CorefLat. Annotazione e modellizzazione per la risoluzione di coreferenze in latino

Eleonora Delfino - Roberta Grazia Leotta
Marco Passarotti - Francesco Mambrini

ABSTRACT

This study is part of a broader research project on coreference phenomena in the Latin language, entitled *CorefLat*. This paper focuses specifically on the publication as Linked Open Data of a set of coreference annotations performed on a selection of Latin texts. The annotations are applied to texts already available as Linked Open Data within the *LiLa Knowledge Base*, a collection of interoperable linguistic resources for Latin. *CorefLat* systematically identifies and tags entities and mentions, establishing relational links between them. The annotated corpus covers various historical periods and literary genres, including Augustine's *Confessiones*, Plautus' *Curculio*, Caesar's *De Bello Gallico*, and Seneca's *Medea*, offering a balanced dataset suitable for broad linguistic analysis. In this paper we also provide quantitative data on the annotations carried out so far, by showing some patterns and distributions of the linguistic phenomena within the dataset. Building on this, the paper

[*] Una versione precedente e non aggiornata di questo studio è stata pubblicata negli atti telematici del workshop SEM-DH, organizzato nell'ambito della Extended Semantic Web Conference (ESWC), svoltasi a Portorose (Slovenia) nel giugno 2025. La presente versione è stata invece presentata al 23° Colloquio Internazionale di Linguistica Latina, tenutosi presso l'Università di Udine nel giugno 2025. Gli autori desiderano ringraziare, in primo luogo, Giovanni Moretti, responsabile informatico del progetto, per la gestione dei dati, l'elaborazione dei risultati ottenuti e per l'indispensabile contributo in fase di modellizzazione. In secondo luogo, si ringraziano i due revisori anonimi per i preziosi consigli e i puntuali suggerimenti. Il contributo è frutto della collaborazione dei quattro autori; a fini scientifici, si precisa che Eleonora Delfino ha curato il §1 e il §3.2, Roberta G. Leotta il §2 e il §3.1, Francesco Mambrini il §4.1, mentre Marco Passarotti il §4.2 e il §4.3.

Lingue antiche e moderne 14 (2025) ISSN 2281-4841
DOI: 10.4424/lam142025-3

describes how coreference phenomena are encoded as Linked Open Data using standard classes and object properties from the POWLA framework.

1. Introduzione e stato dell'arte

La coreferenza e l'anafora costituiscono fenomeni ampiamente studiati nell'ambito della linguistica latina. Una vasta gamma di contributi scientifici ne ha approfondito lo studio da prospettive diverse. Particolare attenzione è stata rivolta agli aspetti sintattici, a quelli semantici, alla struttura informativa, alle funzioni pragmatiche, nonché alla rilevanza di questi fenomeni nell'analisi stilistica e testuale[1].

Nel presente contributo, l'indagine sulle coreferenze e sulle anafore è considerata dal punto di vista della linguistica computazionale e, nello specifico, per fini di sviluppo di risorse linguistiche. L'articolo presenta i risultati del progetto *CorefLat* che ha realizzato l'annotazione di un insieme di testi latini a livello di *Coreference Resolution* (d'ora in poi CR) e *Anaphora Resolution* (d'ora in poi AR).

CR e AR sono due compiti distinti di Trattamento Automatico del Linguaggio (TAL), i cui termini talvolta vengono impiegati nella letteratura scientifica in modo ambiguo. In questo articolo si adottano le definizioni proposte da Sukthanker *et al.* (2020), secondo cui la CR consiste nell'identificazione di tutte le menzioni che fanno riferimento alla stessa entità del mondo reale all'interno di un testo o di testi diversi (*cross-document CR*), anche qualora esse differiscano per forma o struttura grammaticale. La nozione di CR include riferimenti che non seguono necessariamente la progressione lineare del testo, come, ad esempio, la catafora. La AR, invece, concerne l'identificazione della relazione anaforica tra un termine e il suo antecedente, dal quale dipende la sua interpretazione (cfr. Sukthanker *et al*. 2020; Poesio *et al*. 2016).

[1] Si vedano, tra gli altri, Bolkenstein (2000); Joffre (2007); Spevak (2007); Longrée (2010); Pieroni (2011).

I compiti di CR e AR vertono entrambi sull'individuazione di menzioni linguistiche che rimandano ad altro, con la differenza che, nella CR, menzione ed entità sono collegate da una relazione di identità, mentre, nell'AR, esse sono connesse tramite relazioni testuali di natura sintattica, logica o metonimica. Vi sono anche casi di anafore coreferenziali, come nell'esempio: "Mi sono accorto di non avere più l'*orologio*, devo aver*lo* perso al mercato"[2], in cui sussiste identità tra la menzione (aver*lo*) e l'antecedente (*orologio*). Come già rilevato, però, non tutte le anafore sono coreferenziali: "Ieri ho comprato un *orologio*. Il *cinturino* è in pelle" costituisce un caso di relazione anaforica indiretta[3], in quanto *cinturino* e *orologio* non designano la stessa entità e pertanto non sono coreferenti.

Adattando queste nozioni a esigenze operative relative all'arricchimento di un insieme di testi letterari latini con metadati relativi a CR e AR, nel progetto *CorefLat* si è deciso di annotare tutte le menzioni che rimandano a un'entità espressa nel testo, indipendentemente dall'ordine in cui si trovano, come sarà illustrato più dettagliatamente nel §3.

Procediamo ora a delineare sinteticamente lo stato dell'arte relativo a CR/AR nella ricerca contemporanea in TAL[4], soffermandoci in particolare sugli studi e gli strumenti applicati alle lingue classiche.

Già nel 1983 Roberto Busa sottolineava la mancanza di ricerche sistematiche sulla risoluzione pronominale (cfr. Nyhan – Passarotti 2019). A partire dagli anni Novanta, la ricerca sulla CR e AR automatica si orientò verso approcci basati su algoritmi di *machine learning*. Ciò fu possibile grazie allo sviluppo di *corpora* arricchiti con annotazioni CR/AR e a iniziative come il *Message Understanding Conference* (MUC; Chinchor 1998) e l'*Automatic Content Evaluation* (ACE; Doddington *et al.* 2004). Successivamente, diversi *corpora* annotati hanno

[2] Gli esempi sono tratti da Ježek – Sprugnoli (2023: 60).

[3] Questi casi sono stati descritti con il termine di *associative anaphoras* cfr. Kleiber (1999), o, in TAL, *bridging anaphora* cfr. Caselli (2009).

[4] Per ragioni di spazio, questa sezione offre soltanto una sintesi dello stato della questione; per una rassegna più ampia e dettagliata della letteratura di riferimento si rimanda a Poesio *et al*. (2023).

esteso la copertura linguistica includendo il tedesco (Hinrichs *et al*. 2004), il giapponese (Iida *et al*. 2007), lo spagnolo (Recasens – Martí 2010), il ceco (Nedoluzhko *et al.* 2016) e l'italiano (Minutolo *et al.* 2022). È inoltre opportuno ricordare che anche progetti, come *Universal Anaphora*[5] ed *Enhanced Dependencies* (ED)[6] nell'ambito del framework *Universal Dependencies* (UD)[7], hanno dedicato crescente attenzione ai fenomeni coreferenziali, lavorando a un modello condiviso di annotazione multilingue.

Per quanto concerne le lingue classiche, risorse fondamentali sono l'*Ancient Greek and Latin Dependency Treebank* (AGLDT)[8], che comprende estratti di testi greci e latini dell'età classica, e l'*Index Thomisticus Treebank* (IT-TB)[9], che raccoglie testi latini medievali di Tommaso d'Aquino. L'annotazione sintattica di entrambi i *corpora* si fondava inizialmente su uno schema affine al livello analitico del *Prague Dependency Treebank* (PDT; Bamman *et al*. 2008). Ai fini della nostra ricerca, è interessante notare che entrambe le *treebank* includono un sottoinsieme di dati annotato a livello tectogrammaticale del PDT (cfr. Mambrini 2013; Passarotti 2014; Passarotti – González Saavedra 2017). Tale livello consente di rappresentare la struttura sintattica profonda delle frasi ed estende l'annotazione a ulteriori fenomeni, quali *semantic role labeling*, struttura informativa, ellissi e, appunto, coreferenza.

Per quanto riguarda il latino, che costituisce il *focus* del presente studio, circa 45.000 *token* tratti da AGLDT e IT-TB sono stati arricchiti con annotazione tectogrammaticale, comprendendo estratti da Sallustio, Cesare, Cicerone (AGLDT) e Tommaso d'Aquino (IT-TB). Nonostante la loro rilevanza, tali annotazioni CR/AR restano tuttavia disomogenee, poiché oltre metà delle occorrenze proviene dalla *Summa*

[5] https://universalanaphora.github.io/UniversalAnaphora/.
[6] https://universaldependencies.org/u/overview/enhanced-syntax.html.
[7] https://universaldependencies.org/.
[8] https://perseusdl.github.io/treebank_data/.
[9] http://itreebank.marginalia.it.

contra Gentiles di Tommaso d'Aquino (circa 27.000) e dalla *In Catilinam* di Sallustio (circa 10.936) (cfr. González Saavedra – Passarotti 2018).

Per compensare questa disparità, nell'ambito del progetto *CorefLat* abbiamo sviluppato un insieme di annotazioni CR/AR più ampio ed omogeneo. Tali annotazioni sono state realizzate per essere pubblicate come *Linked Open Data* (LOD) e pertanto sono stati selezionati testi latini già disponibili come LOD all'interno della *LiLa Knowledge Base*[10].

Il presente contributo illustra il processo di annotazione e di pubblicazione come LOD dell'insieme dei dati arricchiti con CR/AR fornito da *CorefLat*. La struttura dell'articolo è la seguente: il §2 offre una breve introduzione alla *LiLa Knowledge Base*; il §3 presenta le linee guida di annotazione di *CorefLat* e i primi risultati ottenuti; il §4 si concentra sulla modellizzazione adottata per descrivere i dati e sulla pubblicazione di *CorefLat* come LOD (§4.1), esamina alcuni esempi (§4.2) e propone uno studio di caso che mostra l'interazione tra *CorefLat* e le altre risorse collegate a *LiLa* (§4.3). Infine, il §5 presenta alcune considerazioni conclusive e prospettive di ricerca.

2. La *LiLa knowledge base*

Il progetto *LiLa – Linking Latin* (Passarotti *et al*. 2020) ha ottenuto un *ERC Consolidator Grant* (2018-2023) con l'obiettivo di integrare le risorse linguistiche esistenti per il latino all'interno di una *Knowledge Base* (KB) pubblicata come LOD, al fine di garantirne l'interoperabilità in rete.

La *LiLa Knowledge Base* è stata sviluppata adottando gli standard consolidati per la pubblicazione dei dati nel contesto del *Semantic Web*, in conformità ai principi del cosiddetto paradigma dei *Linked Data* (cfr. Berners-Lee 1996). Ciascun dato contenuto nelle risorse linguistiche interconnesse nella KB è pertanto dotato di un URI (*Uniform Resource*

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

Identifier) univoco e persistente, pubblicato sul Web come URL mediante protocollo HTTP, così da assicurarne rintracciabilità e accessibilità. L'impiego di standard del Web, quali il modello di dati RDF (*Resource Description Framework*) e il linguaggio di interrogazione SPARQL[11], consente a *LiLa* di creare collegamenti tra URI distinti e di favorire il riuso dei dati.

La KB di *LiLa* si avvale inoltre di alcune ontologie preesistenti per rappresentare i (meta)dati delle risorse latine in essa interconnesse. Tra le principali ontologie integrate figurano POWLA (*Portable Linguistic Annotation with OWL*), un'ontologia concepita per esprimere dati e metadati testuali come LOD (cfr. Chiarcos *et al.* 2012); OLiA per l'annotazione linguistica (*Ontologies of Linguistic Annotation*), un insieme di ontologie che permettono di rappresentare e mettere in corrispondenza le categorie linguistiche (cfr. Chiarcos – Sukhareva 2015) e OntoLex-Lemon per i dati lessicali (cfr. McCrae *et al.* 2017).

Nell'architettura della KB di *LiLa*, un ruolo centrale è svolto dai lemmi, che costituiscono i nodi di connessione principali tra risorse lessicali e testuali. Tale architettura, fortemente basata sul lessico, si fonda sul presupposto che le risorse testuali sono costituite da occorrenze di parole (*token*), mentre le risorse lessicali descrivono le proprietà delle parole nelle rispettive voci lessicali. La KB si basa quindi sulla cosiddetta *Lemma Bank*, un insieme di circa 200.000 lemmi latini (forme canoniche di citazione delle unità lessicali), pubblicati come LOD. Tale insieme ha origine dalla base lessicale dell'analizzatore morfologico per il latino *LEMLAT 3.0* (cfr. Passarotti *et al.* 2017) e viene costantemente ampliato sulla base dell'integrazione di nuove risorse nella KB. L'interoperabilità è assicurata dal fatto che tutte le voci lessicali e i *token* sono collegati al lemma corrispondente contenuto nella *Lemma Bank*, rendendo così possibile una piena integrazione tra le diverse risorse.

[11] https://www.w3.org/TR/rdf-sparql-query/. L'*endpoint SPARQL* di *LiLa* è accessibile all'indirizzo https://lila-erc.eu/sparql/.

3. *CorefLat*: INISIEME DEI DATI ANNOTATI

3.1. Linee guide e criteri di annotazione

La presente sezione offre una panoramica dell'insieme dei dati annotati nell'ambito del progetto *CorefLat*, con particolare attenzione alle linee guida che hanno orientato il processo di annotazione.

Tale processo si fonda, innanzitutto, sull'individuazione di due elementi: l'*Entity*, ossia le entità cui si fa riferimento, e la *Mention*, vale a dire le espressioni linguistiche che riprendono o anticipano un'entità[12]. Nell'ambito di *CorefLat* vengono quindi identificate le relazioni tra una *Mention* e una *Entity*, piuttosto che le catene di coreferenze. Le prime rappresentano il legame specifico tra due o più espressioni che rinviano alla stessa entità all'interno di un testo e si distinguono dalle catene coreferenziali per ampiezza e struttura. Queste ultime, infatti, consistono in sequenze di molteplici espressioni referenziali che rimandano alla medesima entità. A titolo esemplificativo, si considerino le seguenti frasi: "Maria ama il suo gatto. Lei si prende cura di lui con molta attenzione e lui si diverte a giocare con lei". Le coreferenze relative a "Maria" possono essere annotate sia sotto forma di relazioni sia sotto forma di catene. Nel primo caso, le relazioni coreferenziali si articolano secondo uno schema uno a uno tra la menzione e l'entità a cui si riferisce: *suo → Maria; lei → Maria; lei → Maria*. Nel secondo caso, invece, l'annotazione a catena comporta il collegamento dei vari elementi che si riferiscono alla stessa entità: *suo → lei → lei → Maria*.

Alla luce di quanto esemplificato, si è scelto, nell'ambito di *CorefLat*, di circoscrivere l'annotazione a un numero limitato di tipologie di relazioni coreferenziali[13]. Tale scelta è in linea con l'obiettivo di arric-

[12] Si precisa che il processo di annotazione è stato condotto seguendo le linee guida del *corpus* GUM, adottate anche nell'ambito del progetto Universal Anaphora (UA), con l'obiettivo di garantire coerenza tra le diverse risorse linguistiche arricchite con annotazioni di CR e AR. Cfr. https://wiki.gucorpling.org/gum/entities.

[13] È opportuno precisare che le etichette qui fornite svolgono una funzione operativa nel processo di annotazione, ma non figurano nell'*output* della risorsa. Come verrà

chire con annotazioni coreferenziali una variegata collezione di testi latini in modo quanto più possibile uniforme e coerente. In (1)-(4) vengono forniti alcuni esempi delle tipologie di coreferenze annotate in *CorefLat*.

1. Coreferenze *one-to-one*: una *Mention* (composta da un solo *token*) rimanda (o anticipa) a un'entità anch'essa rappresentata da un singolo *token*. Come verrà discusso più in dettaglio nel §3.2, questa tipologia di coreferenza rappresenta la categoria di gran lunga più frequente nell'insieme dei dati annotati.

a. *hoc proelio trans Rhenum nuntiato Suebi, qui ad ripas Rheni venerant, domum reverti coeperunt* "Annunciata questa battaglia al di là del Reno, i Suebi, che erano giunti sulle rive del Reno, cominciarono a ritornare in patria" (Caes. *Gal*. 1, 54).
b. *invocat te, domine.* "Invoca te, Dio" (Aug. *Conf*. 1, 1, 1).
c. *Laudes tuae, domine, laudes tuae per scripturas tuas suspenderent palmitem cordis mei* "Le tue lodi, Signore, le tue lodi innalzino, attraverso le tue Scritture, il ramo del mio cuore" (Aug. *Conf*. 1, 17, 27).
d. *Quoniam itaque et ego sum, quid peto ut venias in me, qui non essem nisi esses in me?* "Dunque, poiché anch'io esisto, perché chiedo che tu venga in me, che non sarei se non fossi in me?" (Aug. *Conf*. 1, 2, 8).
e. *Quo usque, quaeso, ad hunc modum / inter nos amore utemur semper surrupticio*? "Fino a quando, ti prego, in questo modo tra noi vivremo un amore sempre furtivo" (Pl. *Curc*. 1, 204-205).

L'anafora pronominale costituisce, nel nostro *corpus*, il caso prototipico della coreferenza che qui abbiamo definito come *one-to-one*: in queste relazioni la *Mention* è per l'appunto rappresentata da un pronome, come accade con il pronome relativo *qui* nell'esempio (1a) e con

illustrato più avanti (§4), la risorsa si limita infatti a stabilire relazioni tra entità e menzioni.

il pronome *te* nell'esempio (1b). In questa categoria includiamo anche le relazioni anaforiche che coinvolgono due parole piene e identiche. Tali contesti possono essere interpretati come delle anafore stilistiche, realizzate attraverso la ripetizione di un medesimo lessema, in cui la seconda occorrenza rinvia alla prima, come in (1c). In questo caso, la prima occorrenza del lessema funziona come *Entity*, mentre la seconda svolge il ruolo di *Mention*.

Gli esempi (1d) e (1e) sono stati annotati come casi particolari. In (1d), i pronomi di prima persona rinviano all'autore Agostino, il cui nome non compare mai esplicitamente nel primo libro delle *Confessiones*. In (1e), invece, il pronome di prima persona plurale rinvia a Planesio e Fedromo, due dei principali personaggi del *Curculio* di Plauto, menzionati nel testo a notevole distanza dall'occorrenza di *nos*[14]. Per queste due tipologie di relazioni coreferenziali si è ritenuto opportuno introdurre un'etichetta operativa, quella di *external entity*. Tale etichetta comprende tutte le istanze di coreferenza che richiedono al lettore (così come all'annotatore) di compiere un'inferenza basata sulla conoscenza del mondo e del testo o sul contesto conversazionale.

2. *Split antecedent*[15]: la *Mention* presenta antecedenti multipli, vale a dire che un singolo elemento referenziale rinvia a (o anticipa) più di un'entità. Nella nostra annotazione tale fenomeno può manifestarsi attraverso delle menzioni che rinviano a sintagmi nominali congiunti (2a) o disgiunti (2b) o attraverso delle menzioni che rinviano a elementi di un elenco (2c).

[14] Per garantire la coerenza dell'annotazione, è stata fissata una soglia arbitraria: quando una menzione dista più di cinque frasi dall'entità cui si riferisce, essa viene collegata a un'*external entity*. Tale soglia è definita in termini di frasi piuttosto che di *token*, in conformità con la prassi consolidata negli studi di CR e AR, nei quali la frase costituisce l'unità primaria di analisi. Per una discussione più approfondita su questo argomento, si rimanda a Delfino *et al.* (2024).

[15] Per un'analisi puntuale di questa tipologia di coreferenza si rimanda a Sukthanker *et al.* (2020: 141) e ai riferimenti ivi citati.

a. *An vero caelum et terra, quae fecisti et in quibus me fecisti, capiunt te?* "O forse il cielo e la terra, che tu hai creato e nei quali mi hai creato, possono contenerti?" (Aug. *Conf.* 1, 2, 2).
b. *Nec mater mea vel nutrices meae sibi ubera implebant, sed tu mihi per eas dabas alimentum infantiae* "E non mia madre né le mie nutrici riempivano da sé il loro seno, ma Tu, attraverso di loro, mi davi il nutrimento dell'infanzia" (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* "Tutta la Gallia è divisa in tre parti: una la abitano i Belgi, un'altra gli Aquitani, la terza quelli che nella loro lingua sono chiamati Celti e nella nostra Galli. Tutti costoro differiscono tra loro per lingua, istituzioni e leggi" (Caes. *Gal.* 1, 1, 1).

3. *Multiword antecedent*: l'entità coinvolta nella relazione di coreferenza è costituita da più di un *token*. Nella nostra annotazione questa tipologia ricorre principalmente in due contesti: quando l'entità è un nome proprio strutturato secondo il sistema onomastico romano (*tria nomina*), come illustrato in (3a); oppure quando l'entità è realizzata come sintagma nominale accompagnato da un modificatore o da uno specificatore che ne restringe semanticamente il riferimento, come illustrato in (3b).

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* "Pertanto, prima di tentare qualsiasi cosa, ordina che sia convocato presso di lui Divizziaco e, fatti allontanare gli interpreti abituali, parla con lui tramite Gaio Valerio Troucillo, notabile della Gallia Narbonense, suo amico, al quale accordava la massima fiducia in tutte le cose" (Caes. *Gal.* 1, 19, 3).

b. *Comprecor vulgus silentum vosque ferales deos et Chaos caecum atque opacam Ditis umbrosi domum* "Vi supplico, o popolo silenzioso, voi dei funesti e il Caos cieco e la tenebrosa dimora di Dite ombroso" (Sen. *Med.* 740-741).

Definite le tipologie di relazioni coreferenziali oggetto dell'annotazione, occorre illustrare i testi finora scelti per la realizzazione del progetto *CorefLat*. Per garantire l'applicabilità delle linee guida esposte *supra* ai diversi contesti linguistici e stilistici, è stato selezionato un insieme eterogeneo di testi che variano per genere letterario e periodo storico, con l'obiettivo di fornire un campione quanto più possibile rappresentativo della lingua latina. Allo stato attuale, *CorefLat* include una commedia arcaica (*Curculio* di Plauto), un estratto da un testo storiografico di età classica (il primo libro del *De Bello Gallico* di Cesare), una tragedia di età imperiale (*Medea* di Seneca) e un passo da un'opera filosofica tardoantica (il primo libro delle *Confessiones* di Agostino), per un totale di 25.965 *token*[16].

L'annotazione è stata condotta manualmente da due annotatrici, con il supporto del *Content Annotation Tool* (CAT), uno strumento flessibile e personalizzabile[17]. I (meta)dati sono stati dapprima salvati in formato XML e successivamente trasformati in modo automatico nel formato CoNLL-U Plus[18], seguendo le linee guida indicate dal progetto UA (cfr. Lenzi *et al.* 2012). La disamina dei testi è stata equamente distribuita tra le due annotatrici; tuttavia, per valutare la qualità delle annotazioni prodotte, si è stabilito che entrambe lavorassero autonomamente sulle ultime cinquanta frasi del primo libro delle *Confessiones* di Agostino. In questo modo, è stato possibile calcolare l'accordo tra le

[16] I dati del latino classico provengono dal corpus *Opera Latina* del LASLA, che contiene oltre 1,7 milioni di parole tratte sia da testi di latino classico sia di latino tardo (https://lasladb.uliege.be/OperaLatina/), mentre gli esempi di latino tardo sono tratti da *The Latin Library* (http://www.m.thelatinlibrary.com/).

[17] Ringraziamo Giovanni Moretti per il suo supporto con il *Content Annotation Tool*, su cui per maggiori approfondimenti si rimanda a Lenzi *at al.* (2012).

[18] https://universaldependencies.org/ext-format.html.

due, misurandolo mediante il coefficiente di Dice: una metrica di similarità ampiamente impiegata nell'ambito della linguistica computazionale (cfr. Dice 1945; Sørensen 1948) che varia da 0 (assenza di sovrapposizione) a 1 (identità perfetta degli insiemi). Dopo aver verificato che in ogni contesto, per entrambe le annotatrici, gli elementi annotati interessassero gli stessi *token*, sono stati calcolati i punteggi di similarità per le entità (0,817) e per le menzioni (0,824). I valori risultanti sono in linea con quelli attesi per il compito di risoluzione delle coreferenze (cfr. Cohen *et al.* 2017; Hendrickx *et al.* 2008; Nedoluzhko *et al.* 2009).

3.2. Primi risultati e considerazioni iniziali

In questa sezione vengono illustrati i primi risultati dell'annotazione e viene preliminarmente discusso come e in quale misura le coreferenze varino nei testi selezionati. In questa fase del progetto, tali osservazioni sono utili per orientare e affinare le successive tappe della ricerca, in particolare per quanto riguarda la scelta di ulteriori testi da annotare. Cominciamo prendendo in esame la Tabella 1, che presenta i risultati organizzati per testo.

Testo	**Coreferenze *one-to-one***	***Split Antecedent***	***Multiword Antecedent***	**Coref. Tot.**	**Token Tot.**
Pl. *Curc*.	228	4	0	232	5,853
Caes. *Bell.* I	630	67	26	723	8,272
Sen. *Med*.	142	2	3	147	5,707
Aug. *Conf*. I	399	14	5	418	6,133
Tot.	1399	87	34	1,520	25,965

Tabella 1: Risultati dell'annotazione, organizzati in base ai testi

Un primo aspetto da sottolineare riguarda il fatto che le coreferenze *one-to-one*, come da noi definite, rappresentano la tipologia più diffusa nell'insieme dei testi annotati. Si osserva, inoltre, come i testi in prosa

manifestino una densità di fenomeni coreferenziali sensibilmente superiore rispetto alle opere di carattere teatrale, come mostrato nella Tabella 2.

Testo	Tot. Coref.	Tot. Token	Coref./Token (%)
Pl. *Curc*.	232	5,853	3.96
Caes. *Bell.* I	723	8,272	8.74
Sen. *Med*.	147	5,707	2.58
Aug. *Conf*. I	418	6,133	6.82

Tabella 2: Distribuzione della densità coreferenziale nei testi.

In particolare, si osserva che la narrazione storiografica di Cesare si configura come il testo con il più elevato numero di coreferenze – anche per quanto riguarda i casi di *split antecedent* e di *multiword antecedent*. Tale tendenza può essere interpretata come il risultato congiunto di convenzioni proprie del genere storiografico e di tratti stilistici d'autore: i frequenti riferimenti di Cesare a gruppi etnici e il ricorso sistematico alle formule onomastiche romane (*tria nomina*) contribuiscono infatti alla formazione di strutture referenziali più complesse. La variazione del fenomeno coreferenziale tra i diversi generi si riscontra altresì – sia pure con una distribuzione differente – quando vengono presi in considerazione anche i dati relativi alle menzioni associate a delle *external entity*, come mostra la Tabella 3. Qui i dati sono presentati in termini assoluti e normalizzati per 1.000 *token*, consentendo un confronto più equilibrato tra testi di lunghezza differente[19].

[19] È stato effettuato un test chi-quadro di indipendenza per esaminare la relazione tra i testi e le occorrenze delle relazioni con *internal entity* e delle relazioni con *external entity*. Il valore di p è < 0.00001 e il risultato è significativo a $p < .05$. L'ipotesi nulla, secondo cui la distribuzione proporzionale delle relazioni con *external entity* rispetto a quelle con *internal entity* sia la stessa nei quattro testi (indipendenza), può quindi essere respinta.

Testo	Relazioni con *External entity*	Relazioni con *Internal entity*	Tot. token	Ext./ 1000 token	Int./ 1000 token
Pl. *Curc*.	454	232	5,853	77.6	39.6
Caes. *Bell.* I	57	723	8,272	6.9	8.4
Sen. *Med*.	132	147	5,707	23.1	25.8
Aug. *Conf*. I	244	418	6,133	39.8	68.2

Tabella 3: Panoramica delle coreferenze che coinvolgono *external entity* nell'insieme di dati annotati, in termini assoluti e normalizzati per 1.000 *token*.

Nel caso dei testi drammatici si osservano frequenze particolarmente elevate di relazioni con *external entity*, con il *Curculio* che da solo raggiunge un totale di 454 occorrenze. Tale dato può essere plausibilmente interpretato alla luce della dimensione performativa del genere, in cui i personaggi, fisicamente presenti sulla scena, sono facilmente identificabili per il pubblico. Ci sembra poi interessante evidenziare la divergenza tra i due testi in prosa. In particolare, le *Confessiones* mostrano una frequenza di relazioni con *external entity* marcatamente differente rispetto al *De Bello Gallico*. Il testo agostiniano, infatti, presenta una distribuzione delle relazioni con *external entity* più vicina a quella delle opere teatrali che a quella del testo storiografico. Tale tendenza può essere ricondotta alla natura delle *Confessiones*, un'opera di prosa filosofica strutturata come un discorso diretto tra l'autore e Dio, e che quindi condivide con i testi drammatici la forma prevalentemente dialogica.

La variazione emersa da questi primi risultati sembra essere in linea con le ricerche – tra le quali citiamo quelle di Bolkestein (2000), Pieroni (2011) e Longrée (2004) – che hanno già dimostrato che i fenomeni referenziali sono fortemente influenzati da variabili pragmatiche e testuali e conseguentemente tendono a oscillare in funzione del genere, del contesto comunicativo e delle strategie stilistiche dell'autore. Questi risultati preliminari confermano quindi come, al fine di costruire un campione significativo di testi arricchito con coreferenze per la lingua

latina, sia essenziale garantire la massima diversificazione nella scelta di testi e autori[20].

4. Pubblicare *CorefLat* in *LiLa*

4.1. Modellizzazione

Questa sezione illustra come è stato modellizzato l'insieme dei dati arricchiti con coreferenze. Le soluzioni adottate mirano a collegare il testo annotato alla *LiLa* KB e a garantire l'interoperabilità dell'annotazione coreferenziale con gli altri *Linguistic Linked Open Data* in *LiLa*. In primo luogo, la *LiLa CorefLat Ontology* è un'ontologia OWL che estende il framework POWLA[21], condiviso anche dagli altri *corpora* annotati in *LiLa* (cfr. Mambrini *et al.* 2022)[22].

Al livello più alto di astrazione, la *CorefLat Ontology* introduce una classe chiamata *Coreference Element*[23], che funge da macrocategoria per tutte le entità e le relazioni coinvolte nell'annotazione coreferenziale. La classe *Coreference Element* include come sottoclassi *Entity*, *Mention*, *Coreference Unit* e *Coreference Relation*.

- Le prime due sottoclassi servono a rappresentare rispettivamente il ruolo di entità e di menzione nella relazione coreferenziale.

[20] Quest'analisi rappresenta un punto di partenza per approfondimenti futuri. In particolare, sembra promettente esplorare questi dati anche in relazione ad altri fenomeni referenziali, come la deissi. Siamo particolarmente grati al Revisore 2 per aver sottolineato il potenziale del nostro studio in questa direzione, che, allo stadio attuale del progetto, non è stato possibile sviluppare.

[21] Il framework POWLA definisce quattro concetti di base per descrivere i *corpora*: *document*, *layer*, *node* e *relation*. Mentre i primi due figurano anche nell'ontologia che rappresenta i dati disponibili in *LiLa* (http://lila-erc.eu/ontologies/lila_corpora/), *node* e *relation* risultano particolarmente appropriati per modellizzare le informazioni annotate da *CorefLat*.

[22] L'ontologia è disponibile qui: https://lila-erc.eu/lodview/ontologies/lila_coref/.

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

- La classe *Coreference Relation*[24], sfruttando le caratteristiche delle relazioni in POWLA[25], etichetta e orienta la relazione tra entità e menzioni.
- La classe *Coreference Unit* è stata pensata per modellizzare la relazione di coreferenza non fra specifici *token*, ma fra elementi più astratti, le *Coreference Unit*; queste ci consentono di rappresentare anche i casi di *multiword antecedent*, in cui un elemento coreferenziale è costituto da più *token* (cfr. §4.2).
- L'ontologia *CorefLat* si serve di alcune proprietà per descrivere le relazioni tra i *token* e le *Coreference Unit*, oltre che specificare il ruolo di queste ultime all'interno della *Coreference Relation*.
- La proprietà *hasCoreferenceTerminal*[26] collega ciascuna *Coreference Unit* ai *token* coinvolti nella relazione di coreferenza[27].
- Le proprietà *hasCoreferenceSource* e *hasCoreferenceTarget*[28] collegano la relazione di coreferenza alla sua *source* (che, per convenzione, corrisponde alla *Coreference Unit Mention*) e al suo *target* (che, per convenzione, corrisponde alla *Coreference Unit Entity*).
- Le proprietà *hasMention* e *hasEntity*, sono sottoproprietà rispettivamente di *hasCoreferenceSource* e *hasCoreferenceTarget*. Queste hanno come *range* le classi *Mention* ed *Entity* e stabiliscono un'interpretazione più restrittiva della coreferenza, concepita come relazione diretta da una menzione verso un'entità. Tuttavia, gli utenti dell'ontologia possono scegliere se aderire a questa interpretazione

[24] http://purl.org/powla/powla.owl#Relation.
[25] POWLA adotta un approccio reificato, secondo il quale tutte le relazioni vengono istanziate come risorse RDF, dotate di un proprio URI. Ciò consente, ad esempio, di specificare quale annotatore abbia etichettato una relazione coreferenziale oppure di associare a quell'etichetta un valore che indichi il livello di attendibilità.
[26] http://lila-erc.eu/ontologies/lila_coref/hasCoreferenceTerminal.
[27] Si noti che l'ordine lineare dei *token* nelle risorse testuali pubblicate in *LiLa* è rappresentato grazie alle relazioni simmetriche *next* e *previous* di POWLA, che collegano i nodi testuali in una catena. La sequenza dei *token* all'interno delle unità di coreferenza può quindi essere espressa utilizzando queste due proprietà di POWLA.
[28] http://lila-erc.eu/ontologies/lila_coref/hasCoreferenceSource;http://lila-erc.eu/ontologies/lila_coref/hasCoreferenceTarget

rigorosa oppure adottare un modello più flessibile, che includa semplicemente le *Coreference Unit*.

Infine, per completare la presentazione della *LiLa CorefLat Ontology*, è necessario introdurre il nodo extratestuale, concepito per armonizzare e riconoscere le diverse *Entity-Type Coreference Unit*[29] e, al tempo stesso, permettere in futuro dei task di risoluzione di coreferenza *cross-document.* Questo nodo, collegato alle entità tramite la proprietà *itsdf:taLentReef* dell'ontologia ITS (*Internationalization Tag Set*)[30], svolge la funzione di aggregatore: semplifica le *query* all'interno della risorsa e facilita il collegamento dei dati annotati con altre fonti di conoscenza, come le risorse enciclopediche *DBpedia*[31] o *Wikidata*[32], attraverso le proprietà di mapping del *Simple Knowledge Organization System* (SKOS), come *skos:exactMatch*. In tal modo diventa possibile arricchire tali nodi con proprietà ereditate transitivamente da tali risorse.

4.2 Esempi in LOD

Questa sezione illustra come la modellizzazione dei dati realizzata all'interno del progetto *CorefLat* venga applicata negli esempi (1)-(3) presentati in §3.1[33].

[29] Questo nodo extratestuale è infatti utile per rappresentare le entità che abbiamo definito come *external entity*, si veda l'esempio (1e).
[30] https://github.com/w3c/itsrdf.
[31] https://www.dbpedia.org/.
[32] https://www.wikidata.org/.
[33] Gli esempi (1b)-(1d) non saranno spiegati nel dettaglio, poiché seguono la modellizzazione di (1a). Analogamente, gli esempi (2b) e (2c) ricalcano la modellizzazione (2a).

La Figura 1 mostra la rappresentazione dell'esempio (1a), in cui sussiste una relazione di coreferenza tra il *token qui* (*Mention*) e il *token domine* (*Entity*)[34].

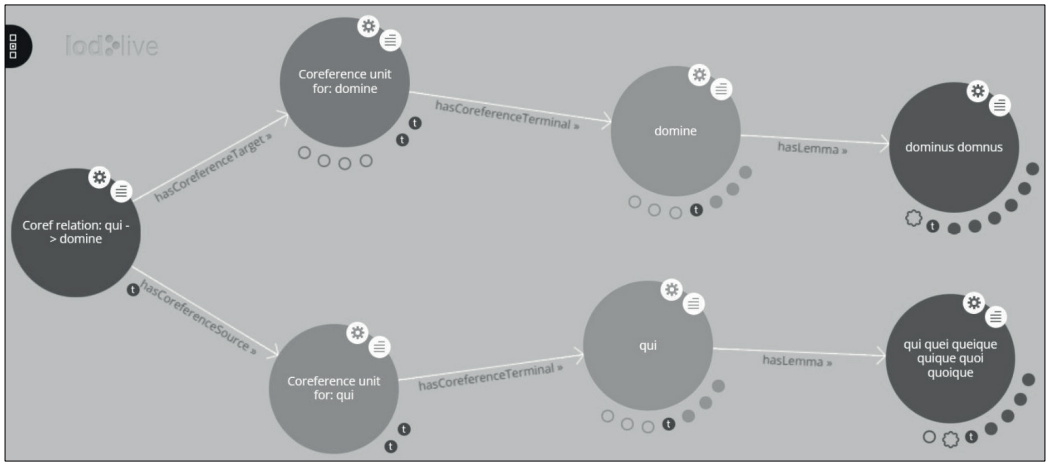

Figura 1. Visualizzazione LODLive della relazione coreferenziale fra *qui* - *domine* in (1a).

Il *token domine* è l'oggetto della proprietà *hasCoreferenceTerminal*, il cui soggetto è la *Coreference Unit Entity* corrispondente a *domine*. Lo stesso vale per il *token qui*, che è l'oggetto della proprietà *hasCoreferenceTerminal*, il cui soggetto è la *Coreference Unit Mention* corrispondente a *qui*. Queste *Coreference Unit* sono collegate attraverso un nodo che reifica la loro relazione. Questo nodo è di tipo *Coreference Relation* e funge da soggetto di due proprietà:

1. *hasCoreferenceSource*, che ha come oggetto la *Coreference Unit Mention* corrispondente a *qui*;
2. *hasCoreferenceTarget*, che ha come oggetto la *Coreference Unit Entity* corrispondente a *domine*.

[34] Le visualizzazioni sono generate tramite un'istanza dell'applicazione web LodLive, eseguita su un *server* del progetto *LiLa* https://lila-erc.eu/lodlive/.

Infine, entrambi i *token* sono collegati al rispettivo lemma nella *Lemma Bank* tramite la proprietà *lila:hasLemma*[35].

La Figura 2 mostra l'approccio adottato nella nostra ontologia per rappresentare i casi di *split antecedent*, in cui la stessa *Mention* instaura una *Coreference Relation* con due *Entity* differenti. Si ricorderà che nell'esempio (2a), una singola *Mention* (*quae*) faceva riferimento a due *Entity* distinte (*caelum* e *terra*).

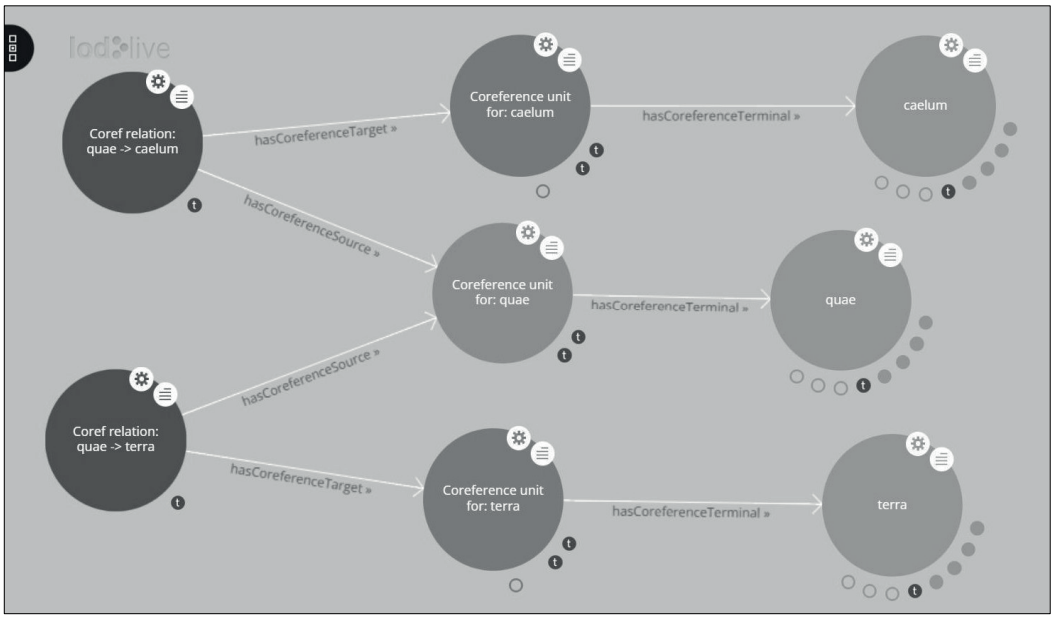

Figura 2. Visualizzazione LODLive delle relazioni coreferenziali *quae* - *caelum* e *quae* - *terra*: un caso di *split antecedent.*

Come si può osservare nella *Figura 2*, il *token quae* è l'oggetto della proprietà *hasCoreferenceTerminal*, il cui soggetto è la *Coreference Unit Mention* corrispondente a *quae*. Questa *Coreference Unit*, a sua volta, è l'oggetto della proprietà *hasCoreferenceSource* per due distinte *Coreference Relations*: *CorefRelation quae → caelum* e *CorefRelation quae → terra*. La prima *Coreference Relation* è l'oggetto della pro-

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

prietà *hasCoreferenceTarget*, il cui soggetto è la *Coreference Unit Entity* corrispondente a *caelum*. Analogamente, la seconda *Coreference Relation* è l'oggetto della proprietà *hasCoreferenceTarget*, il cui soggetto è la *Coreference Unit Entity* corrispondente a *terra*.

Entrambe le *Coreference Unit* fungono da soggetti della proprietà *hasCoreferenceTerminal,* che ha come oggetti rispettivamente i *token caelum* e *terra*.

Come abbiamo visto, i contesti di *split antecedent* si differenziano dai casi di *multiword antecedent*, come in (3a), dove i tre *token Caius Valerius Troucillus* indicano la stessa entità. Questo contesto viene modellizzato come l'esempio (1a), con l'unica differenza sostanziale che la *Coreference Unit Entity* è soggetto della proprietà *hasCoreferenceTerminal* per tre volte: l'oggetto della prima è il *token Caium*, l'oggetto della seconda è il *token Valerium*, e l'oggetto della terza è il *token Troucillum*.

Ricordiamo qui che proprio per casi come questi si è ritenuto necessario optare, all'interno della nostra modellizzazione, per una concettualizzazione astratta delle *Coreference Relation* basata sulle *Coreference Unit*, piuttosto che sui singoli *token*. Se la *Coreference Relation* fosse stata stabilita tra *token*, non sarebbe stato possibile distinguere un caso di *multiword* come quello osservato in (3a) da un caso di *split antecedent*, come quello visto in (2a).

4.3 Studio di caso

Questa sezione mostra l'interazione tra *CorefLat* e le altre risorse già presenti all'interno della *LiLa* KB. Ad esempio, risulta particolarmente interessante osservare come *CorefLat* possa interagire con la risorsa lessicale *Latin WordNet*, disponibile in *LiLa*[36].

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

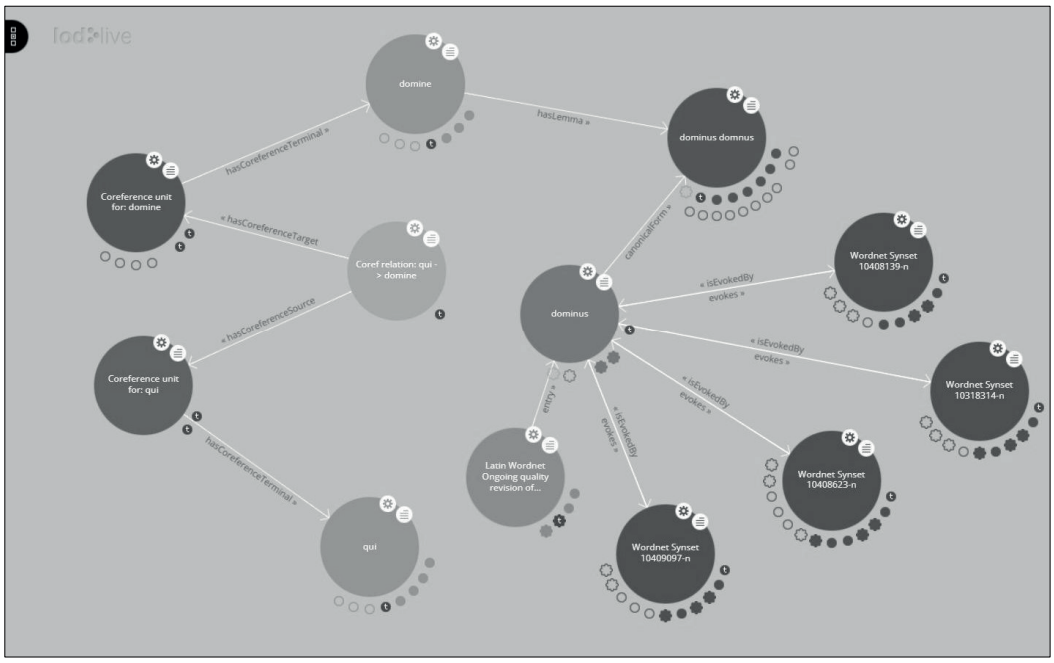

Figura 3. Visualizzazione LodLive dell'interoperabilità tra *CorefLat* e *Latin WordNet* nella *LiLa* KB.

Nella parte sinistra della Figura 3 si vede che il *token domine* (da un passo delle *Confessiones* di Agostino) è il *terminal* di una *Coreference Unit* di tipo *Entity*, la quale costituisce il *target* di una *Coreference Relation*. Nella parte destra della Figura 3, si vede come, in virtù dell'architettura *lemma-centred* di *LiLa*, il *token domine* risulti collegato al lemma *dominus* della *LiLa Lemma Bank*, che è a sua volta collegato alle voci lessicali del *Latin WordNet*. Ogni voce lessicale nella risorsa è associata ai *synset* corrispondenti[37], com'è per *dominus* in figura.

Generalizzando da questo esempio, è possibile formulare una *query SPARQL* per recuperare tutti i *synset* evocati dalle voci lessicali associate ai lemmi collegati a *token* coinvolti in una *Coreference Relation*.

[37] «Ogni *synset*, identificato da un codice univoco, è formato da parole che sono considerate sinonimi e che codificano uno specifico senso, spiegato con una breve glossa e qualche esempio» (Ježek – Sprugnoli (2023: 141)).

Questa *query* consente l'estrazione dei lemmi e dei *synset* corrispondenti in un formato a due colonne, fornendo inoltre, per ciascun lemma, il numero di *token* coinvolti in una *Coreference Relation*, come mostrato nella Figura 4[38].

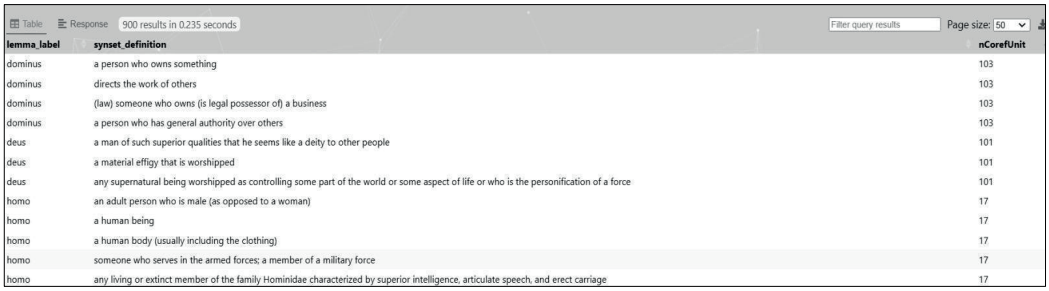

Table | Response | 900 results in 0.235 seconds | Filter query results | Page size: 50

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

Figura 4. *Output* di una *query SPARQL* che combina *CorefLat* e il *Latin WordNet* in *LiLa.*

5. Conclusioni e prospettive future

Riassumendo, questo lavoro ha introdotto *CorefLat*, una nuova risorsa progettata per supportare l'analisi della coreferenza in latino e per garantire una più ampia interoperabilità tra risorse all'interno della *LiLa Knowledge Base*.

Il progetto è ancora in corso e offre ampie possibilità di sviluppo, che andranno in queste direzioni: (*i.*) ampliamento dell'estensione dei dati annotati, includendo una gamma più vasta di generi testuali e di periodi storici, e (*ii*.) utilizzo dei dati arricchiti per addestrare modelli automatici di CR/AR per il latino, con una valutazione delle loro prestazioni sia su materiali *in-domain* sia *out-of-domain*.

In merito al punto (*i.*), precisiamo che, nella scelta dei testi per l'espansione dei dati disponibili con annotazione coreferenziale, si cercherà di favorire l'integrazione con altri livelli di informazione lingui-

[38] Si veda l'appendice per una visualizzazione dell'*output* della *query SPARQL* e del relativo codice.

stica. Più nello specifico, arricchire i testi che già presentano annotazione sintattica con informazioni coreferenziali permetterebbe di realizzare una risorsa più completa e versatile, adatta a supportare un ventaglio più ampio di ricerche future[39].

Infine, riconosciamo che il fenomeno della relazione coreferenziale, qui trattato unicamente con il fine dell'assemblamento di una risorsa, si rivela assai più complesso e tocca dimensioni che, per questioni di spazio e di *focus* della ricerca, non abbiamo potuto approfondire. Sfruttare la nostra risorsa per contribuire all'indagine di ulteriori aspetti (quali il rapporto tra relazione anaforica e meccanismi deittici, la logoforicità o, più in generale, le implicazioni stilistiche delle relazioni coreferenziali) costituisce tuttora un *desideratum* della ricerca.

Università degli Studi di Udine
Dipartimento di Lingue e Letterature. Comunicazione, Formazione e Società
eleonora.delfino@uniud.it

Università Cattolica del Sacro Cuore di Milano
Dipartimento di Scienze linguistiche e letterature straniere
robertagrazia.leotta@unicatt.it
marco.passarotti@unicatt.it
francesco.mambrini@unicatt.it

[39] In particolare, per quanto riguarda il latino classico, faremo uso di UD Latin-Circse (https://github.com/UniversalDependencies/UD_Latin-CIRCSE), un repository di *treebank* attualmente in fase di sviluppo presso il centro di ricerca CIRCSE dell'Università Cattolica del Sacro Cuore di Milano.

Bibliografia

Bamman, D. –Passarotti, M.C. –Busa, R. –Crane, G.
2008 *The annotation guidelines of the Latin Dependency Treebank and Index Thomisticus Treebank. The treatment of some specific syntactic constructions in Latin*, in *LREC 2008*, European Language Resources Association (ELRA), Marrakech, pp. 71-76.

Berners-Lee T.
1996 *WWW: Past, present, and future*, in «Computer», 29, pp. 69-77.

Bolkenstein, A.M.
2000 *Discourse organization and anaphora in Latin*, in Herring, S.C. – van Reenen, P. – Schoesler, L. (edd.), *Textual Parameters in Older Languages*, John Benjamins Publishing Company, Amsterdam – Philadelphia, pp. 107-138.

Caselli, T.
2009 *Using a generative Lexicon resource to compute bridging anaphora in Italian*, in «Procesamiento del Lenguaje Natural», 42, pp. 71-78.

Chiarcos, C. – Hellmann, S. –Nordhoff, S.
2012 *Linked Data in Linguistics: Representing and Connecting Linguistic Data and Language Metadata*, Springer, Heidelberg.

Chiarcos, C. – Sukhareva, M.
2015 *OLiA – Ontologies of Linguistic Annotation*, in «Semantic Web», 6, pp. 379-386.

Chinchor, N.A.
1998 *Overview of MUC-7*, in *Seventh Message Understanding Conference (MUC-7): Proceedings of a Conference Held in Fairfax, Virginia, April 29 – May 1, 1998*.

Cohen, K.B. – Lanfranchi, A. – Choi, M.J.-Y. – Bada, M. – Baumgartner, W.A. – Panteleyeva, N. – Verspoor, K. – Palmer, M. – Hunter, L.E.
2017 *Coreference annotation and resolution in the Colorado Richly Annotated Full Text (CRAFT) corpus of biomedical journal articles*, in «BMC Bioinformatics», 18, pp. 1-14.

Delfino, E. – Leotta, R.G. – Passarotti, M. – Moretti, G.
2024 *Building CorefLat. a Linguistic Resource for Coreference and Anaphora Resolution in Latin*, in *Proceedings of the 10th Italian Conference on Computational Linguistics (CLiC-it 2024)*, CEUR Workshop Proceedings, Pisa, pp. 273-279.

Dice, L.R.
1945 *Measures of the amount of ecologic association between species*, in «Ecology», 26, pp. 297-302.

Doddington, G. – Mitchell, A. – Przybocki, M. – Ramshaw, L. – Strassel, S. – Weischedel, R.
2004 *The automatic content extraction (ACE) program – tasks, data, and evaluation*, in Lino, M.T. – Xavier, M.F. – Ferreira, F. – Costa, R. – Silva, R. (edd.), *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC '04)*, European Language Resources Association (ELRA), Lisbon, pp. 837-840.

Hendrickx, I. – Bouma, G. – Coppens, F. – Daelemans, W. – Hoste, V. – Kloosterman, G. – Mineur, A.-M. – Van Der Vloet, J. – Verschelde, J.-L.
2008 *A coreference corpus and resolution system for Dutch*, in *LREC 2008*, European Language Resources Association (ELRA), Marrakech, pp. 144-149.

Hinrichs, E. – Kübler, S. – Naumann, K. – Telljohann, H. – Trushkina, J.
2004 *Recent developments in linguistic annotations of the TüBa-D/Z treebank*, Universitätsbibliothek Johann Christian Senckenberg, Frankfurt.

Iida, R. – Komachi, M. – Inui, K. – Matsumoto, Y.
2007 *Annotating a Japanese text corpus with predicate-argument and coreference relations*, in *Proceedings of the Linguistic Annotation Workshop*, Association for Computational Linguistics, Prague, pp. 132-139.

Ježek, E. – Sprugnoli, R.
2023 *Linguistica computazionale. Introduzione all'analisi automatica dei testi*, Il Mulino, Bologna.

Joffre, M.D.
2007 Ipse, *anaphore et deixis*, in Purnel, G. – Denooz, J. (edd.), *Ordre et cohérence en latin (Bruxelles-Liège, 4-9 avril 2005)*, Bibl. Philosophie et Lettres, Liège, pp. 99-110.

Kleiber, G.
1999 *Anaphore associative et relation partie-tout: condition d'aliénation et principe de congruence ontologique*, in «Langue française», pp. 70-100.

Lenzi, V.B. – Moretti, G. – Sprugnoli, R.
2012 *CAT: The CELCT Annotation Tool*, in *LREC 2012*, European Language Resources Association (ELRA), Istanbul, pp. 333-338.

Longrée, D.
2004 *Une approche statistique de la concurrence entre démonstratifs chez les historiens latins (César, Salluste, Tacite)*, in Bodelot, C. (ed.), *Anaphore, cataphore et corrélation en latin, Actes de la Journée d'étude de Linguistique latine, Université Blaise Pascal - Clermont-Ferrand 2, 7 janvier 2003*, Presses universitaires Blaise Pascal, Clermont-Ferrand, pp. 157-178.

2010 *Adverbes de lieu, deixis et anaphore chez les historiens latins*, in «Lingua Latina», 3, pp. 1-15.

Mambrini, F.
2013 *Thucydides 1.89-118: A multi-layer treebank*, in «CHS Research Bulletin», 1.

Mambrini, F. – Passarotti, M. – Moretti, G. – Pellegrini, M.
2022 *The Index Thomisticus Treebank as Linked Data in the LiLa Knowledge Base*, in Calzolari, N. – Béchet, F. – Blach, P. – Choukri, K. – Cieri, C. – Declerck, T. – Goggi, S. – Isahara, H. – Maegaard, B. – Mariani, J. – Mazo, H. – Odijk, J. – Piperidis, S. (edd.), *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, European Language Resources Association (ELRA), Marseille, pp. 4022-4029.

McCrae, J.P. – Bosque-Gil, J. – Gracia, J. – Buitelaar, P. – Cimiano, P.
2017 *The OntoLex-Lemon model: Development and applications*, in *Proceedings of eLex 2017 Conference*, pp. 19-21.

Minutolo, A. – Guarasci, R. – Damiano, E. – De Pietro, G. –
Fujita, H. – Esposito, M.
2022 *A multi-level methodology for the automated translation of a coreference resolution dataset: an application to the Italian language*, in «Neural Computing and Applications», 34, pp. 22493-22518.

Nedoluzhko, A. – Mírovský, J. – Pajas, P.
2009 *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)*, Association for Computational Linguistics, Suntec, pp. 108-111.

Nedoluzhko, A. – Novák, M. – Cinková, S. –
Mikulová, M. – Mírovský, J.
2016 *Coreference in Prague Czech-English Dependency Treebank*, in *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC '16)*, European Language Resources Association (ELRA), Portorož, pp. 169-176.

Nyhan, J. – Passarotti, M.
2019 *One Origin of Digital Humanities: Fr Roberto Busa in His Own Words*, Springer Nature, Cham.

Passarotti, M.
2014 *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)*, Association for Computational Linguistics, Gothenburg, pp. 100-109.

Passarotti, M. – González Saavedra, B.
2017 *The treebanked conspiracy. Actors and actions in* Bellum Catilinae, in Hajič, J. (ed.), *Proceedings of the 16th International Workshop on Treebanks and Linguistic Theories*, Prague, pp. 18-26.

Passarotti, M. – Budassi, M. – Litta, E. – Ruffolo, P.
2017 *The Lemlat 3.0 package for morphological analysis of Latin*, in *Proceedings of the NoDaLiDa 2017 Workshop on Processing Historical Language*, Linköping University Electronic Press, Gothenburg, pp. 24-31.

Passarotti, M. – Mambrini, F. – Franzini, G. – Cecchini, F.M. – Litta, E. – Moretti, G. – Ruffolo, P. – Sprugnoli, R.
2020 *Interlinking through lemmas. The lexical collection of the LiLa Knowledge Base of linguistic resources for Latin*, in «Studi e Saggi Linguistici», 58, pp. 177-212.

Pieroni, S.
2011 *Deixis and Anaphora* in Baldi, P. – Cuzzolin, P. (edd.) *Constituent Syntax: Quantification, Numerals, Possession, Anaphora*, De Gruyter, Berlin – Boston, pp. 389-502.

Poesio, M. – Stuckardt, R. – Versley, Y.
2016 *Anaphora Resolution*, Springer, Berlin – Heidelberg.

Poesio, M. – Yu, J. – Paun, S. – Aloraini, A. – Lu, P. – Haber, J. – Cokal, D.
2023 *Computational Models of Anaphora*, in «Annual Review of Linguistics», 9, pp. 561-587.

Recasens, M. – Martí, M.A.
2010 *Ancora-Co: Coreferentially annotated corpora for Spanish and Catalan*, in «Language Resources and Evaluation», 44, pp. 315-345.

Sorensen, T.
1948 *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*, in «Biologiske skrifter», 5, pp. 1-34.

Spevak, O.
2007 *L'anaphore, la deixis et l'ordre des constituants en latin*, in «Latomus», 66/4, pp. 853-870.

Sukthanker, R. – Soujanya, P. – Cambria, E. – Thirunavukarasu, R.
2020 *Anaphora and coreference resolution: A review*, in «Information Fusion», 59, pp. 139-162.

Appendice

Presentiamo di seguito la *query SPARQL* per ottenere i *synset* delle voci lessicali in *Latin WordNet* collegate ai lemmi della *Lemma Bank* di *LiLa*, i cui *token* corrispondono ai *terminal* di una *Coreference Unit Entity*, unitamente al numero di *Coreference Unit* in cui tali *token* sono coinvolti. I *synset* di *WordNet* sono considerati istanze della classe 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 .
}
```