CorefLat. Coreference Resolution for Latin as Linked Open Data

Eleonora Delfino[1,*,†], Roberta Grazia Leotta[2,†], Francesco Mambrini[2,†], Marco Passarotti[2,†] and Giovanni Moretti[2,†]

[1]*Università degli Studi di Udine, Via Palladio 8, 33100 Udine, Italia*
[2]*Università Cattolica del Sacro Cuore, Largo Gemelli 1, 20123 Milano, Italia*

Abstract

This paper presents the publication as Linked Open Data of a set of coreference and anaphora annotations (called *CorefLat*) performed on a set of Latin texts. Annotations are made on texts already available as Linked Open Data as part of the *LiLa Knowledge Base* of interoperable linguistic resources for Latin. By adopting a lemma-centered architecture and established guidelines for annotation inspired by those of the GUM corpus, *CorefLat* systematically identifies and tags entities and mentions, creating relational links. The annotated corpus covers multiple periods and genres, including Augustine's *Confessiones*, Plautus' *Curculio*, Caesar's *De Bello Gallico*, and Seneca's *Medea*, ensuring a balanced dataset for broader linguistic analysis. The publication of *CorefLat* as Linked Open Data relies on an OWL ontology that extends the POWLA framework, thus enabling interoperability with diverse linguistic resources within *LiLa*. We detail how coreference relations, including phenomena such as anaphora, cataphora, split antecedents, and multiword units, are encoded through specialized classes and object properties.

Keywords

Latin, Linguistic Linked Open Data, Coreference and Anaphora Resolution, Linguistic Resources

1. Introduction and Related Work

Coreference Resolution and Anaphora Resolution (CR and AR) have been central to Natural Language Processing (NLP) since the 1960s, but were long considered complex tasks requiring advanced knowledge and inference tools. In 1983, Roberto Busa highlighted the lack of research work for pronoun resolution, asking if tools existed to "automatically link pronouns to their antecedents" [1].

By the 1990s, following the empirical turn that hit the NLP world in those years, research on automatic CR and AR shifted to stochastic approaches based on machine learning algorithms. Such turn was possible thanks to the development of corpora enriched with CR/AR annotations, through initiatives such as the Message Understanding Conference (MUC) [2] and the Automatic Content Evaluation (ACE) [3]. While these corpora primarily consist of English news texts, they also include Arabic and Chinese datasets. A key resource in this domain is OntoNotes, a large-scale annotated corpus spanning multiple genres[1] and languages, proposed in the CoNLL-2012 shared task [4]. The NXT-format Switchboard Corpus [5] is also enriched with coreference annotation: it consists of a dataset of informal telephone conversations in English, annotated with syntactic, prosodic, and, as said, coreference information. Several annotated corpora have extended the linguistic coverage to include German [6], Japanese [7], Italian [8], Spanish [9], and Czech [10]. Other useful resources for CR/AR studies include parallel corpora such as ParCorFull [11] and ParCorFull2.0 [12], which provide full coreference annotation across

SemDH 2025: Second International Workshop of Semantic Digital Humanities co-located with ESWC 2025, June 01-02, 2025, Portoroz, Slovenia

*Corresponding author.
†These authors contributed equally.
eleonora.delfino@uniud.it (E. Delfino); robertagrazia.leotta@unicatt.it (R. G. Leotta); francesco.mambrini@unicatt.it (F. Mambrini); marco.passarotti@unicatt.it (M. Passarotti); giovanni.moretti@unicatt.it (G. Moretti)
0009-0002-5947-5011 (E. Delfino); 0009-0004-5631-1032 (R. G. Leotta); 0000-0003-0834-7562 (F. Mambrini); 0000-0002-9806-7187 (M. Passarotti); 0000-0001-7188-8172 (G. Moretti)

[1]Although it encompass only practical and informational texts.

multiple languages and are particularly valuable for studying cross-linguistic coreference phenomena and improving machine translation.

While most resources focus exclusively on practical or informational texts, some are specifically dedicated to the study of literary texts, such as DramaCoref, a neural network system for CR on German theater plays [13] and LitBank [14], a dataset of coreference annotations for literary English texts. Moreover, it is worth mentioning that in the Universal Dependencies (UD) framework, Enhanced Dependencies (ED)[2] extend the basic syntactic representations by incorporating additional relational information, such as controlled arguments, propagated conjunct dependencies, and referential links. This enriched annotation partially supports coreference and anaphora resolution by explicitly encoding certain grammatical relations that contribute to reference tracking. ED annotations are available for multiple languages and are primarily applied to texts from treebanks covering a range of genres, including news articles, legal texts, and web data, depending on the specific UD corpus. Nevertheless, their ability to fully resolve coreference remains limited compared to dedicated coreference-annotated corpora. ED primarily provide advanced information that can support coreference resolution but do not systematically or exhaustively resolve it.

For Classical languages, fundamental resources are the Ancient Greek and Latin Dependency Treebank (AGLDT),[3] which includes excerpts from Ancient Greek and Latin texts of the Classical era, and the *Index Thomisticus* Treebank (IT-TB),[4] which features Medieval Latin texts of Thomas Aquinas.

The syntactic annotation of both corpora was originally based on a scheme resembling that of the analytical layer of the Prague Dependency Treebank (PDT) [15]. Both the treebanks feature a small subset of data annotated at the so-called tectogrammatical layer of the PDT [16, 17, 18]. This annotation layer captures the underlying syntactic structure of sentences (while the analytical layer represents surface syntax). Through tectogrammatical annotation, the treebank is enhanced with a range of annotation tasks, including semantic role labeling, information structure, and ellipsis and coreference resolution.

As for Latin, which is the focus of this paper, approximately 45,000 tokens out of the AGLDT and the IT-TB are available enhanced with tectogrammatical annotation, covering excerpts from Sallust, Caesar, Cicero (AGLDT), and Thomas Aquinas (IT-TB). Despite its significance, this set of CR/AR annotations for Latin remains unbalanced, with more than half of its tokens originating from Aquinas' *Summa contra Gentiles* (approx. 27,000 words) and Sallust's *In Catilinam* (approx. 10,936 words) [19]. To mitigate this imbalance, we developed *CorefLat*, a more diverse and balanced set of CR/AR annotations for Latin, which will ultimately incorporate a broader selection of Classical and Late Latin texts.

Since the CR/AR annotations provided by *CorefLat* are intended for publication as Linked Open Data (LOD), they were performed on a corpus of Latin texts that is already available as LOD. This corpus is part of the *LiLa Knowledge Base* of interoperable linguistic resources for Latin published as LOD.[5] Annotating directly within *LiLa* ensures seamless interoperability with the (meta)data of other resources published therein.

This paper outlines the process of publication as LOD of the CR/AR annotations provided by *CorefLat*. The paper is structured as follows. Section 2 offers a brief introduction to the *LiLa Knowledge Base*; Section 3 presents an overview of *CorefLat*'s annotation guidelines. Section 4 describes the ontology used to describe the data and for the publication of *CorefLat* as LOD (4.1), details a few examples that informed its development (4.2), and illustrates a case study showing the research potential of integrating *CorefLat* into *LiLa* (4.3).

[2] https://universaldependencies.org/u/overview/enhanced-syntax.html.
[3] https://perseusdl.github.io/treebank_data/.
[4] http://itreebank.marginalia.it.
[5] https://lila-erc.eu.

2. The LiLa Knowledge Base

The *LiLa - Linking Latin* project [20] was awarded an ERC Consolidator Grant (2018-2023) to integrate existing linguistic resources for Latin in a Knowledge Base to ensure their online interoperability.

The *LiLa Knowledge Base* (KB) was developed upon established standards for the publication of data in the Semantic Web, fitting the principles of the so-called Linked Data paradigm [21]. Accordingly, each data point in the linguistic resources interlinked in the KB is assigned a unique and persistent URI (Uniform Resource Identifier) published on the Web as URL using the HTTP protocol, to ensure its findability and accessibility. By employing web standards such as the RDF (Resource Description Framework) data model [22] and the SPARQL query language,[6] *LiLa* facilitates the creation of links between distinct URIs and the reuse of data.

The *LiLa KB* leverages a few existing ontologies to represent the (meta)data of the Latin resources interlinked therein. Key ontologies integrated into the KB include POWLA for corpus data (Portable Linguistic Annotation with OWL, an ontology designed to express any textual data and metadata as LOD) [23], OLiA for linguistic annotation (Ontologies of Linguistic Annotation, a set of ontologies that allow to express and map linguistic categories) [24], and Ontolex-Lemon for lexical data [25].

In the architecture of the *LiLa KB*, lemmas play the core role, as the pivotal connection points among both lexical and textual resources. Such a highly lexically-based architecture is based upon a simple yet efficient assumption: textual resources consist of word occurrences (tokens) and lexical resources describe word properties in lexical entries. Following the lemma's foundational role in *LiLa*, the KB is based on the so-called *Lemma Bank*, a collection of approximately 200,000 Latin lemmas (canonical citation forms of lexical items) published as LOD. This collection originates from the lexical base of the LEMLAT 3.0 morphological analyzer for Latin [26] and it is constantly extended as far as new resources are added to the KB. Interoperability is achieved by linking all lexical entries and corpus tokens to their corresponding lemma in the *Lemma Bank*, thus enabling seamless integration across resources.

3. Building *CorefLat*. Guidelines and Annotation

This section provides an overview of the *CorefLat* data set, with a particular focus on the guidelines that informed the annotation process.[7]

The annotation process was performed following the guidelines of the GUM corpus,[8] which are also employed in the Universal Anaphora (UA) project,[9] aiming for consistency across linguistic resources enhanced with CR/AR annotation. In coreference annotation, a distinction is made between Entities, which are referred, and Mentions, which refer back or forward to an Entity. *CorefLat*'s approach emphasizes relationships rather than chains of Entities and Mentions. The difference between coreference relations and coreference chains lies in their scope and structure:

- coreference relations refer to the specific linguistic connection between two or more expressions that refer to the same entity in a discourse. For example, in the sentence "Maria loves her cat. She takes good care of it", the pronoun *she* is in a coreference relation with *Maria*, and *it* refers to *her cat*.
- coreference chains are sequences of multiple referring expressions that all refer to the same entity throughout a text. A chain is made up of multiple coreference relations. For example, in "Maria loves her cat. She takes good care of it. The feline enjoys playing with her", the chain consists of (*Maria* → *she* → *her*), and (*her cat* → *it* → *the feline*).

[6] https://www.w3.org/TR/rdf-sparql-query/. *LiLa*'s SPARQL endpoint can be accessed at https://lila-erc.eu/sparql/.

[7] Annotation was performed manually using the customizable Content Annotation Tool (CAT). (Meta)data were first saved in XML and then converted automatically into the CoNLL-U Plus format (https://universaldependencies.org/ext-format.html), following the recommendations provided by the UA project [27].

[8] https://wiki.gucorpling.org/gum/entities.

[9] https://universalanaphora.github.io/UniversalAnaphora/.

In *CorefLat*, we annotate coreferences as relations and we select a limited set of fundamental types of coreference. Such a limited set aligns with our objective of building a foundational Latin corpus enhanced with coreference annotations. In examples (1) to (4) we provide a detailed review of the types of annotations available in *CorefLat*.

(1) **Anaphora**: a mention referring back to an entity. This type of coreference constitutes the most frequently represented category in our corpus, accounting for 1,222 instances out of the 1,520 annotated coreferences within the texts. [10]

a. ***domine qui*** *et semper vivis.*
'**Lord** (you) **who** live for ever'.[11]
(Aug. *Conf.* 1.6.8)

b. ***Laudes*** *tuae, domine,* ***laudes*** *tuae per scripturas tuas suspenderent palmitem cordis mei.*
'Your **praises**, Lord, your **praises** throughout your Scriptures would have supported the vine shoot of my heart'.
(Aug. *Conf.* 1.17.27)

c. *Quo usque, quaeso, ad hunc modum / inter* ***nos*** *amore utemur semper surrupticio?*
'How much longer, please, will **we** always conduct our love affair in secret?'[12]
(Pl. *Curc.* 1, 204-205)

Pronominal anaphora are the prototypical case of anaphora, where the mention is represented by a pronoun, like the relative pronoun *qui* in (1a). Beside this type of anaphora, we annotate one type of anaphoric relation involving two (identical) content words, where the latter refers back to the former, as in (1b). Thus, the first utterance of the content word functions as the entity, while the second serves as the mention.

During the annotation process we got through cases of coreference relation where the entity is implicit, leading to the mention being anaphorically linked to an external entity (cf. 4.1). The personal pronoun *nos* in (1c) refers to *Planesium* and *Phaedromus*, two of the main characters from Plautus' comedy *Curculio*. These characters, however, are not mentioned in close textual proximity to the pronoun. This is an instance of long-distance coreference, a phenomenon that presents a challenge in CR/AR, as there is no strict upper limit on the number of sentences after which a mention can no longer be linked to its entity [30]. However, modern NLP methods have proven highly effective in addressing this issue, successfully linking mentions to their entities across spans exceeding 200 sentences [31]. Literary texts, the primary focus of *CorefLat*, seem to exhibit long-distance coreference more frequently than other textual genres [32], thus making it crucial to devote particular attention to this phenomenon. To ensure consistency in annotation, we set a threshold: when a mention exceeds five sentences from its entity, it is connected to an external entity. This threshold is sentence-based rather than token-based, aligning with the standard practice in CR/AR studies, where sentences serve as the primary unit of analysis.

(2) **Cataphora**: a mention referring forward to an entity. This type of coreference is the second most frequent in our corpus, with 177 occurrences, out of the total of 1,520 coreferential relations annotated in the texts.

a. *invocat* ***te***, ***domine****.*
'invokes **you**, **Lord**'
(Aug. *Conf.* 1.1.1)

(3) **Split antecedents**: the mention has multiple antecedents, so one mention refers back or forward to more than one entity. In *CorefLat*, this linguistic phenomenon is underrepresented, with only 87 occurrences of the 1,520 annotated coreferential relations in the texts. However, it might

[10] Mentions referring to external entities are excluded from these counts, see (1c) and the corresponding discussion for further details.
[11] All translations of Augustine's *Confessiones* are taken from [28].
[12] All translations of Plautus' *Curculio* are taken from [29].

manifest in two structural patterns: as mentions referring to either conjoint (3a) or disjoint noun phrases (3b) and as mentions referring to previously listed nouns (3c).

a. *An vero **caelum** et **terra**, **quae** fecisti et in **quibus** me fecisti, capiunt te?*
'**Heaven** and **earth**, **which** you made, and in **which** you made me, encompass you?'
(Aug. *Conf.* 1.2.2)

b. *Nec **mater** mea vel **nutrices** meae **sibi** ubera implebant, sed tu mihi per **eas** dabas alimentum infantiae.*
'Neither my **mother** nor my **nurses** filled (their) breasts for **themselves**, but you gave the nourishment of infancy to me through **them**.'
(Aug. *Conf.* 1.6.7)

c. *Gallia est omnis divisa in partes tres, quarum unam incolunt **Belgae**, aliam **Aquitani**, tertiam qui ipsorum lingua Celtae, nostra **Galli** appellantur. **Hi** omnes lingua, institutis, legibus inter se differunt.*
'Gaul is a whole divided into three parts, one of which is inhabited by the **Belgae**, another by the **Aquitani**, and a third by a people called in their own tongue Celtae, in the Latin **Galli**. All **these** are different one from another in language, institutions, and laws.'[13]
(Caes., *Gal.*, 1.1.1)

(4) **Multiword antecedents**: the entity involved in the coreference relation consists of more than one token. In *CorefLat*, this type of coreference is the least represented, accounting for only 34 occurrences of the 1,520 coreferential relations annotated in the texts. However, it primarily occurs in two contexts: when the entity is a proper noun following the Roman onomastic system (*tria nomina*), as illustrated in (4a), or when the entity is linguistically realized as a noun phrase consisting of a noun and a modifier or specifier that semantically restricts the reference of a noun by specifying a subset of possible referents (4b).

a. *Itaque prius quam quicquam conaretur, Diviciacum ad se vocari iubet et, cotidianis interpretibus remotis, per **C. Valerium Troucillum**, principem Galliae provinciae, familiarem suum, **cui** summam omnium rerum fidem habebat, cum eo conloquitur*;
'Therefore, before attempting anything in the matter, Caesar ordered Diviciacus to be summoned to his quarters, and, having removed the regular interpreters, conversed with him through the mouth of **Gaius Valerius Trocillus**, a leading man in the Province of Gaul and his own intimate friend, in **whom** he had the utmost confidence upon all matters.'
(Caes., *Gal.*, 1.19.3)

b. *Comprecor vulgus silentum **vos**que **ferales deos** et Chaos caecum atque opacam Ditis umbrosi domum.*
'I invoke the thronging silent dead, and **you the gods of the grave**, and sightless Chaos, and the shadowy home of dark-enshrouded Dis.'[14]
(Sen. *Med.*, 740-741)

In (4b), *ferales deos* functions as multiword antecedent of the pronoun *vos*. In this instance, the semantics of the noun phrase is constrained by a modifier. In addition to illustrating a case of a multiword antecedent, (4b) also demonstrates a cataphoric relationship between an entity and a mention, as the pronoun *vos* precedes the noun phrase.

To ensure the applicability of the annotation guidelines explained above across different linguistic and stylistic contexts, a diversified set of texts was selected, comprising works from various genres and periods to provide a balanced representation of Latin traditions.[15] So far, *CorefLat* includes a passage from a Late Antique philosophical work (the first book of Augustine's *Confessiones*), an archaic comedy

[13] All translations of Caesar's *De Bello Gallico* are taken from [33].
[14] Seneca *Medea*'s translations are taken from [34].
[15] The decision to diversify the corpus by including texts from different genres is also driven by the interest in analyzing how the coreference phenomenon occurs and manifests across various textual domains. For initial observations on this topic, see Delfino et al. [35].

(Plautus' *Curculio*), an excerpt from a Classical historiographical text (the first book of Ceaser's *De Bello Gallico*), and a Classical tragedy (Seneca's *Medea*), for a total of 25,965 tokens.[16]

The workload was equally distributed between two annotators. To assess inter-annotator agreement, however, both annotators annotated the final 50 sentences of the first book of Augustine's *Confessiones*. Agreement was measured using the Dice coefficient, a similarity metric widely employed in NLP ([36], [37]), which ranges from 0 (indicating no overlap) to 1 (indicating identical sets). After confirming that the annotated markables spanned the same tokens for both annotators in all cases, we computed the similarity scores for entities (0.817) and mentions (0.824), both of which are comparatively high and acceptable for this task ([38], [39], [40]).

4. Publishing CorefLat in LiLa

4.1. Modeling

This section explains how we modeled the information in the coreference annotation. The adopted solutions aim to link the annotated text to the *LiLa KB*, and to ensure the semantic interoperability of the coreference annotation with the other Linguistic Linked Open Data in *LiLa*.

The *LiLa CorefLat* Ontology is an OWL ontology that extends the POWLA framework [23] shared also by the other annotated corpora in *LiLa* [41].[17] At the highest level of abstraction, the *CorefLat* Ontology introduces a class called `Coreference Element`,[18] which serves as the foundational set for all entities and relations involved in coreference annotation. The class `Coreference Element` includes `Entity`, `Mention`, `Coreference Unit` and `Coreference Relation` as subclasses.

The POWLA ontology defines four primitive concepts to describe corpora: documents, layers, nodes and relations. While the former two are used in the *LiLa* corpus ontology to structure the texts and model the structural metadata [41],[19] nodes and relations are particularly relevant to model the information annotated by *CorefLat*. As said in Section 3, we define coreference as a relation between tokens that play the role of entities and mentions. As relations in POWLA are described as labeled, directed edge, the class is very suitable to express this notion.[20] Moreover, POWLA adopts a reified approach, whereby all relations are instantiated as RDF resources, provided with their own URI. This modeling strategy allows users of the *CorefLat* Ontology to make statements such as attributions to annotators, or degree of certitude about any coreference link, if one so wishes.

In POWLA, every unit of linguistic analysis is defined as an instance of the class Node. In our framework, coreference units often span across multiple tokens, as in (4a) and (4b), and are therefore best conceptualized as phrases encompassing one or more tokens in the text. The class of `powla:Nonterminal`, the subclass of Node used for phrases or chunks disjoint from the class of `Terminal` (used for actual base segments of texts), is once again very well suited to express the concept. The classes of `Coreference Relation` and `Coreference Units` are thus defined as subclasses of both *CorefLat*'s top concept `Coreference Element` and of the POWLA classes Node and Relation.

The classes of `Entity` and `Mention`, on the other hand, do not align with concepts in POWLA, and were abstracted from the guidelines described in Section 3. In our ontology, they are defined as disjoint subclasses of `Coreference Element` only that allow annotators to further specify the role of each coreference unit in the coreference relation, whether they serve as referred (entity) or the referring element (mention).

The object properties of the *CorefLat* ontology serve the purpose of expressing the relations between the corpus tokens and the coreference units, as well as the role of the units with the reified coreference

[16] The Classical Latin data originate from the *Opera Latina* corpus by LASLA, which contains over 1.7 million words from both Classical and Late Latin texts (https://lasladb.uliege.be/OperaLatina/), while Late Latin examples are sourced from *The Latin Library* http://www.m.thelatinlibrary.com/.

[17] The ontology is available at https://lila-erc.eu/lodview/ontologies/lila_coref/.

[18] http://lila-erc.eu/ontologies/lila_coref/CoreferenceElement.

[19] http://lila-erc.eu/ontologies/lila_corpora/.

[20] http://purl.org/powla/powla.owl#Relation.

relation. The property `hasCoreferenceTerminal`,[21] in particular, connects the coreference unit with each of the corpus tokens that make up the phrase.[22] The properties `hasCoreferenceSource` and `hasCoreferenceTarget`,[23] subproperties of `powla:hasSource` and `powla:hasTarget`, link the reified coreference relation to, respectively, the source and the target of the directed edge. Moreover, the ontology defines two additional properties `hasMention` and `hasEntity` as subproperties of, respectively, `hasCoreferenceSource` and `hasCoreferenceTarget`. These two properties define, respectively, the classes `Mention` and `Entity` as their range and presuppose and enforce a stricter interpretation of a coreference relation as a directed edge going from a mention to an entity. Users of the ontology are free either to adhere to the stricter interpretation or to adopt a looser model of relation that only involves coreference units.

Finally, in order to facilitate the harmonization and the recognition of the various entity-type coreference units, while also enabling cross-document coreference tasks in the future, we decided to introduce an extra-textual node linked to the entities via the `itsrdf:taIdentRef` property of the Internationalization Tag Set (ITS) Ontology.[24] This node serves as an aggregator for all the entities present within the text, and provides a connector between the textual element and the extralinguistic entities and concepts referred to in texts.

Indeed, in addition to simplifying queries within the document by grouping all entity-type Coreference Units under distinct nodes, this extra-textual node also enables the expansion and interoperability of the annotated resource with other knowledge sources. The extra-textual entity nodes created by the project form a separate knowledge base of entities referred to in the annotated texts. These entities can be mapped to other encyclopedic resources, like DBpedia,[25] or Wikidata,[26] via e.g. the mapping properties of the Simple Knowledge Organization System (SKOS), such as `skos:exactMatch`. In this way, it becomes possible to connect the various "supernodes" to external gazetteers, and to enrich the nodes with properties that are transitively inherited from these external resources.

4.2. Examples LODified

This Section illustrates how the modeling of the data by *CorefLat* applies, by detailing the representation of the examples (1)-(4) presented above.[27] To begin with, in Example (1a), we annotated the token *qui* as an anaphoric reference to the token *domine*. Figure 1 visualizes how this relation is modeled using the classes and properties defined in our ontology. The visualizations are generated using an instance of the web application LodLive running on a server of the *LiLa* project.[28]

The token *domine* (yellow in Figure 1) is the object of the property `hasCoreferenceTerminal`, whose subject is the `CoreferenceUnitEntity` for *domine* (orange in Figure 1). The same applies to the token *qui* (yellow as token *domine*[29]), which is the object of the property `hasCoreferenceTerminal`, whose subject is the `CoreferenceUnitMention` for *qui* (lilac in Figure 1[30]). Those coreference units are related through the reification of their relationship, which is represented by a node. This node is of type `CoreferenceRelation` (burgundy in Figure 1) and serves as the subject of two properties: (*i*) `hasCoreferenceSource`, which has as its object the `CoreferenceUnitMention` for *qui*, and

[21] http://lila-erc.eu/ontologies/lila_coref/hasCoreferenceTerminal.

[22] Note that the linear order of the tokens in the textual resources published in *LiLa* is captured thanks to POWLA's symmetric relations `next` and `previous` that connect the text nodes in a chain. The sequence of the tokens within the coreference units can thus be expressed using these two properties of POWLA.

[23] http://lila-erc.eu/ontologies/lila_coref/hasCoreferenceSource; http://lila-erc.eu/ontologies/lila_coref/hasCoreferenceTarget.

[24] https://github.com/w3c/itsrdf.

[25] https://www.dbpedia.org/.

[26] https://www.wikidata.org/.

[27] Examples (1b) and 1 will not be explained in detail, as they follow the same modeling as Example (1a). Likewise, Examples (3b) and (3c) mirror (3a), while Example (4b) aligns with (4a).

[28] https://lila-erc.eu/lodlive/.

[29] In LODLive every class or property is represented through the same colour, although the colours are not pre-established, so they might change from a visualization to another.

[30] It has to be noted that this node has a different colour from the `CoreferenceUnit` for *domine*, because *qui* is also an instance of `Mention`, while *domine* is an instance of `Entity`.

Figure 1: LODLive view of the anaphoric relation *qui* - *domine* shown in (1a).

(*ii*) `hasCoreferenceTarget`, which has as its object the `CoreferenceUnitEntity` for *domine*. Finally, both tokens are linked to their corresponding lemma in the Lemma Bank via the property `lila:hasLemma` (both in purple in Figure 1).[31]

Figure 2: LODLive view of the anaphoric relations *quae* - *caelum* and *quae* - *terra*: the split antecedents case shown in (3a).

Example (3a) illustrates a case of split antecedent, in which a single mention (*quae*) refers to two distinct entities (*caelum* and *terra*). The challenge posed by this structure lies in the fact that the same mention establishes a coreference relationship with two different entities. The approach used in our ontology to model such cases is depicted in Figure 2.

In Figure 2, the token *quae* (yellow node) is the object of the property `hasCoreferenceTerminal`, whose subject is the `CoreferenceUnitMention` for *quae* (lilac in Figure 2). This coreference unit

[31] http://lila-erc.eu/ontologies/lila/hasLemma.

is, in turn, the object of the property `hasCoreferenceSource` for two distinct coreference relations: `CorefRelation` *quae* → *caelum* and `CorefRelation` *quae* → *terra*. Since these coreference relations belong to the same class, they share the same color in Figure 2 (burgundy). The first coreference relation is the object of the property `hasCoreferenceTarget`, whose subject is the `CoreferenceUnitEntity` for *caelum* (orange in Figure 2). Similarly, the second coreference relation is the object of the property `hasCoreferenceTarget`, whose subject is the `CoreferenceUnitEntity` for *terra* (also orange in Figure 2). Both coreference units serve as subjects of the property `hasCoreferenceTerminal`, with their respective objects being the tokens *caelum* and *terra* (both yellow in Figure 2).

Example (4a) (*per Caium Valerium Troucillum* [...] *cui*...), instead, highlights the challenge of representing an entity consisting of multiple tokens within our ontological framework. This example demonstrates the importance of an abstract conceptualization of coreference relations based on coreference units rather than tokens, given that *Caius Valerius Troucillus* represents the typical three-element structure of the Roman onomastic system. If the coreference relation had been established between tokens, it would not have been possible to distinguish a 'multiword' case such as this one in (4a) from a split antecedent case like the one observed in (3a). This example is modeled in the same way as the one in (1a), with the only substantial difference being that the `CoreferenceUnitEntity` is the subject of the property `hasCoreferenceTerminal` three times: the object of the first one is the token *Caium*, the object of the second is the token *Valerium*, and the object of the third is the token *Troucillum*.

4.3. Use Case

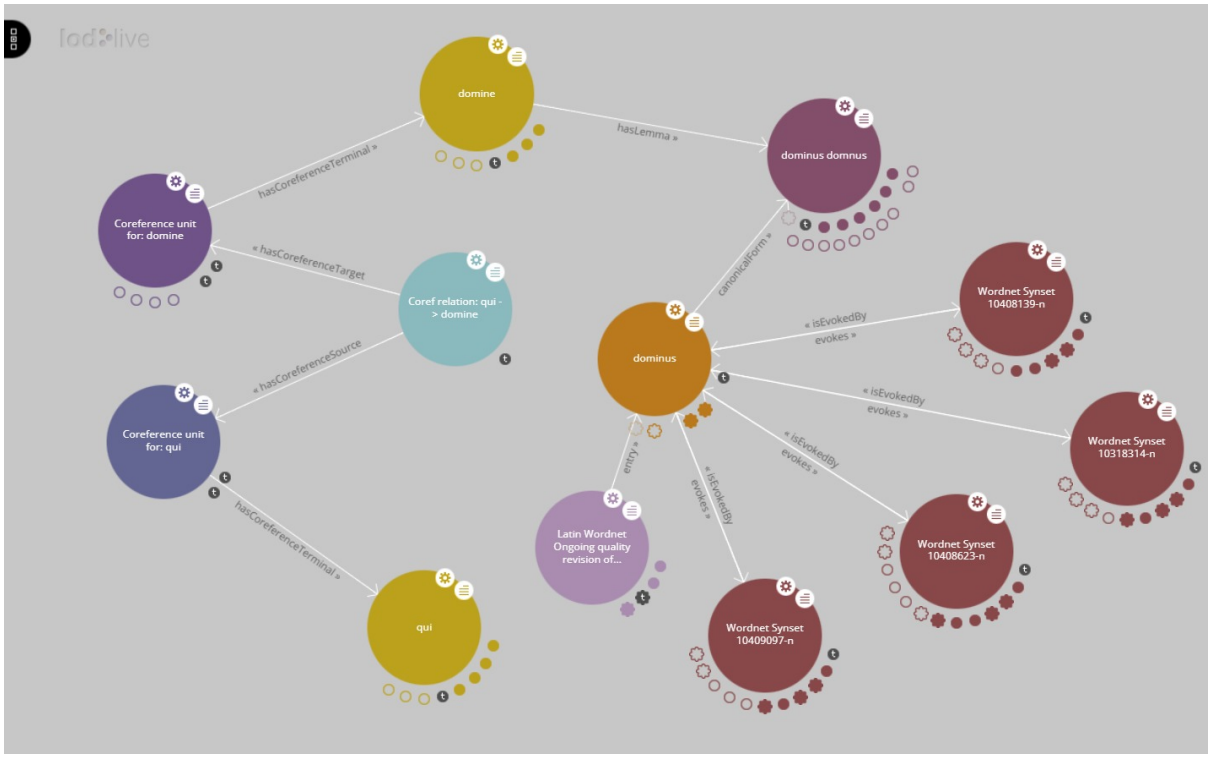

Figure 3: LodLive view of the interoperability between *CorefLat* and *WordNet* in the *LiLa Knowledge Base*.

This section examines an example of research opportunities facilitated by linking our resource *CorefLat* to the *LiLa Knowledge Base*.

Linking *CorefLat* to *LiLa* enables interoperation between *CorefLat* and all the resources already integrated therein. For instance, it is particularly interesting to observe how *CorefLat* can interact with lexical resources available in *LiLa*, such as the *Latin WordNet*.[32]

[32] http://lila-erc.eu/data/lexicalResources/LatinWordNet/Lexicon

Figure 3 illustrates an example of how a token, *domine*, linked to the lemma *dominus*, from Augustine's *Confessiones*, is the terminal belonging to a Coreference Unit of type Entity, which is the target of a Coreference Relation. Following the lemma-centred architecture of *LiLa*, lexical entries in the *Latin WordNet* are linked to the corresponding lemma in the *LiLa Lemma Bank* (as the canonical form of citation for the entry). As such, the lexical entry for *dominus* evokes a set of synsets.

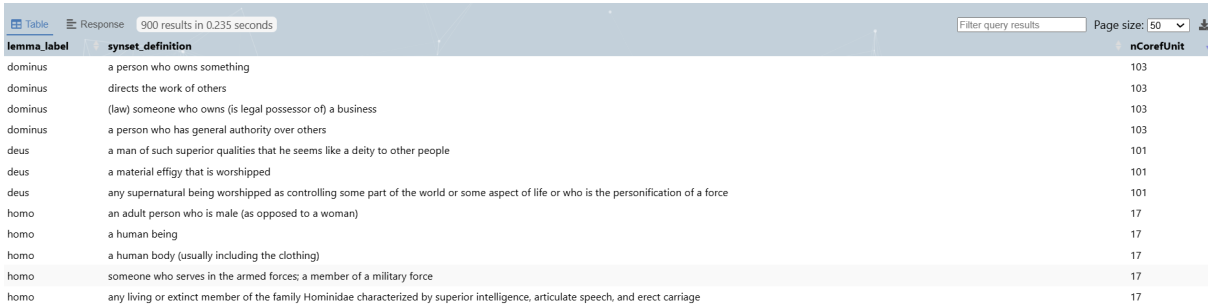

Table | Response | 900 results in 0.235 seconds | Filter query results | Page size: 50

lemma_label	synset_definition	nCorefUnit
dominus	a person who owns something	103
dominus	directs the work of others	103
dominus	(law) someone who owns (is legal possessor of) a business	103
dominus	a person who has general authority over others	103
deus	a man of such superior qualities that he seems like a deity to other people	101
deus	a material effigy that is worshipped	101
deus	any supernatural being worshipped as controlling some part of the world or some aspect of life or who is the personification of a force	101
homo	an adult person who is male (as opposed to a woman)	17
homo	a human being	17
homo	a human body (usually including the clothing)	17
homo	someone who serves in the armed forces; a member of a military force	17
homo	any living or extinct member of the family Hominidae characterized by superior intelligence, articulate speech, and erect carriage	17

Figure 4: The output of a SPARQL query joining *CorefLat* and the Latin WordNet in LiLa

Generalizing from this example, it is possible to formulate a SPARQL query to retrieve all synsets evoked by lexical entries associated with lemmas linked to tokens involved in a coreference relation. This query allows for the extraction of lemmas and their corresponding synsets in a two-column format, while also providing, for each lemma, the number of tokens involved in a coreference relation, as shown in Figure 4. See the appendix for a visualization of the SPARQL query output and the corresponding code listing.

5. Conclusion and Future Work

This work has introduced *CorefLat*, a new resource designed to support coreference and anaphora analysis in Latin and to ensure smooth data integration within the *LiLa Knowledge Base*.

By leveraging existing standards for Linked Open Data, *CorefLat* enables wider interoperability and fosters cross-resource research. Future developments will center on (a) increasing the scope of the annotated corpus to include a broader range of textual genres and historical periods, and (b) exploiting the expanded dataset to train automatic CR/AR models for Latin, evaluating their performance both on in-domain and out-of-domain material.

To conclude, it would be advisable to annotate with coreference texts that have already been enriched with syntactic annotation in accordance with the Universal Dependencies guidelines. Specifically, regarding Classical Latin, we will make use of the *UD Latin-Circse*,[33] a treebank repository currently under development by the CIRCSE Research Centre in Milan. The repository contains both prose and poetry texts from different periods. At present, it includes three texts taken from the *Opera Latin* corpus by LASLA, namely Seneca's *Hercules Furens* and *Agamemnon*, and Tacitus' *Germania*.

Acknowledgments

This contribution is funded by the PRIN-2022 project "Textual Data and Tools for Coreference Resolution in Latin" (CUP J53D23013680008), a project carried out jointly by the Università Cattolica del Sacro Cuore in Milan and by the University of Udine.

[33] https://github.com/UniversalDependencies/UD_Latin-CIRCSE

Declaration on Generative AI

During the preparation of this work, the authors used X-GPT-4 for grammar and spelling check. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] J. Nyhan, M. Passarotti, One Origin of Digital Humanities: Fr Roberto Busa in His Own Words, Springer, 2019.

[2] N. A. Chinchor, Overview of MUC-7, in: Seventh Message Understanding Conference (MUC-7): Proceedings of a Conference Held in Fairfax, Virginia, April 29 - May 1, 1998, 1998. URL: https://aclanthology.org/M98-1001.

[3] G. Doddington, A. Mitchell, M. Przybocki, L. Ramshaw, S. Strassel, R. Weischedel, The automatic content extraction (ACE) program – tasks, data, and evaluation, in: M. T. Lino, M. F. Xavier, F. Ferreira, R. Costa, R. Silva (Eds.), Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04), European Language Resources Association (ELRA), Lisbon, Portugal, 2004, pp. 837–840. URL: https://aclanthology.org/L04-1011/.

[4] S. Pradhan, A. Moschitti, N. Xue, O. Uryupina, Y. Zhang, CoNLL-2012 shared task: Modeling multilingual unrestricted coreference in OntoNotes, in: S. Pradhan, A. Moschitti, N. Xue (Eds.), Joint Conference on EMNLP and CoNLL - Shared Task, Association for Computational Linguistics, Jeju Island, Korea, 2012, pp. 1–40. URL: https://aclanthology.org/W12-4501/.

[5] S. Calhoun, J. Carletta, J. M. Brenier, N. Mayo, D. Jurafsky, E. Shriberg, M. Walker, The nxt-format switchboard corpus: a rich resource for investigating the syntax, semantics, pragmatics and prosody of dialogue, Language Resources and Evaluation 44 (2010) 387–419. doi:10.1007/s10579-010-9120-1.

[6] E. Hinrichs, S. Kübler, K. Naumann, H. Telljohann, J. Trushkina, et al., Recent developments in linguistic annotations of the TüBa-D/Z treebank, Universitätsbibliothek Johann Christian Senckenberg, 2004.

[7] R. Iida, M. Komachi, K. Inui, Y. Matsumoto, Annotating a japanese text corpus with predicate-argument and coreference relations, in: Proceedings of the linguistic annotation workshop, 2007, pp. 132–139.

[8] A. Minutolo, R. Guarasci, E. Damiano, G. De Pietro, H. Fujita, M. Esposito, A multi-level methodology for the automated translation of a coreference resolution dataset: an application to the italian language, Neural Computing and Applications 34 (2022) 22493–22518.

[9] M. Recasens, M. A. Martí, Ancora-co: Coreferentially annotated corpora for spanish and catalan, Language resources and evaluation 44 (2010) 315–345.

[10] A. Nedoluzhko, M. Novák, S. Cinková, M. Mikulová, J. Mírovskỳ, Coreference in prague czech-english dependency treebank, in: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), 2016, pp. 169–176.

[11] E. Lapshinova-Koltunski, C. Hardmeier, P. Krielke, ParCorFull: A parallel corpus annotated with full coreference, 2018. URL: http://hdl.handle.net/11372/LRT-2614, LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

[12] E. Lapshinova-Koltunski, P. A. Ferreira, E. Lartaud, C. Hardmeier, ParCorFull2.0: a parallel corpus annotated with full coreference, in: N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, J. Odijk, S. Piperidis (Eds.), Proceedings of the Thirteenth Language Resources and Evaluation Conference, European Language Resources Association, Marseille, France, 2022, pp. 805–813. URL: https://aclanthology.org/2022.lrec-1.85/.

[13] J. Pagel, N. Reiter, DramaCoref: A hybrid coreference resolution system for German theater

plays, in: M. Ogrodniczuk, S. Pradhan, M. Poesio, Y. Grishina, V. Ng (Eds.), Proceedings of the Fourth Workshop on Computational Models of Reference, Anaphora and Coreference, Association for Computational Linguistics, Punta Cana, Dominican Republic, 2021, pp. 36–46. URL: https://aclanthology.org/2021.crac-1.4/. doi:10.18653/v1/2021.crac-1.4.

[14] D. Bamman, O. Lewke, A. Mansoor, An annotated dataset of coreference in English literature, in: N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis (Eds.), Proceedings of the Twelfth Language Resources and Evaluation Conference, European Language Resources Association, Marseille, France, 2020, pp. 44–54. URL: https://aclanthology.org/2020.lrec-1.6/.

[15] D. Bamman, M. C. Passarotti, R. Busa, G. Crane, The annotation guidelines of the Latin Dependency Treebank and Index Thomisticus Treebank. The treatment of some specific syntactic constructions in Latin, in: LREC 2008, ELDA, 2008, pp. 71–76.

[16] F. Mambrini, Thucydides 1.89-118: A multi-layer treebank, CHS Research Bulletin 1 (2013). URL: http://nrs.harvard.edu/urn-3:hlnc.essay:MambriniF.Thucydides_1.89-118_Multi-layer_Treebank.2013.

[17] M. Passarotti, From syntax to semantics. first steps towards tectogrammatical annotation of latin, in: Proceedings of the 8th Workshop on Language Technology for Cultural Heritage, Social Sciences, and humanities (LaTeCH), 2014, pp. 100–109.

[18] M. Passarotti, B. González Saavedra, The treebanked conspiracy. actors and actions in bellum catilinae, in: J. Hajič (Ed.), Proceedings of the 16th International Workshop on Treebanks and Linguistic Theories, Prague, Czech Republic, 2017, pp. 18–26. URL: https://aclanthology.org/W17-7605/.

[19] B. G. Saavedra, M. Passarotti, Using tectogrammatical annotation for studying actors and actions in sallust's bellum catilinae, The Prague Bulletin of Mathematical Linguistics 111 (2018) 5–28.

[20] M. Passarotti, F. Mambrini, G. Franzini, F. M. Cecchini, E. Litta, G. Moretti, P. Ruffolo, R. Sprugnoli, Interlinking through lemmas. the lexical collection of the lila knowledge base of linguistic resources for latin, Studi e Saggi Linguistici 58 (2020) 177–212.

[21] T. Berners-Lee, Www: Past, present, and future, Computer 29 (1996) 69–77.

[22] O. Lassila, Resource description framework (rdf) model and syntax specification w3c working draft 08 october 1998, http://www. w3. org/1998/10/WD-rdf-syntax-19981008 (1998).

[23] C. Chiarcos, S. Nordhoff, S. Hellmann, Linked Data in Linguistics, Springer, 2012.

[24] C. Chiarcos, M. Sukhareva, Olia–ontologies of linguistic annotation, Semantic Web 6 (2015) 379–386.

[25] J. P. McCrae, J. Bosque-Gil, J. Gracia, P. Buitelaar, P. Cimiano, The ontolex-lemon model: development and applications, in: Proceedings of eLex 2017 conference, 2017, pp. 19–21.

[26] M. Passarotti, M. Budassi, E. Litta, P. Ruffolo, The lemlat 3.0 package for morphological analysis of latin, in: Proceedings of the NoDaLiDa 2017 workshop on processing historical language, 2017, pp. 24–31.

[27] V. B. Lenzi, G. Moretti, R. Sprugnoli, Cat: the celct annotation tool., in: LREC, 2012, pp. 333–338.

[28] W. W. Augustine, Confessions, Vol. 2: Books 9-13 (Loeb Classical Library, No. 27), 1912.

[29] P. Nixon, et al., Plautus, Vol. II: Casina. The Casket Comedy. Curculio. Epidicus. The Two Menaechmuses (Loeb Classical Library), William Heinemann; GP Putnam's Sons, 1917.

[30] R. Simone, Fondamenti di linguistica, volume 9, Laterza Bari, 1990.

[31] H.-L. Trieu, A.-K. D. Nguyen, N. Nguyen, M. Miwa, H. Takamura, S. Ananiadou, Coreference resolution in full text articles with bert and syntax-based mention filtering, in: Proceedings of the 5th workshop on BioNLP open shared tasks, 2019, pp. 196–205.

[32] R. Thirukovalluru, N. Monath, K. Shridhar, M. Zaheer, M. Sachan, A. McCallum, Scaling within document coreference to long texts, Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021 (2021) 3921–3931.

[33] G. J. Caesar, The Gallic War, volume 72 of *Loeb Classical Library*, Harvard University Press, Cambridge, MA, 1917. URL: https://www.loebclassics.com/view/LCL072/1917/volume.xml.

[34] Seneca, Tragedies, Volume I: Hercules. Trojan Women. Phoenician Women. Medea. Phaedra,

volume 62 of *Loeb Classical Library*, Harvard University Press, Cambridge, MA, 2018.
[35] E. Delfino, R. G. Leotta, M. Passarotti, G. Moretti, et al., Building coreflat a linguistic resource for coreference and anaphora resolution in latin, in: CEUR WORKSHOP PROCEEDINGS, volume 3878, CEUR-WS, 2024.
[36] L. R. Dice, Measures of the amount of ecologic association between species, Ecology 26 (1945) 297–302.
[37] T. Sorensen, A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on danish commons, Biologiske skrifter 5 (1948) 1–34.
[38] K. B. Cohen, A. Lanfranchi, M. J.-y. Choi, M. Bada, W. A. Baumgartner, N. Panteleyeva, K. Verspoor, M. Palmer, L. E. Hunter, Coreference annotation and resolution in the colorado richly annotated full text (craft) corpus of biomedical journal articles, BMC bioinformatics 18 (2017) 1–14.
[39] I. Hendrickx, G. Bouma, F. Coppens, W. Daelemans, V. Hoste, G. Kloosterman, A.-M. Mineur, J. Van Der Vloet, J.-L. Verschelde, A coreference corpus and resolution system for dutch., in: LREC, 2008.
[40] A. Nedoluzhko, J. Mírovskỳ, P. Pajas, The coding scheme for annotating extended nominal coreference and bridging anaphora in the prague dependency treebank, in: Proceedings of the Third Linguistic Annotation Workshop (LAW III), 2009, pp. 108–111.
[41] F. Mambrini, M. Passarotti, G. Moretti, M. Pellegrini, The Index Thomisticus Treebank as Linked Data in the LiLa Knowledge Base, in: N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, J. Odijk, S. Piperidis (Eds.), Proceedings of the Thirteenth Language Resources and Evaluation Conference, European Language Resources Association, Marseille, France, 2022, pp. 4022–4029. URL: https://aclanthology.org/2022.lrec-1.428.

Appendix

SPARQL query to retrieve the synsets of those lexical entries of the *Latin WordNet* that are linked to lemmas in the *Lemma Bank* whose tokens are the terminals of a coreference unit entity involved in a coreference relation, together with the number of coreference units in which the tokens associated with such a lemma are involved. WordNet synsets are instances of the class `ontolex:lexicalConcept`. Endpoint: https://lila-erc.eu/sparql/.

```
PREFIX skos: <http://www.w3.org/2004/02/skos/core#>
PREFIX ontolex: <http://www.w3.org/ns/lemon/ontolex#>
PREFIX lime: <http://www.w3.org/ns/lemon/lime#>
PREFIX lila: <http://lila-erc.eu/ontologies/lila/>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX dc: <http://purl.org/dc/elements/1.1/>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX powla: <http://purl.org/powla/powla.owl#>
PREFIX lila_coref: <http://lila-erc.eu/ontologies/lila_coref/>

SELECT distinct ?lemma_label ?synset_definition (count(?coref_unit)
   as ?nCorefUnit)
WHERE {
  ?coref_relation rdf:type lila_coref:CoreferenceRelation ;
                  lila_coref:hasCoreferenceTarget ?coref_unit ;
                  rdfs:label ?coref_relation_label .
  ?coref_unit rdfs:label ?coref_unit_label ;
              lila_coref:hasCoreferenceTerminal ?token .
```

```
  ?token rdfs:label ?token_label ;
         lila:hasLemma ?lemma .
  ?lemma rdfs:label ?lemma_label .
  <http://lila-erc.eu/data/lexicalResources/LatinWordNet/Lexicon>
     lime:entry ?lex_entry_LWN .
  ?lex_entry_LWN ontolex:canonicalForm ?lemma ;
                 ontolex:evokes ?synset .
  ?synset skos:definition ?synset_definition .
}
```