Linked Open Treebanks.
Interlinking Syntactically Annotated Corpora in the LiLa Knowledge Base of Linguistic Resources for Latin

Francesco Mambrini, Marco Passarotti
CIRCSE Research Centre
Università Cattolica del Sacro Cuore
Largo Gemelli, 1 - 20123 Milan, Italy
{francesco.mambrini}{marco.passarotti}@unicatt.it

Abstract

In spite of the current availability of large collections of treebanks that can be used and queried from one common place on the web, we are still far from achieving a real interconnection, both between treebanks themselves and with other (kinds of) linguistic resources. However, making resources interoperable is a crucial requirement to maximize the contribution of each single resource, as well as to account for the linguistic complexity of the texts provided by (annotated) corpora and particularly by treebanks. This paper describes how dependency treebanks are interlinked in a Knowledge Base of linguistic resources for Latin based on Linked Open Data practices and standards. The Knowledge base is built to make linguistic resources interact by integrating all types of annotation applied to a particular word/text into a common representation.

1 Introduction and Motivation

Dependency treebanks for Latin have a history that goes back to 2006. For it was in that year that the first two projects kicked off: the Latin Dependency Treebank (LDT) (Bamman and Crane, 2006), featuring a small selection of texts by Classical authors (currently around 50k nodes), and the *Index Thomisticus* Treebank (IT-TB) (Passarotti, 2011), based on works written in the XIIIth century by Thomas Aquinas (approximately 400k nodes). Later on, a third Latin treebank was created in the context of the PROIEL project (Haug and Jøhndal, 2008), which includes the entire New Testament in Latin (the so called *Vulgata* by Jerome) and texts from the Classical era (for a total of around 250k nodes). Most recently, a syntactically annotated corpus of original VIIIth-IXth century charters from Central Italy, called Late Latin Charter Treebank (LLCT; around 250k nodes), was made available (Korkiakangas and Passarotti, 2011). While the LDT, the IT-TB and the LLCT have shared the same manual for syntactic annotation since the beginning of their respective projects (Bamman et al., 2007), the PROIEL treebank follows a slightly different style (Haug, 2010). Currently, all the Latin treebanks except the LLCT are available also in the Universal Dependencies collection (UD) (Nivre et al., 2016).[1]

The existence of four treebanks for an ancient language like Latin is not surprising, reflecting the large diachronic (as well as diatopic) span of Latin texts, which are spread across a time frame of more than two millennia and in most areas of the Mediterranean and of what is called Europe today. Since Latin has represented for a long time a kind of *lingua franca*, the variety of its textual typologies is wide, including scientific treaties, literary works, philosophical texts and official documents. This aspect makes it impossible to build one textual corpus that alone can be sufficiently representative of "Latin", just because there are too many varieties of Latin, which can be even very different from each other.[2]

In order to cope with such a large variety, several collections of Latin texts are today available in digital format, like for instance the *Perseus Digital Library* [3] and the collection of Medieval Italian Latinity *ALIM*.[4]

Besides textual resources, the centuries-old tradition of Latin lexicography resulted in the current availability of several digitized dictionaries, like for instance the Lewis-Short dictionary available at Perseus

[1] http://universaldependencies.org/

[2] For instance, Ponti and Passarotti (2016) show the dramatic decrease of the accuracy rates of a dependency parsing pipeline trained on the IT-TB when applied on texts of the Classical era taken from the LDT.

[3] http://www.perseus.tufts.edu/hopper/

[4] http://www.alim.dfll.univr.it/

and the *Thesaurus Linguae Latinae* by the Bayerische Akademie der Wissenschaften in Munich.[5] A small *Latin WordNet* including around 9,000 lemmas is also available (Minozzi, 2010), as well as a derivational morphology lexicon called *Word Formation Latin* (WFL) (Litta et al., 2016).

Just like for most other (both modern and ancient) languages, the interoperability issues imposed by the different formats, tag sets and annotation criteria of the linguistic resources for Latin severely limit their potential for exploitation and use. Indeed, linking linguistic resources to one another would maximize their contribution to linguistic analysis at multiple levels, be those lexical, morphological, syntactic, semantic or pragmatic. Thus, presently there is a growing interest in the interoperability of (annotated) corpora, lexical resources and Natural Language Processing (NLP) tools (Ide and Pustejovsky, 2010). So far, this was partially approached by building large infrastructures and databases of linguistic resources, like CLARIN,[6] DARIAH,[7] META-SHARE,[8] and EAGLE.[9] In the treebank area, the UD collection includes more than 100 treebanks sharing the same annotation guidelines and provides different tools for querying the treebanks on-line.[10] A relevant initiative of this kind is the Norwegian *Infrastructure for the Exploration of Syntax and Semantics* (INESS) (Rosén et al., 2012), which offers an open and easy-to-use platform for building, accessing, searching and visualizing treebanks through a web browser.[11]

These collections and infrastructures enable to use and query various resources and tools from one common place on the web, but they do not provide a real interconnection between them, thus failing to achieve their interoperability. Instead, making linguistic resources interoperable requires that all types of annotation applied to a particular word/text get integrated into a common representation that enables access to the linguistic information conveyed in a linguistic resource or produced by an NLP tool (Chiarcos, 2012, p. 162). Particularly, by applying the principles of Linked Data to linguistic resources[12] "it is possible to follow links between existing resources to find other, related data and exploit network effects" (Chiarcos et al., 2013, p. iii).[13] Despite their rich annotation (ranging from tokenization to syntactic analysis), treebanks alone cannot account for the linguistic complexity of the texts they include, which requires that information provided by different (and currently available) textual and lexical resources is interlinked and, thus, exploited to the best.

To this aim, the *LiLa: Linking Latin* project (2018-2023)[14] was launched with the objective to interlink the wealth of linguistic resources and NLP tools for Latin developed thus far, in order to bridge the gap between raw language data, NLP and knowledge description (Declerck et al., 2012, p. 111). LiLa addresses this challenge by building a collection of several data sets described using the same vocabulary and linked together, namely a Linked (Open) Data Knowledge Base of the linguistic resources (and NLP tools) for Latin currently available from different providers under various licences.

After a brief description of the basic architecture of the LiLa Knowledge Base (Section 2), this paper focuses on the inclusion of three dependency treebanks for Latin into LiLa (namely, the IT-TB in two versions, PROIEL and the LLCT), presenting an example of a complex query crossing the treebanks and the other linguistic resources included so far in the Knowledge Base (Section 3).

2 The LiLa Knowledge Base

In order to achieve interoperability between linguistic resources and NLP tools, the LiLa Knowledge Base makes use of a set of Semantic Web and Linguistic Linked Open Data standards. These include ontologies to

[5] http://www.thesaurus.badw.de/

[6] http://www.clarin.eu

[7] http://www.dariah.eu

[8] http://www.meta-share.org/

[9] http://www.eagle-network.eu

[10] SETS treebank search (http://bionlp-www.utu.fi/dep_search); PML Tree Query (http://lindat.mff.cuni.cz/services/pmltq/); Kontext (http://lindat.mff.cuni.cz/services/kontext/corpora/corplist); Grew-match (http://match.grew.fr/).

[11] http://clarino.uib.no/iness/page

[12] See Tim Berners-Lee's note at https://www.w3.org/DesignIssues/LinkedData.html.

[13] The *Linguistic Linked Open Data cloud* http://linguistic-lod.org/llod-cloud is a good example of a set of interconnected linguistic resources.

[14] https://lila-erc.eu/

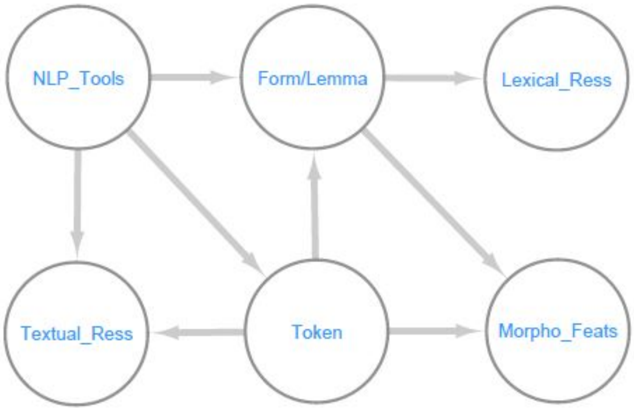

Figure 1: The basic architecture of the LiLa Knowledge Base.

describe linguistic annotation (OLiA (Chiarcos and Sukhareva, 2015)), corpus annotation (NIF (Hellmann et al., 2013), conll-rdf (Chiarcos and Fäth, 2017)) and lexical resources (Lemon (Buitelaar et al., 2011), Ontolex[15]). The Resource Description Framework (RDF) (Lassila et al., 1998) is used to encode graph-based data structures to represent linguistic annotations in terms of triples, made of a predicate connecting two nodes (a subject and its object). The SPARQL language is used to query the data recorded in the form of RDF triples (Prud'Hommeaux et al., 2008).

The LiLa Knowledge Base is highly lexically-based, striking a balance between feasibility and granularity: its basic assumption is that textual resources are made of (occurrences of) words, lexical resources describe properties of words, and NLP tools process words. Figure 1 presents the basic architecture of the LiLa Knowledge Base, showing its main components and their relations. The **Lemma** is the key node type in LiLa. A Lemma is an (inflected) **Form** conventionally chosen as the citation form of a lexical item. Lemmas occur in **Lexical Resources** as canonical forms of lexical entries. Forms, too, can occur in lexical resources, like in a lexicon containing all of the forms of a language, as for instance in Tombeur (1998). The occurrences of Forms in real texts are **Tokens**, which are provided by **Textual Resources**. Finally, **NLP tools** process either Forms regardless of their contextual use (e.g., a morphological analyzer), or Tokens (e.g., a PoS-tagger), or texts in Textual Resources (e.g., a tokenizer). Forms, Lemmas and Tokens can be assigned **Morphological Features**, like part of speech and gender.

Since lemmas serve as the optimal interface between lexical resources, (annotated) corpora and NLP tools, the core of the LiLa Knowledge Base is a collection of citation forms for Latin. Interoperability can be achieved by linking the entries in lexical resources and the corpus tokens pointing to the same lemma.[16] The collection of citation forms of LiLa is built on top of the set of lemmas used by the morphological analyzer for Latin Lemlat (Passarotti et al., 2017).[17] Lemlat relies on a lexical basis resulting from the collation of three Latin dictionaries (Georges and Georges, 1913 1918; Glare, 1982; Gradenwitz, 1904) for a total of 40,014 lexical entries and 43,432 lemmas, as more than one lemma can be included in one lexical entry. This lexical basis was recently further enlarged by adding the *Onomasticon* provided by the 5th edition of Forcellini dictionary (Budassi and Passarotti, 2016) and the entries from a large reference glossary for Medieval Latin, namely the *Glossarium Mediae et Infimae Latinitatis* (du Cange et al., 1883 1887; Cecchini et al., 2018), leading to a total of around 150,000 lemmas.

The linguistic resources currently linked in the LiLa Knowledge Base are stored in a triplestore using the Jena framework.[18] The Fuseki component exposes the data as a SPARQL end-point accessible over HTTP. The current prototype of the LiLa RDF triplestore database connects the following resources for Latin: (a) the collection of lemmas provided by Lemlat, (b) the WFL lexicon, and (c) three treebanks (four by version):

[15] https://www.w3.org/community/ontolex/

[16] On the process of harmonization of the different lemmatization strategies for Latin in LiLa, see Mambrini and Passarotti (Forthcoming).

[17] https://github.com/CIRCSE/LEMLAT3

[18] A prototype of the LiLa triple store is available at https://lila-erc.eu/data/.

(c.1) PROIEL in its UD version (release 2.3), (c.2-3) the IT-TB in both its UD 2.3 and original version, and (c.4) a selection of 3,900 sentences (105,380 tokens) of the LLCT.

3 Interlinking and Querying Treebanks in LiLa

In this section, we discuss how we integrated the Latin treebanks into the LiLa Knowledge Base and how the linked data obtained by connecting the treebank tokens to the other resources support complex queries crossing through different linguistic resources.

3.1 Linked Treebanks

The Latin treebanks currently integrated into LiLa have been converted into RDF triples. As an example, Figure 2 represents a first result in the conversion and linking process. The figure shows a three-word sentence from the *Vulgata* (*Matt.* 6.10), taken from the UD 2.3 version of the PROIEL corpus: *veniat regnum tuum* ("thy kingdom come"). The UD 2.3 tree for this sentence is shown in Figure 3.[19]

Tokens and sentences are defined using the NIF vocabulary. In the current, preliminary stage of the Knowledge Base, some information on the tokens, such as the list of morphological features, is still registered as a simple string of text. For instance, in Figure 2 this is the case of the string "Case=Nom|Gender=Neut|Number=Sing", which is linked to the PROIEL token with ID `s15924_2` (for the word *regnum* "kingdom") via the relation `conll:FEAT`, linking the morphological features taken from files in the CoNLL-U format of UD.[20]

Other types of tagging (such as syntactic dependencies, or sentence boundaries) are expressed by links between the nodes for tokens or sentences. For example, in Figure 2, this is represented by the linking between the token `s15924_2` (*regnum*) and the token `s15924_1` (*veniat* "come") via the relation `conll:HEAD`, representing that in the sentence the word *veniat* is the head of the word *regnum*, as can be seen from the tree in Figure 3.

Finally, a third group of linguistic annotations, like the part of speech, directly relate tokens to concepts from an ontology of linguistic data (OLiA).[21] In Figure 2, this is shown by the edge connecting the token `s15924_2` (*regnum*) to the concept node `olia:CommonNoun`.

Tokens are connected to the appropriate Lemma nodes recorded in the LiLa Knowledge Base. In Figure 2, for instance, the token `s15924_2` (*regnum*) is linked to lemma `34146`, which has written representation *regnum*. Via this connection, it becomes possible to access all the other information that is also pointing to that lemma. In the figure, the lemma `34146` is connected to a node for a lexical base (`1133`), the same to which also lemmas *rex* "king" (`34799`) and *regno* "to rule, to be king" (`34145`) are attached. This means that lemmas *regno*, *regnum* and *rex* belong to the same "word formation family", i.e. a set of lemmas sharing the same lexical base. The lemma *regnum* is also formed with the suffix "-n" (represented by the node `affix:111` in Figure 2), the same found in e.g. *fanum* "shrine" (not shown here for reasons of space). In the collection of citation forms included in LiLa, all the lemmas formed with the suffix "-n" are linked to `affix:111` via the relation `lemlat_base:hasSuffix`, thus allowing to retrieve them in the Knowledge Base. The information about lexical bases and affixes is available thanks to the connection of the WFL lexicon in LiLa.

3.2 Querying LiLa

In this section, we provide an example of the types of queries that the LiLa Knowledge Base can already support. As mentioned, one single query can extract data from all the multiple corpora and lexical resources linked to LiLa's collection, and can also combine syntactic, lexical and morphological information beyond the type of annotation explicitly recorded in a single corpus.

[19] In Figure 3, each node apart from the root is assigned its part of speech and a dependency relation. In the tree, the `nsubj` relation is used for nominal subjects, while `nmod` for nominal modifiers. The full list of dependency relations used in UD v2 is available at `https://universaldependencies.org/u/dep/index.html`.

[20] On the CoNLL-U format used in the UD treebanks see `https://universaldependencies.org/format.html`.

[21] A shallow conversion from the CoNLL-U format to RDF was obtained with the help of conll-rdf. The application also allows to design custom SPARQL Update queries to link the RDF representation of the corpus to other resources.

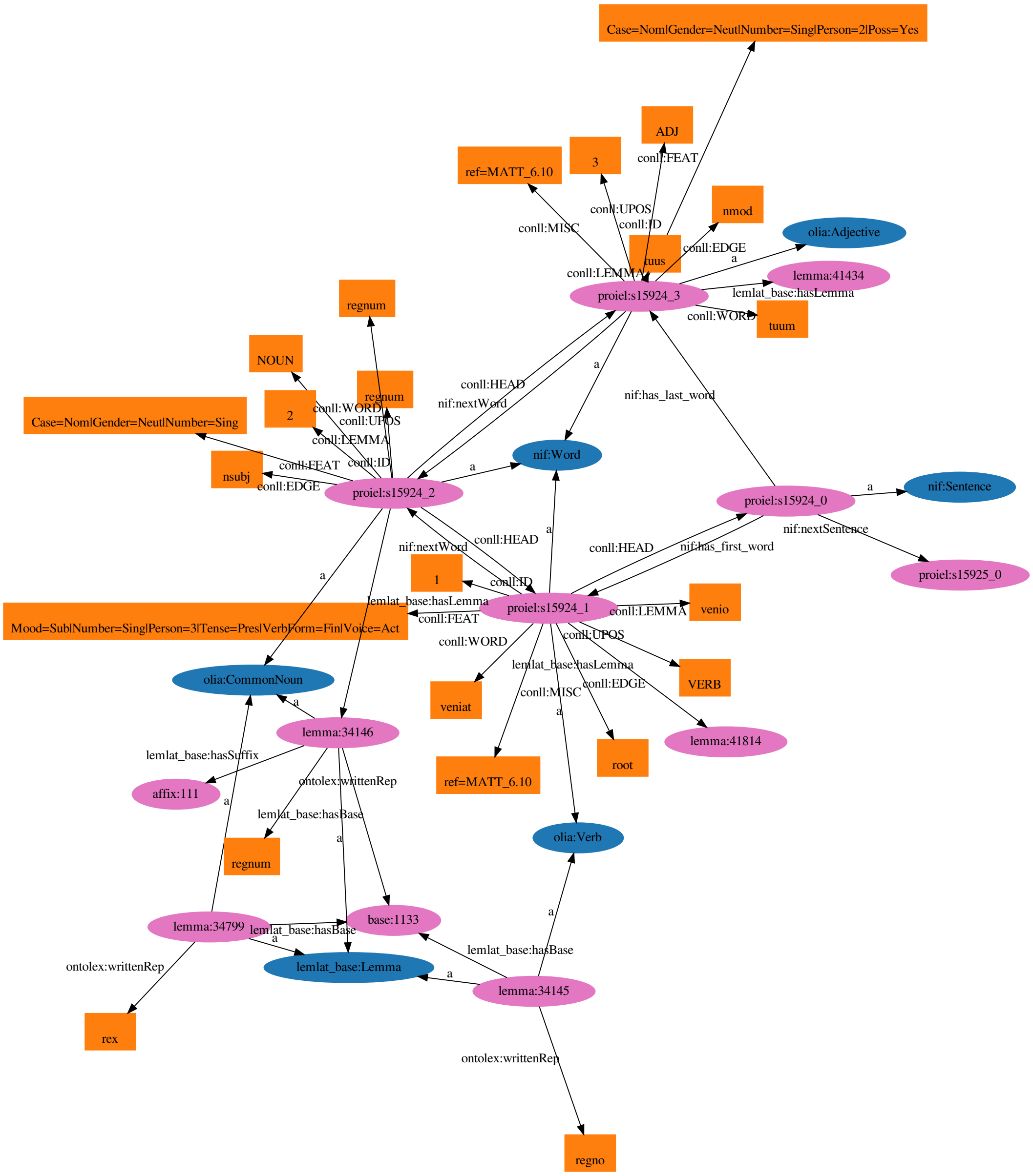

Figure 2: A sentence from PROIEL as RDF triples in the LiLa Knowledge Base.

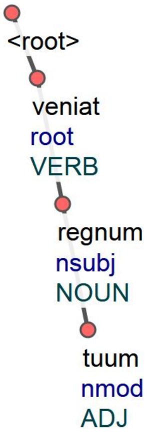

Figure 3: The UD 2.3 tree of *veniat regnum tuum* from PROIEL.

Consider, for instance, the case of a researcher interested in the relation between the syntactic role of subject and the semantic role of agent in Latin. One possible approach to study the question would be to start by collecting and analyzing the sentences where nouns formed with a typical morpheme for agent nouns like "-(t)or" (common to several Indo-European languages) are attested as subject of an active verb.

Though the number of linguistic resources currently interlinked in LiLa is still small, it is already possible to design a single SPARQL query to extract this information from our RDF versions of PROIEL, IT-TB (UD version) and LLCT. In what follows, we illustrate the results of a query that asks for an active (or deponent) verb governing a noun with the syntactic relation of subject in the three treebanks. By leveraging the connection between lemmas and the affixes in WFL, we add the additional constraint that the noun must be formed with the suffix "-(t)or". This information, which is not encoded into the original treebanks, is now accessible thanks to the architecture based on Linked Open Data that LiLa adopts.

The query allows us to extract 143 passages, with 80 different verbs and 58 agent nouns. One sample of the results, a sentence from Cicero's *Letters to Atticus* (4.4a.2) retrieved from PROIEL, is reported in Example (1).

(1) **gladiatores** audio **pugnare** mirifice.
'I hear that your gladiators fight superbly.'

The subject-verb bigrams resulting from the query highlight interest lexical aspects in the language of the three corpora. As it is to be expected from the documentary nature of the texts provided by the LLCT treebank, the 10 occurrences found in this corpus all involve legal actors and events: the most frequent subject (4 occurrences) is *rector*, the priest responsible for a rural church. The other actors are: *dispensator* "treasurer", *fideiussor* "bail", *genitor* "parent" and *imperator* "emperor".

In the ITTB, on the other hand, the most frequent couplet is the one formed by the noun *commentator* "interpreter" and the verb *dico* "to say" (21 cases), where the assertions of a scholar are reported and discussed. Indeed, the verbs pointing to intellectual activities of scholars dominate in the results from the corpus of Thomas Aquinas: in addition to the most frequent *dico* (22), other intellectual verbs include *respondeo* "to reply" (3 instances), *fingo* "to imagine" (2), and *intendo* "to mean" (2).

Finally, PROIEL, which is more balanced between different genres, offers a more varied set of subject-verb couplets in its 57 results. As in Example (1), where the noun *gladiator* "gladiator" is coupled with the verb *pugnare* "to fight", we find several nouns and verbs from everyday life, or from the domain of the professions and human activities. Thus, for instance, we find 4 cases of *fossor* "digger, ditcher" joined with verbs like *includo* "to shut in" and *incumbo* "to press upon", or 6 cases of *pastor* "herdsman, shepherd" with verbs like *fugio* "to flee" and *secludo* "to shut off".

4 Conclusions and Future Work

In this paper, we have described how we interlinked three dependency treebanks for Latin (one available in two versions) into a Knowledge Base of linguistic resources based on Linked Open Data practices and standards. Linking resources of different kind (such as corpora and lexica) makes it possible to exploit their potential to the best. Indeed, single resources tend to focus on a limited set of linguistic features (e.g. morphology and syntax for treebanks), which are in most cases insufficient to provide a full analysis of the textual or lexical data. Making interoperable the still scattered and unconnected resources that are currently available for Latin (as well as for many other languages) is a way to approach the data from the various layers of annotation that such resources provide.

Our work of interlinking the linguistic resources for Latin has just begun. In the near future, we plan to integrate into the LiLa Knowledge Base two other lexical resources, namely an etymological dictionary (de Vaan, 2008) and the Latin WordNet. Interlinking these resources with the textual occurrences of their lemmas (enriched with syntactic annotation in treebanks) will enable the users of LiLa to run complex queries crossing different kinds of linguistic features. Given that the set of interlinked resources will grow in the coming years, the chain of connection can be continued indefinitely; as long as new lexical resources are connected to the Knowledge Base, all the connections from any corpus token to their nodes will become explorable in the network.

Acknowledgements

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme - Grant Agreement No 769994.

References

David Bamman and Gregory Crane. 2006. The design and use of a Latin dependency treebank. In *Proceedings of the Fifth Workshop on Treebanks and Linguistic Theories (TLT 2006)*, pages 67–78, Prague, Czech Republic. Univerzita Karlova.

David Bamman, Marco Passarotti, Gregory Crane, and Savina Raynaud. 2007. Guidelines for the syntactic annotation of latin treebanks. *Tufts University Digital Library*.

Marco Budassi and Marco Passarotti. 2016. Nomen omen. Enhancing the Latin morphological analyser Lemlat with an onomasticon. In *Proceedings of the 10th SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH)*, pages 90–94, Berlin, Germany. Association for Computational Linguistics.

Paul Buitelaar, Philipp Cimiano, John McCrae, Elena Montiel-Ponsoda, and Thierry Declerck. 2011. Ontology lexicalisation: The lemon perspective. In *Proceedings of the Workshops. 9th International Conference on Terminology and Artificial Intelligence*, pages 33–36.

Flavio Cecchini, Marco Passarotti, Paolo Ruffolo, Marinella Testori, Lia Draetta, Martina Fieromonte, Annarita Liano, Costanza Marini, and Giovanni Piantanida. 2018. Enhancing the latin morphological analyser lemlat with a medieval latin glossary. In *Proceedings of the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018). 10-12 December 2018, Torino*, pages 87–92.

Christian Chiarcos and Christian Fäth. 2017. CoNLL-RDF: Linked Corpora Done in an NLP-Friendly Way. In Jorge Gracia, Francis Bond, John P. McCrae, Paul Buitelaar, Christian Chiarcos, and Sebastian Hellmann, editors, *Language, Data, and Knowledge*, pages 74–88, Cham. Springer International Publishing.

Christian Chiarcos and Maria Sukhareva. 2015. OLiA - Ontologies of Linguistic Annotation. *Semantic Web Journal*, 6(4):379–386.

Christian Chiarcos, Philipp Cimiano, Thierry Declerck, and John P McCrae. 2013. Linguistic linked open data (llod). introduction and overview. In *Proceedings of the 2nd Workshop on Linked Data in Linguistics (LDL-2013): Representing and linking lexicons, terminologies and other language data*, pages i–xi.

Christian Chiarcos. 2012. Interoperability of corpora and annotations. In *Linked Data in Linguistics*, pages 161–179. Springer.

Michiel de Vaan. 2008. *Etymological Dictionary of Latin and the other Italic Languages*. Leiden & Boston: Brill.

Thierry Declerck, Piroska Lendvai, Karlheinz Mörth, Gerhard Budin, and Tamás Váradi. 2012. Towards linked language data for digital humanities. In *Linked Data in Linguistics*, pages 109–116. Springer.

Charles du Fresne du Cange, Bénédictins de Saint-Maur, Pierre Carpentier, Louis Henschel, and Léopold Favre. 1883–1887. *Glossarium mediae et infimae latinitatis*. Niort, France.

Karl Ernst Georges and Heinrich Georges. 1913–1918. *Ausführliches lateinisch-deutsches Handwörterbuch*. Hahn, Hannover, Germany.

Peter GW Glare. 1982. *Oxford Latin dictionary*. Clarendon Press. Oxford University Press, Oxford, UK.

Otto Gradenwitz. 1904. *Laterculi Vocum Latinarum: voces Latinas et a fronte et a tergo ordinandas*. Hirzel, Leipzig, Germany.

Dag TT Haug and Marius Jøhndal. 2008. Creating a parallel treebank of the old Indo-European Bible translations. In *Proceedings of the Second Workshop on Language Technology for Cultural Heritage Data (LaTeCH 2008)*, pages 27–34, Marrakesh, Morocco. European Language Resources Association (ELRA).

Dag Haug. 2010. Proiel guidelines for annotation. *Retrieved April*, 23:2013.

Sebastian Hellmann, Jens Lehmann, Sören Auer, and Martin Brümmer. 2013. Integrating NLP using Linked Data. In *12th International Semantic Web Conference, Sydney, Australia, October 21-25, 2013*.

Nancy Ide and James Pustejovsky. 2010. What does interoperability mean, anyway. *Toward an Operational.*

Timo Korkiakangas and Marco Passarotti. 2011. Challenges in annotating medieval latin charters. *Journal for Language Technology and Computational Linguistics*, 26(2):103–114.

Ora Lassila, Ralph R. Swick, World Wide, and Web Consortium. 1998. Resource description framework (rdf) model and syntax specification.

Eleonora Litta, Marco Passarotti, and Chris Culy. 2016. Formatio formosa est. building a word formation lexicon for latin. In *Proceedings of the third italian conference on computational linguistics (clic–it 2016)*, pages 185–189.

Francesco Mambrini and Marco Passarotti. Forthcoming. Harmonizing different lemmatization strategies for building a knowledge base of linguistic resources for latin. In *Proceedings of the 13th Linguistic Annotation Workshop (LAW XIII)*, Florence, Italy.

Stefano Minozzi. 2010. The latin wordnet project. In *Latin Linguistics Today. Latin Linguistics Today. Akten des 15. Internationalen Kolloquiums zur Lateinischen Linguistik*, pages 707–716.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajič, Christopher Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. 2016. Universal Dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 1659–1666, Portorož, Slovenia. European Language Resources Association (ELRA).

Marco Passarotti, Marco Budassi, Eleonora Litta, and Paolo Ruffolo. 2017. The Lemlat 3.0 Package for Morphological Analysis of Latin. In *Proceedings of the NoDaLiDa 2017 Workshop on Processing Historical Language*, volume 133, pages 24–31. Linköping University Electronic Press.

Marco Passarotti. 2011. Language resources. The state of the art of Latin and the *Index Thomisticus* treebank project. In Marie-Sol Ortola, editor, *Corpus anciens et Bases de données*, number 2 in ALIENTO. Échanges sapientiels en Méditerranée, pages 301–320, Nancy, France. Presses universitaires de Nancy.

Edoardo Maria Ponti and Marco Passarotti. 2016. Differentia compositionem facit. A slower-paced and reliable parser for Latin. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 683–688, Portorož, Slovenia. European Language Resources Association (ELRA).

Eric Prud'Hommeaux, Andy Seaborne, et al. 2008. Sparql query language for rdf. w3c. *Internet: https://www.w3.org/TR/rdf-sparql-query/[Accessed on February 27th, 2019].*

Victoria Rosén, Koenraad De Smedt, Paul Meurer, and Helge Dyvik. 2012. An open infrastructure for advanced treebanking. In *META-RESEARCH Workshop on Advanced Treebanking at LREC2012*, pages 22–29. Hajič, Jan.

Paul Tombeur. 1998. *Thesaurus formarum totius Latinitatis: a Plauto usque ad saeculum XXum.* Turnhout: Brepols.