Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 4022–4029
Marseille, 20-25 June 2022

The Index Thomisticus Treebank as Linked Data in the LiLa Knowledge Base

Francesco Mambrini, Marco Passarotti, Giovanni Moretti, Matteo Pellegrini
Università Cattolica del Sacro Cuore
Milan, Italy
{francesco.mambrini,marco.passarotti,giovanni.moretti,matteo.pellegrini}@unicatt.it

Abstract

Although the Universal Dependencies initiative today allows for cross-linguistically consistent annotation of morphology and syntax in treebanks for several languages, syntactically annotated corpora are not yet interoperable with many lexical resources that describe properties of the words that occur therein. In order to cope with such limitation, we propose to adopt the principles of the Linguistic Linked Open Data community, to describe and publish dependency treebanks as LLOD. In particular, this paper illustrates the approach pursued in the LiLa Knowledge Base, which enables interoperability between corpora and lexical resources for Latin, to publish as Linguistic Linked Open Data the annotation layers of two versions of a Medieval Latin treebank (the Index Thomisticus Treebank).

Keywords: Treebank, Linguistic Linked Open Data, Latin

1. Introduction and Motivation

Linguistic annotation on textual corpora is an invaluable support for studying historical languages like Latin. Language learning and corpus-based research are the two most obvious applications. Since Latin has a very rich morphology, lemmatization and morphological-feature annotation are particularly important for tasks like word search or for vocabulary acquisition. The distinctively "free(er)" word order of Latin, as compared to many modern languages like Italian or English, also greatly complicates the syntactic analysis of texts for modern readers (or even for parsers fine-tuned for modern languages).

While services and libraries that support lemmatization and/or morphological analysis for Latin have been available for decades,[1] a series of Latin treebanks, with word-by-word account of the syntax and morphology of Latin texts, have been published only in recent years. Latin's long and rich history is well reflected also in the spectrum of existing treebanks, which include texts of different genres and periods. The treebank developed within the Perseus project (Bamman, D. et al., 2017), consists of a small (about 53,000 tokens) selection of texts of the Classical period, from 1st century BC to 1st century AD (Bamman and Crane, 2011); the Latin portion of the PROIEL treebanks (Haug, D. et al., 2018) contains the *Vulgata* by Jerome (4th century), plus some Classical and Late Latin texts, for a total of about 200,000 tokens (Haug and Jøhndal, 2008). The Late Latin Charter Treebank (LLCT) (Korkiakangas, T., 2020) is the only one featuring non-literary texts, as it is entirely composed of charters written in Tuscany in the 8th and 9th century, for a total of 242,000 tokens (Cecchini et al., 2020b). Moreover, two treebanks display texts from a single author, namely Thomas Aquinas (13th century) in the Index Thomisticus Treebank (ITTB) (Passarotti, M. et al., 2021), the largest of all Latin treebanks amounting at about 450,000 tokens (Passarotti, 2019) – and the Latin works by Dante Alighieri (13th-14th century) in UDante (Cecchini, F. et al., 2021) – about 55,000 tokens (Cecchini et al., 2020a).

The advantages of such syntactically annotated corpora are many, especially since they support more advanced applications like treebank-based linguistic studies (Korkiakangas, 2017), or the development of trained models for stochastic NLP tools (Ponti and Passarotti, 2016). Still, even not considering the non-trivial investment of time and resources that their construction requires, the usability of the available treebanks is limited by a series of intrinsic factors. Firstly, projects adopt a variety of different formats, tagsets and guidelines for each level of annotation. The result is that the existing annotation cannot be queried consistently and simultaneously, but each query must be converted and adapted to the local schema of each project. Secondly, lexical resources that describe those very words that are attested in the corpora are not structurally connected to the treebanks. Thus, if readers want to obtain all the available information published on the web about a specific word used in a corpus (such as its meaning(s), translation(s) into several languages, the related WordNet synsets, etc.), they will have to perform a specific query using separate web services.

For treebank annotation, a successful answer to the first shortcoming is offered by the Universal Dependencies (UD) project (Nivre et al., 2016). UD is an open-access and collaborative effort to allow for cross-linguistically consistent annotation of morphology (i.e., parts of

[1] See for instance: LEMLAT (Passarotti, M. et al., 2020), and the modules for Latin in UDPipe (Straka, M., and Straková, J., 2021) and CLTK (Johnson, K.P. et al., 2021). See Passarotti et al. (2020) for a more detailed list.

speech and other grammatical features) and syntax (using dependency relations) in treebanks for different languages. Currently, 122 languages are included in the project, for a total of 217 treebanks. The five Latin corpora cited above are all distributed with the latest version of UD (2.9) (Zeman, D. et al., 2021). Four of them (Perseus, PROIEL, ITTB, LLCT) are natively annotated in a specific formalism and then converted to the UD schema; UDante was natively annotated using the UD guidelines.

While UD does provide a suitable shared formalism for the simultaneous interrogation of the included treebanks, the second problem mentioned above is still not solved. Furthermore, while all non-native projects maintain specific converters to map the original annotation to the UD guidelines, no comprehensive alignment of the linguistic vocabularies in use is provided.

In order to cope with these limitations, we propose to adopt the principles of the Linguistic Linked Open Data (LLOD) community, as well as some of the ontologies in use to describe the published LLOD. This paper illustrates the approach adopted in the context of the "LiLa - Linking Latin" project[2] to publish all the annotation layers of the ITTB as LLOD; as the ITTB is available in two different annotation schemes, namely the original and its conversion to UD, we also show how both sets of syntactic annotations have been linked to the corpus tokens. Furthermore, since the treebanks published in LiLa make extensive use of ontologies developed by the LLOD community (presented in Section 2), we discuss how this proves helpful to improve the conceptual interoperability of UD's linguistic annotation.

The paper is organized as follows: Section 2 discusses some related works on the subject. Section 3 briefly describes the general architecture of the LiLa Knowledge Base, which enables the interoperability between corpora and lexical resources for Latin. Section 4 illustrates how the ITTB has been published as LLOD, focusing on the modeling of the corpus architecture (4.1), of syntactic annotation (4.3), and of morphological features (4.4). Section 5 presents a possible use case supported by our architecture, while Section 6 highlights the conclusions and the plans for future work.

2. Related Work: LLOD, Treebanks and Universal Dependencies

Given the success of UD and the growing popularity of LLOD to support interoperability between linguistic resources, several potential interactions between the two initiatives have already been explored.

The suite of tools CoNLL-RDF (Chiarcos and Fäth, 2017; Chiarcos et al., 2021) allows users to convert from the UD format CoNLL-U[3], as well as any other tab-separated columnar formats, to RDF triples. The output sentences and tokens are described using the NIF vocabulary (Hellmann et al., 2013). The software also supports the enrichment of the converted files with concepts from other LLOD ontologies, including OLiA.

The Ontologies of Linguistic Annotation (OLiA) is a set of OWL ontologies designed to mediate between different vocabularies used for corpora annotation (Chiarcos and Sukhareva, 2015b). Instead of relying on the idea of aggregating the existing terminologies into one central repository, or of promoting a centralized standard to supersede all local projects, OLiA leverages the power of Linked Data and of ontologies to link and harmonize the various terms into a wide network of linguistic concepts (Chiarcos et al., 2020). OLiA adopts a modular architecture: while an 'annotation model' defines the terms used locally by single projects and makes connections between them and the tagset explicit, the OLiA Reference Model formalizes a general ontology of linguistic concepts that the annotation models can refer to. A 'linking model' formulates the connection between the terms in the various annotation models and the general concepts in the OLiA Reference Model.

Chiarcos et al. (2020) present an attempt to formalize the vocabulary of UD in three different annotation models for OLiA, dedicated respectively to the UD tags for parts of speech, for morphological features and for dependency relations. As the UD guidelines[4] are constantly being discussed and revised, the authors decided to build their ontologies by scraping the concepts directly from the project's online documentation.[5]

Finally, Passos (2018) discusses and evaluates an OWL ontology with a formal specification for UD annotation, with the aim to support conversion and validation of annotated data. The ontology, however, is not published and limited to UD v.1.

At present, the LLOD Cloud (Chiarcos et al., 2012) includes several UD corpora, which however only display a shallow conversion to the RDF data model obtained via CoNLL-RDF, without any linking to vocabularies for annotation.[6]

In the present paper, we adopt the approach of Chiarcos et al. (2020) and we use the OLiA ontologies to model the annotations of the ITTB. Moreover, by linking the lemmatized tokens to the lemmas of the LiLa Knowledge Base, we connect the treebank (in its two formats) to the rest of the linguistic resources for Latin, both lexical and textual, that are also part of that network.

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

[3] https://universaldependencies.org/format.html

[4] https://universaldependencies.org/guidelines.html

[5] The web-scraper used is also distributed with the OWL files at https://github.com/acoli-repo/olia. Note that the project provides models for both versions 1 and 2 of UD.

[6] See https://linguistic-lod.org/llod-cloud. As for Jan. 15, 2022, however, not all links to the datasets are still active.

By doing this, we intend to build an enhanced model of publication of UD treebanks as LLOD.

3. The LiLa Project

The "LiLa - Linking Latin" project aims to reach interoperability between the wealth of existing lexical and textual resources that have been developed in the last decades for Latin. One of the main problems that LiLa intends to solve is the fact that such resources and tools are often characterized by different conceptual and structural models, which makes it difficult for them to interact with one another.

To this goal, LiLa has undertaken the creation of an open-ended Knowledge Base, following the principles and techniques of the Linked Data paradigm. All content involved or referenced in the linguistic resources that we connect is made unambiguously findable and accessible by assigning an HTTP Uniform Resource Identifier (URI) to each data point. Data reusability and interoperability between resources are achieved by establishing links between different URIs and by using web standards such as: [a] the RDF data model, which is based on triples: (i) a predicate-property connects (ii) a subject (a resource) with (iii) its object (another resource, or a literal) (Lassila and Swick, 1998); [b] SPARQL, a query language specifically devised for RDF data.[7] Furthermore, the LiLa Knowledge Base makes reference to classes and properties of already existing ontologies to model the relevant information. The main ones are POWLA for corpus data (Chiarcos, 2012), OLiA for linguistic annotation (Chiarcos and Sukhareva, 2015a), and OntoLex-Lemon for lexical data (Buitelaar et al., 2011; McCrae et al., 2017).

Within this framework, the core of the Knowledge Base is the so-called Lemma Bank,[8] a collection of lemmas – defined as the canonical form a lexical item, i.e. its citation form – taken from the database of the morphological analyzer LEMLAT (Passarotti, M. et al., 2020) (Passarotti et al., 2017). Textual and lexical resources are thus made interoperable by connecting their tokens and entries, respectively, to the corresponding lemma in the Lemma Bank.

4. The Index Thomisticus Treebank as Linguistic Linked Open Data

4.1. UD Corpora into LLOD

While CoNLL-RDF and NIF provide a very convenient model for transitioning from the CoNLL-U format to a graph model, as well as a flexible series of pointers to identify the annotated tokens, we have chosen to rely on a third ontology to model the corpus data themselves, namely POWLA (Chiarcos et al., 2012). In contrast to the more limited integration proposed by Chiarcos et al. (2021), our treebanks published as LLOD make a more extensive use of that vocabulary, on account of the several advantages that POWLA offers especially for dealing with ancient texts and canonical authors.

Indeed, firstly, POWLA provides classes and properties to describe the stratification of documents and subsections within a corpus. Whereas for most standard corpora in Computational Linguistics (and in UD) the internal subdivisions of the corpus might not be relevant, researchers in Historical Linguistics and Digital Humanities are sensible to, or even directly interested in, the differences between author and author or text and text.

Secondly, especially for heavily studied and frequently annotated works like the canon of Classical Latin literary authors, POWLA allows corpus providers to group different types of annotations in different layers. One consequence of this, which is particularly compelling when dealing with a treebank converted into UD from a different format, is the possibility to publish both morpho-syntactic annotations (i.e., both the original and the UD-converted treebank) as linked to the same underlying tokens and sentences, provided that both presuppose the same sentence splitting and tokenization. Our focus is therefore on representing the stratification of different layers of interpretation and annotation on the same corpus of texts, rather than in providing seamless conversion between the annotations serialized in CONLL-U format and linked data.

In our representation, the ITTB is an instance of the POWLA's `Corpus` class. The internal subdivisions of the corpora can be expressed by means of the instances of the class `Document` and the two inverse object properties 'has sub-document'/'has super-document'. In our case, the documents are represented by the works of the authors that are included in the collections, which, for the ITTB, means only Aquinas' *Summa Contra Gentiles* (SCG).

Each document can be linked to one or more layers of annotations. In the case of the treebank in question, we partition the morpho-syntactic annotation into three layers, that store parts of speech, morphological features, and dependency relations respectively. (Note that a fourth layer is used to group tokens that belong to the same unit within the citation hiearchy, e.g. paragraph 10 of chapter 8 of book 4). For the ITTB, the syntactic level is represented by two layers, one holding the syntactic relations based on the original schema, inspired by the analytical layer of the Prague Dependency Treebank (PDT) (Bohmová et al., 2001)[9], and the other the UD annotation.

The other two POWLA's classes that we use are `Node` and `Relation`. The former is used to define the annotated tokens as terminals and the sentences as root nodes. Individuals belonging to the latter class, on

[7] LiLa's SPARQL endpoint can be accessed at: `https://lila-erc.eu/sparql/`.

[8] `http://lila-erc.eu/lodview/data/id/lemma/LemmaBank`.

[9] `https://ufal.mff.cuni.cz/pdt2.0/doc/manuals/en/a-layer/html/index.html`

the other hand, represent a dependency relation, in the form of a tuple of head and dependent node, where the head is either a terminal or the root node of a sentence.

4.2. The OLiA UD Annotation Model

OLiA annotation models represent a tagset by instantiating the concepts in a series of OWL named individuals and corresponding classes. In the case of the UD v.2 vocabulary, the concepts used in each layer of annotation (denoted by the tags mandated by the UD guidelines) are expressed as classes, while the language-specific realization of each of them is defined as a named individual. Thus, for instance, the part of speech 'adposition' is a class,[10] while the concept of adposition as used in the annotation of the Italian UD treebanks[11] is an individual belonging to that class.

The localized Latin named individuals, representing the tags used in the UD annotation of Latin treebanks, are generally supported in the latest distribution of OLiA's UD annotation models, even though only a minority of them have a dedicated documentation page (from which, as said, the concepts in the OLiA model have been scraped). As a rule, the language-specific documentation is reserved to the very few language-specific expansions introduced in the Latin treebanks. Thus, several universal or widely used tags (e.g. the features for the ablative case, the feminine gender or indeed the POS adposition quoted above) do not have a dedicated web page in the UD guidelines, but they can be represented nonetheless in the annotation model. In such cases, the newly introduced terms follow the same naming convention as the other tags.[12]

4.3. Syntactic Annotation

The UD version of the ITTB includes 450,515 relations that connect dependent nodes to either other treebank nodes or the sentence root. These head-dependent arcs form the directed, acyclic syntactic graphs that represents the structure of the sentence. While the UD schema allows annotators to add additional relations (named "enhanced dependencies") that are not subject to the treeness constraint, this enhanced representation is not used in the current UD distribution of the ITTB.

The head-dependent relations are represented in our data as instances of the class `Dependency Relation`, which we define as a subclass of POWLA's `Relation`. Dependency relations are connected with head and dependent nodes via two dedicated properties, named "has head" and "has dependent", which specify POWLA's "has source" and "has target" properties.

[10] `https://universaldependencies.org/u/pos/ADP`.

[11] `https://universaldependencies.org/it/pos/ADP`.

[12] For instance, the localized 'Case=Abl' feature has the URI `https://universaldependencies.org/la/feat/Case#Abl`, even if a page with this URL does not currently exist in the UD domain.

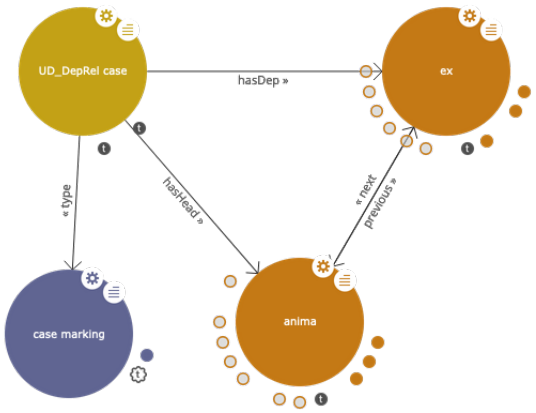

Figure 1: A syntactic relation in the UD schema.

While it would be possible to express the syntactic relation via an annotation node and the local Latin dependency tag (see below Section 4.4), we chose to define each relation as an instance of the appropriate OLiA class. To give an example, two nodes from the ITTB representing the prepositional phrase *ex anima* 'from (the) soul' (SCG 2.57.6) are linked in the UD representation by a syntactic relation where, according to the UD guidelines, the noun is governing the preposition via the *case* dependency relation.[13] This link is expressed in our LLOD representation as an instance both of a syntactic relation and of UD's "case marking" concept (Figure 1).[14] This modeling strategy is conceptually impeccable, because that specific relation is, in fact, precisely an instance of the "case marking" relation defined in UD. It also provides the advantage that we can directly link the relation with the class whose URI is defined within the UD namespace (`https://universaldependencies.org/u/dep/`). This allows us to bypass language-specific tags when they are not needed: such solution seems to be more in tune to the UD's aim of providing a unified tagset for language annotation.[15]

The same token for the preposition *ex* 'from' has a radically different set of dependency relations in the original format of the ITTB, where prepositions are licensed to govern the nouns; treatment of the preposition-noun nexus is in fact one of the main differences between

[13] `http://lila-erc.eu/data/corpora/ITTB/depAnnotation/UD/005.SCG*LB2.CP-5++7.N.-6.21-3.23-2W3`.

[14] `https://universaldependencies.org/u/dep/case`.

[15] Note that one *does* need language-specific classes, in case of relation subtypes that are not defined in the `https://universaldependencies.org/u/dep/` namespace. One such example for Latin would be the *ablativus asbsolutus* subtype of the adverbial clause, for which a documentation page is in fact available: `https://universaldependencies.org/la/dep/advcl-abs`.

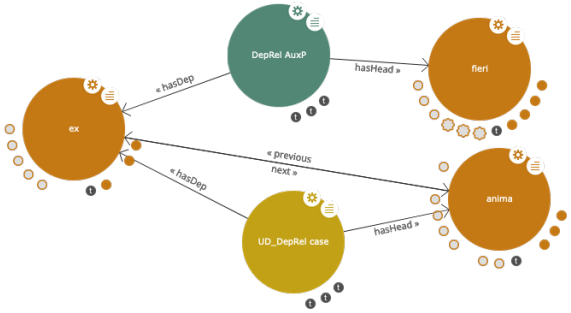

Figure 2: Two different syntactic annotations for a preposition in the ITTB.

the two schemes (Cecchini et al., 2018). In LiLa's representation it is possible to express these diverging interpretations of the preposition-noun construction with two sets of syntactic relations insisting on the same token. Figure 2 visualizes how the two syntactic interpretations are represented in LiLa; the preposition figures as the dependent of two syntactic relations: in one, from the UD treebank, the node that immediately follows it (*anima* 'soul') is the head; in the other, from the original ITTB, the preposition is governed by the verb *fieri* 'be produced/come into being'.

4.4. Morphological Features

In corpus annotation, morphological features are generally defined as additional morphological properties of tokens that are not captured by tags for parts of speech (POS).[16] In UD annotation, they are encoded as a set of key-value pairs, chosen from a list that includes the most commonly attested features and that can also be extended with language-specific expansions; the appropriate combination of features that describe a word is stored in a dedicated column of the CoNLL-U format. In our previous example, for instance, while the morphology of the preposition *ex* is fully accounted for with the POS annotation, the noun *anima* can further be defined as having the values 'feminine', 'singular', and 'ablative' for the features 'gender', 'number' and 'case' respectively. Accordingly, the corresponding line in the CoNLL-U file records the following string in the sixth column: `Case=Abl|Gender=Fem|Number=Sing`.

The CoNLL-RDF suite includes SPARQL queries that link corpus tokens to the universal POS of the UD schema via class instantiation.[17] No solution, however, is offered for morphological features. In fact, class membership does not seem to be the appropriate strategy to account for their relation to tokens. Instead, morphological features are better conceptualized as properties that are predicated about annotation units, rather than classes of tokens.

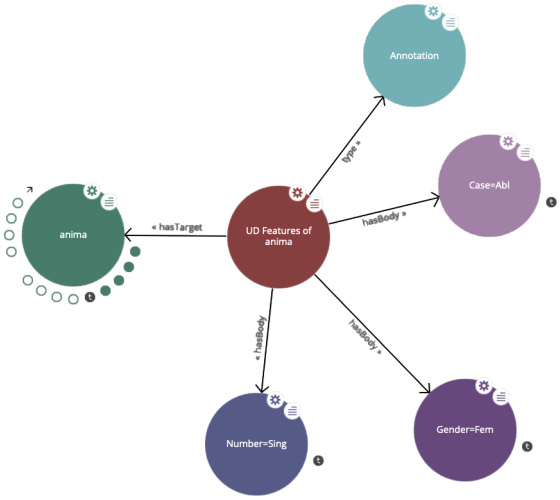

Figure 3: Morphological features of *anima* in the ITTB (SCG 2.57.6).

We propose to model the UD annotation of morphological features using the Web Annotation Data Model, as discussed by Cimiano et al. (2020, 66-69). According to it, annotations are reified as web resources that connect an annotation body (the content of the annotation itself) to a target (the annotated object); the same annotation can have multiple bodies (when it consists of various components) and multiple targets (when the same annotation is repeated on many objects). Considering that OLiA defines the key-value pairs used in UD as named individuals and that corpus tokens are web resources with their own URI, the simplest solution is to take the token as the target and the OLiA concept as the body of the annotation.

Figure 3 exemplifies this model with the usual example of *anima* in SCG 2.57.6 from the ITTB. The node representing the annotation (in the center of the graph) is linked to the target token (to the left) and to the three concepts from the OLiA UD Annotation Model. Just as multiple relations belonging to different schemas can link the same tokens, so different annotations, representing feature tags from various treebanks (such as the original ITTB and its UD version) can be connected to the same annotated words.

5. Use Cases

The biggest advantage that LLOD offers to treebank users is the possibility to query the morpho-syntactic relations together with information recorded in other resources connected to the same knowledge base (Mambrini and Passarotti, 2019). As said, the architecture that we have discussed provides the ad-

[16] This is the definition given by the UD guidelines: `https://universaldependencies.org/u/feat/index.html`.

[17] See the SPARQL files at: `https://github.com/acoli-repo/conll-rdf/tree/master/examples/sparql/link`.

ditional benefit of enabling users to compare multiple treebank annotations on the same tokens.

To give one example, the original ITTB formalism adopts a definition of objects (dependency relation: `Obj`)[18] that is based on verb valency and the distinction of arguments and adjuncts. In contrast, the same notion of in the UD guidelines is grounded on the distinction between core and oblique arguments (Thompson, 1997), and it is annotated by using three separate dependency relations, respectively for (mostly direct) objects acting as the second most core argument of a verb after the subject (`obj`), indirect objects (`iobj`), and oblique nominals (`obl`). According to the valency-based notion, verbs that require three or more arguments to fill their valency slots will be likely to have more than one dependent tagged as `Obj` in the original format; in the passage to UD, all but one of these 'objects' will have to be converted to other relations, a situation that can possibly lead to conversion errors.

LiLa includes a manually compiled valency lexicon for Latin, called Latin Vallex 2.0, that lists the valency frames associated to each sense of any given valency-capable word (Mambrini et al., 2021). The entries in Latin Vallex are linked to the same lemmas in LiLa as the ITTB tokens, so that interoperability between the two resources is ensured.

Using LiLa's SPARQL endpoint,[19] it is possible to extract the set of relations in both schemes where the head is represented by one of the verbs that are licensed to require three or more arguments in Latin Vallex. In this way, LiLa can be effectively used to review the syntactic annotation of verbs based on their argument structure recorded in the valency lexicon.

Figure 4 exemplifies the case showing the double set of annotation of the phrase: *quam Deus indidit creaturis* '(wisdom) that God bestowed on the creatures' (SCG 4.8.10). All the senses of the verb *indo* 'impart, attach to' registered in Latin Vallex 2.0 require three argument slots (left part of the picture). In the original annotation, both the Theme and the Addressee are annotated with the tag 'object' (`Obj`; purple nodes on the right side). Instead, in UD only the Theme (represented by the relative pronoun *quam*) is annotated as `obj`; the Addressee (*creaturis* 'the creatures') is converted to 'oblique', with the subtype (`obl-arg`) reserved to valency arguments.[20]

6. Conclusions and Future Work

In this paper we have presented how the annotation recorded in the ITTB has been made available as LLOD and integrated within the LiLa Knowledge Base of interoperable linguistic resources for Latin.

Our model allows to publish multiple layers of annotations, and even concurring annotations of the same type insisting on the same token, as with the original and converted versions of the ITTB. In this way, we ensure that multiple interpretations given of the same (edition of a) text can be simultaneously accessed.

The solution discussed here presupposes the existence of the appropriate OLiA annotation models for the recorded annotation. A still greater level of interoperability might be reached if *linking* models, anchoring the classes and individuals of the annotation ontology to a common vocabulary, were created, so that the conceptual relations between the terminology used in each project would be made transparent. This is however a very complex goal, given the difficulties in harmonizing notions that deceptively share the same name. The notion of 'object', which, as we saw, is sensibly different in the PDT and UD annotation, is a convenient example of the many pitfalls within this process.

The ITTB is but one of the available treebanks of Latin. Our next step is to extend the coverage of the available syntactically annotated corpora in LiLa. Then, the next target for publication as LLOD and inclusion within the LiLa network of resources will be the UDante treebank.

7. 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.

8. Bibliographical References

Bamman, D. and Crane, G. (2011). The ancient greek and latin dependency treebanks. In *Language Technology for Cultural Heritage*.

Bohmová, A., Hajič, J., Hajičová, E., and Hladká, B. (2001). The Prague Dependency Treebank: A Three-Level Annotation Scenario. In Anne Abeillé, editor, *Treebanks: Building and Using Syntactically Annotated Corpora*, pages 103–127. Kluwer Academic Publishers, Boston.

Buitelaar, P., Cimiano, P., McCrae, J., Montiel-Ponsoda, E., and Declerck, T. (2011). Ontology Lexicalization: The *lemon* Perspective. In *Proceedings of the Workshops-9th International Conference on Terminology and Artificial Intelligence (TIA 2011)*, pages 33–36.

Cecchini, F. M., Passarotti, M., Marongiu, P., and Zeman, D. (2018). Challenges in Converting the Index Thomisticus Treebank into Universal Dependencies. In *Proceedings of the Second Workshop on Universal Dependencies (UDW 2018)*, pages 27–36, Brussels, Belgium, November. Association for Computational Linguistics.

Cecchini, F. M., Sprugnoli, R., Moretti, G., and Passarotti, M. (2020a). UDante: First Steps Towards the Universal Dependencies Treebank of Dante's

[18] `https://ufal.mff.cuni.cz/pdt2.0/doc/manuals/en/a-layer/html/ch03s02x04.html#objvymez`.

[19] `https://lila-erc.eu/sparql/`.

[20] `https://universaldependencies.org/u/dep/obl-arg.html`.

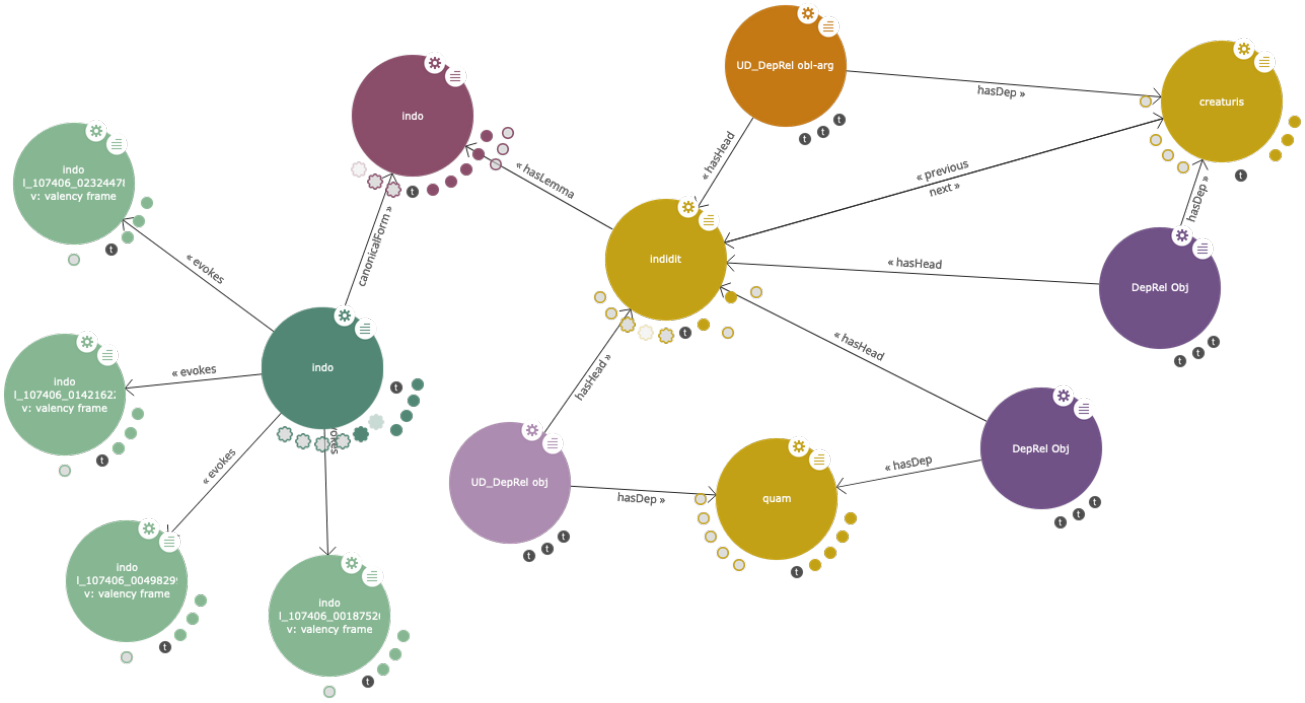

Figure 4: Syntactic dependents of a 3-argument verb in the ITTB, original and UD version.

Latin Works. In *Seventh Italian Conference on Computational Linguistics*, pages 1–7, Bologna. CEUR-WS.org.

Cecchini, F. M., Korkiakangas, T., and Passarotti, M. (2020b). A new latin treebank for universal dependencies: Charters between ancient latin and romance languages. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 933–942.

Chiarcos, C. and Fäth, C. (2017). CoNLL-RDF: Linked corpora done in an NLP-friendly way. In *International Conference on Language, Data and Knowledge*, pages 74–88. Springer.

Chiarcos, C. and Sukhareva, M. (2015a). OLiA – Ontologies of Linguistic Annotation. *Semantic Web*, 6(4):379–386.

Chiarcos, C. and Sukhareva, M. (2015b). Olia – ontologies of linguistic annotation. *Semantic Web*, 6(4):379–386.

Chiarcos, C., Hellmann, S., and Nordhoff, S. (2012). Linking Linguistic Resources: Examples from the Open Linguistics Working Group. In Christian Chiarcos, et al., editors, *Linked Data in Linguistics: Representing and Connecting Language Data and Language Metadata*, pages 201–216. Springer, Berlin, Heidelberg.

Chiarcos, C., Fäth, C., and Abromeit, F. (2020). Annotation Interoperability for the Post-ISOCat Era. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 5668–5677, Marseille, France, May. European Language Resources Association.

Chiarcos, C., Ionov, M., Glaser, L., and Fäth, C. (2021). An Ontology for CoNLL-RDF: Formal Data Structures for TSV Formats in Language Technology. In Dagmar Gromann, et al., editors, *Proceedings of the 3rd Conference on Language, Data and Knowledge (LDK 2021)*, pages 20:1–20:14, Dagstuhl. Schloss Dagstuhl – Leibniz-Zentrum für Informatik.

Chiarcos, C. (2012). POWLA: Modeling Linguistic Corpora in OWL/DL. In Elena Simperl, et al., editors, *The Semantic Web: Research and Applications*, Lecture Notes in Computer Science, pages 225–239, Berlin, Heidelberg. Springer.

Cimiano, P., Chiarcos, C., McCrae, J. P., and Gracia, J. (2020). *Linguistic Linked Data: Representation, Generation and Applications*. Springer, Cham.

Haug, D. T. and Jøhndal, M. (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.

Hellmann, S., Lehmann, J., Auer, S., and Brümmer, M. (2013). Integrating NLP Using Linked Data. In Harith Alani, et al., editors, *The Semantic Web – ISWC 2013*, pages 98–113, Berlin, Heidelberg. Springer Berlin Heidelberg.

Korkiakangas, T. (2017). Spelling variation in historical text corpora: The case of early medieval documentary Latin. *Digital Scholarship in the Humanities*, 33(3):575–591, 11.

Lassila, O. and Swick, R. R. (1998). Resource Description Framework (RDF) Model and Syntax Specification.

Mambrini, F. and Passarotti, M. (2019). Linked Open Treebanks. Interlinking Syntactically Annotated Corpora in the LiLa Knowledge Base of Linguistic Resources for Latin. In *Proceedings of the 18th International Workshop on Treebanks and Linguistic Theories (TLT, SyntaxFest 2019)*, pages 74–81, Paris. Association for Computational Linguistics.

Mambrini, F., Passarotti, M., Litta, E., and Moretti, G. (2021). Interlinking Valency Frames and WordNet Synsets in the LiLa Knowledge Base of Linguistic Resources for Latin. In Mehwish Alam, et al., editors, *Further with Knowledge Graphs. Studies on the Semantic Web 53*, Amsterdam, August. IOS Press.

McCrae, J., Bosque-Gil, J., Gracia, J., Buitelaar, P., and Cimiano, P. (2017). The OntoLex-Lemon Model: Development and Applications. In *Proceedings of eLex*, pages 587–597.

Nivre, J., De Marneffe, M.-C., Ginter, F., Goldberg, Y., Hajic, J., Manning, C. D., McDonald, R., Petrov, S., Pyysalo, S., Silveira, N., et al. (2016). Universal dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1659–1666.

Passarotti, M., Budassi, M., Litta, E., and Ruffolo, P. (2017). The Lemlat 3.0 Package for Morphological Analysis of Latin. In *Proceedings of the NoDaLiDa 2017 Workshop on Processing Historical Language*, pages 24–31.

Passarotti, M., Mambrini, F., Franzini, G., Cecchini, F. M., Litta, E., Moretti, G., Ruffolo, P., and Sprugnoli, R. (2020). Interlinking through Lemmas. The Lexical Collection of the LiLa Knowledge Base of Linguistic Resources for Latin. *Studi e Saggi Linguistici*, 58:177–212.

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

Passos, G. P. (2018). *A formal specification for syntactic annotation and its usage in corpus development and maintenance: a case study in universal dependencies*. PhD Thesis, Universidade Federal do Rio de Janeiro, Rio de Janeiro, August. Accepted: 2020-10-02T22:02:34Z Publisher: Universidade Federal do Rio de Janeiro.

Ponti, E. M. and Passarotti, M. (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'16)*, pages 683–688, Portorož, Slovenia, May. European Language Resources Association (ELRA).

Thompson, S. A. (1997). Discourse Motivations for the Core-Oblique Distinction as a Language Universal. In Akio Kamio, editor, *Studies in Language Companion Series*, volume 36, pages 59–82. Benjamins, Amsterdam.

9. Language Resource References

Bamman, D. et al. (2017). *Perseus Latin Dependency Treebank*. The Perseus Project, Tufts University, ISLRN 578-883-113-369-6.

Cecchini, F. et al. (2021). *UD Latin UDante*. LINDAT / CLARIAH-CZ, `http://hdl.handle.net/11234/1-4611`.

Haug, D. et al. (2018). *The PROIEL Treebank*. `http://proiel.github.io/`.

Johnson, K.P. et al. (2021). *The Classical Language Toolkit (CLTK)*. `http://cltk.org/`, v. 1.0.15.

Korkiakangas, T. (2020). *Late Latin Charter Treebank 2 (LLCT2)*. Zenodo, DOI:10.5281/zenodo.3633614, v. 1.2.

Passarotti, M. et al. (2020). *LEMLAT 3.0*. CIRCSE, Università Cattolica del Sacro Cuore, and Zenodo, DOI:10.5281/zenodo.1492134, v. 3.0.

Passarotti, M. et al. (2021). *Index Thomisticus Treebank*. CIRCSE, Università Cattolica del Sacro Cuore, ISLRN 105-545-284-528-2.

Straka, M., and Straková, J. (2021). *UDPipe*. LINDAT / CLARIAH-CZ, `http://hdl.handle.net/11234/1-1702`.

Zeman, D. et al. (2021). *Universal Dependencies (UD) 2.9*. LINDAT / CLARIAH-CZ, `http://hdl.handle.net/11234/1-4611`, v. 2.9.