://ufal.mff.cuni.cz/pdt2.0/doc/manuals/en/t-layer/html/ch07s02s01.html`).