La Knowledge Base LiLa

Interoperabilità tra risorse testuali e lessicali per il latino

Marco Passarotti
Università Cattolica del Sacro Cuore, Milano, Italia

This paper describes the LiLa Knowledge Base of interoperable linguistic resources for Latin. Resources are made interact by representing their (meta)data through common ontologies, vocabularies, and data categories, in accordance with the principles of the Linked Data paradigm. After presenting the overall architecture of LiLa, which is based on a collection of canonical citation forms of Latin words, the paper details the ontological modeling of each of the lexical and textual resources currently interlinked in the Knowledge Base. Finally, a few perspectives for the near future are sketched.

Keywords: Latin, Linguistic Resources, Linguistic Linked Data, Interoperability

1. Introduzione

Le risorse linguistiche hanno una lunga storia, benché il nome che ad esse si riferisce sia stato introdotto di recente, ovvero negli anni Novanta del Novecento da Antonio Zampolli. Dizionari, lessici, concordanze e raccolte di testi sono stati compilati per secoli, ma solo negli ultimi decenni essi sono stati prodotti (o trasferiti) su supporto elettronico, con il risultato che il loro numero, diffusione e accessibilità sono esplosi, coprendo un numero sempre crescente di lingue.

Nonostante la fase di sviluppo di nuove risorse sia ancora in corso, negli ultimi anni ad essa si sono accostate due nuove linee di lavoro che mirano a far fronte ad altrettante esigenze della comunità linguistica; esigenze sollevate proprio dal notevole aumento del numero (e delle tipologie) di risorse oggi disponibili.

La prima esigenza consiste nella necessità di avere a disposizione luoghi virtuali sul Web dove poter pubblicare, raccogliere, trovare e interrogare le risorse linguistiche. A questa esigenza rispondono oggi infrastrutture che fanno da collettori di risorse fornite dagli sviluppatori. Tra esse, la più importante è certa-

CHIMERA. Romance Corpora and Linguistic Studies 10 (2023), 45-72. ISSN 2386-2629

mente l'infrastruttura CLARIN[1], che ormai è entrata a far parte della quotidianità di tutti coloro che utilizzano e/o sviluppano raccolte di evidenza empirica di natura linguistica. Oltre all'accesso alle risorse, CLARIN offre un pacchetto di strumenti, per lo più di Trattamento Automatico del Linguaggio (TAL), per analizzarle, oltre che una serie di servizi per interrogarle.

La seconda esigenza nasce dalla constatazione che l'interazione tra i (meta)dati linguistici raccolti nelle risorse comporta un valore aggiunto al loro utilizzo. Se, infatti, un'infrastruttura come CLARIN raccoglie e pubblica risorse in un unico luogo, tuttavia ancora non ne consente un uso pienamente 'interoperabile', ovvero tale per cui l'informazione fornita da risorse diverse (per origine e tipo) è interrogabile e utilizzabile in modo congiunto. A tal proposito, nel corso degli ultimi anni, CLARIN ha iniziato a sviluppare metodi e applicazioni a supporto non solo della raccolta, ma anche dell'interoperabilità tra le risorse in essa pubblicate. Ad esempio, alcune risorse pubblicate in CLARIN sono direttamente processabili con strumenti di TAL attraverso servizi online come il Language Resource Switchboard (Zinn 2018) o Weblicht (Hinrichs *et al*. 2010). Inoltre, un primo grado d'interoperabilità tra le risorse di CLARIN è realizzabile a livello dei loro metadati descrittivi attraverso la cosiddetta Component MetaData Infrastructure (CMDI) (Broeder *et al.* 2022), che raccoglie 'componenti', ovvero gruppi di metadati semanticamente coerenti che vengono connessi a un registro di concetti condiviso, chiamato CLARIN Concept Registry (Schuurman *et al.* 2016). Tuttavia, tali concetti non sono (ancora) messi in relazione con quelli di altri schemi/ontologie e, soprattutto, non arrivano a consentire una rappresentazione dei dati più granulari, siano questi lessicali (come, ad esempio, le entrate lessicali dei dizionari) o testuali (le singole parole nei testi), e dunque una piena interoperabilità tra le risorse. Per far fronte a questa esigenza, è andata formandosi nel corso dell'ultimo decennio una comunità che opera ricerca nel campo dei cosiddetti Linguistic Linked Open Data (LLOD), ovvero l'applicazione ai (meta)dati linguistici dei principii del paradigma Linked Data, originariamente sviluppato a supporto del Semantic Web.

Proprio l'applicazione del paradigma Linked Data alle risorse linguistiche della lingua latina per fini di loro interoperabilità è il fondamento metodologico e strutturale del progetto "LiLa. Linking Latin"[2]. Finanziato dallo European Research Council (ERC), l'obiettivo di LiLa è di fare in modo che le molte risorse lessicali e testuali per il latino oggi disponibili in formato digitale possano interagire tra loro sulla Rete, avanzando così lo stato dell'arte della loro pubblica-

[1] http://clarin.eu.
[2] https://lila-erc.eu.

zione che, al momento, è realizzata nell'ambito di biblioteche digitali (come, ad esempio, Perseus[3]), o di singoli progetti editoriali ad accesso libero (come Computational Historical Semantics[4]), o proprietario (come la Library of Latin Texts di Brepols[5]). A tal fine, il progetto ha sviluppato una Knowledge Base in cui le risorse linguistiche sono interoperabili grazie a collegamenti tra le loro componenti resi possibili tramite il ricorso all'applicazione dei principii Linked Data e, nello specifico, attraverso la rappresentazione dei (meta)dati delle risorse con ontologie e vocabolari sviluppati e condivisi dalla comunità LLOD.

Questo articolo descrive lo stato attuale della Knowledge Base LiLa, con una specifica attenzione per la rappresentazione ontologica delle risorse al momento rese interoperabili[6]. Dopo una breve introduzione del paradigma Linked Data (Sezione 2), viene presentata l'architettura di LiLa e, in particolare, la raccolta di lemmi latini che ne costituisce la componente portante (Sezione 3). Quindi, le singole risorse attualmente incluse in LiLa sono descritte, dettagliando il modo in cui ciascuna di esse è modellizzata in termini ontologici (Sezione 4). Infine, l'articolo riporta alcune prospettive di lavoro e conclusioni (Sezione 5).

2. Il paradigma Linked Data

Introdotto da Tim Berners-Lee *et al.* (2001), il concetto di Semantic Web si fonda sull'assunto che i documenti pubblicati nel World Wide Web vengano associati a informazioni e metadati strutturati in modo tale da consentirne l'interrogazione e l'interpretazione semantica da parte non solo di esseri umani, ma anche di agenti automatizzati.

Tale strutturazione è realizzata in forma di Linked Data, che rappresentano le colonne portanti del Semantic Web, inteso come una rete di dati. Diversamente da un Web fatto di ipertesti, in cui i collegamenti non sono semanticamente interpretabili, il Semantic Web è costituto da link tra "oggetti" associati a un identificativo unico e persistente (URI: Uniform Resource Identifier). I collegamenti tra gli oggetti sono semanticamente interpretabili in quanto rappresentati attraverso vocabolari di descrizione della conoscenza (il più possibile condivisi) registrati in forma di ontologie.

[3] http://www.perseus.tufts.edu/hopper/.
[4] http://www.comphistsem.org.
[5] https://www.brepols.net/series/llt-o.
[6] Alcuni dei contenuti di questo articolo sono riportati (e, in parte, estesi) anche in Passarotti & Mambrini (2021).

Il paradigma Linked Data è fondato su quattro principii definiti da Berners-Lee stesso[7]:

1. usare URI come "nomi per le cose" ("names for things") al fine di identificarle in modo unico e persistente. Le "cose" con cui si ha a che fare se si trattano (meta)dati linguistici in Linked Data sono oggetti linguistici, come ad esempio occorrenze di parole in testi, entrate lessicali in dizionari, o insiemi di parti del discorso;
2. usare HTTP URI, per consentire alle persone (e alle macchine) di "cercare le cose" sul Web ("to look up things");
3. usare standard come RDF e SPARQL per fornire informazione utile su quanto è identificato da una URI, ai fini di rappresentazione e ricerca dei (meta)dati. RDF (Resource Description Framework) (Lassila & Swick, 1998) è il data model che sta alla base del Semantic Web. In base a esso, l'informazione nel Semantic Web è organizzata e rappresentata in termini di triple, ovvero relazioni tra un Soggetto e un Oggetto attraverso una Proprietà (ovvero, un Predicato diadico). Le classi cui appartengono i Soggetti e gli Oggetti, così come la semantica delle Proprietà, sono stabilite da ontologie condivise dalle diverse comunità che arricchiscono e utilizzano il Semantic Web. SPARQL (SPARQL Protocol And RDF Query Language)[8] è un linguaggio di interrogazione per (meta)dati rappresentati in RDF;
4. includere link ad altre URI, in modo da consentire alle persone (e alle macchine) di "scoprire più cose" ("to discover more things").

Applicare i principii del paradigma Linked Data a (meta)dati tratti da risorse linguistiche e pubblicarli sul Web comporta una serie di beneficii (Chiarcos *et al.* 2013). Innanzitutto, a livello di rappresentazione e modellizzazione dei (meta)dati, RDF è un modello molto versatile e, quindi, adatto per rappresentare metadati come, ad esempio, quelli veicolati dai vari livelli di annotazione disponibili nelle risorse linguistiche (morfologia, sintassi, lemmatizzazione etc.). Inoltre, proprio l'adozione di un data model comune (RDF) consente di mettere in atto sia un'interoperabilità di tipo strutturale (o sintattico), consistente nell'abilità di sistemi diversi di processare dati scambiati utilizzando protocolli e formati condivisi (come HTTP e URI), sia un'interoperabilità concettuale (o semantica), ovvero l'abilità di un sistema d'interpretare automaticamente e

[7] https://www.w3.org/DesignIssues/LinkedData.
[8] https://www.w3.org/TR/rdf-sparql-query/.

semanticamente l'informazione scambiata, utilizzando un insieme comune di classi e categorie dei dati definite in ontologie e vocabolari (Ide & Pustejovsky 2010). A ciò si aggiunga un alto grado di dinamicità dei (meta)dati: infatti, dal momento che chi fornisce i (meta)dati delle risorse linguistiche pubblicate in LLOD li può gestire e mantenere localmente sul proprio server, è possibile dare accesso sempre alla versione più recente della risorsa. Infine, il mondo LLOD è un ecosistema sviluppato e mantenuto da un'ampia e vivace comunità scientifica che adotta e produce strumenti e pratiche comuni; tra le iniziative in corso è particolarmente meritevole di menzione la COST Action Nexus Linguarum: European network for Web-centred linguistic data science[9].

3. La Knowledge Base LiLa

Questa sezione descrive la Knowledge Base LiLa, che consiste in una raccolta di risorse linguistiche (sia lessicali che testuali) per la lingua latina rese interoperabili in Linked Data sul Web tramite la loro rappresentazione attraverso comuni ontologie e vocabolari di descrizione della conoscenza.

3.1 L'architettura di LiLa

L'architettura di LiLa (Passarotti *et al.* 2020) si fonda sul semplice assunto che tutto ciò che fa parte di LiLa ha a che fare con le parole. La Figura 1 mostra, nella parte bassa, le fonti dei (meta)dati che LiLa rende interoperabili. Nello specifico, esse sono:

- le risorse lessicali, come i dizionari o i lessici, che descrivono proprietà di parole e sono costitute da entrate lessicali;
- le risorse testuali, come i corpora e le biblioteche digitali, che forniscono testi e includono occorrenze di parole in essi (tecnicamente nominate token);
- gli strumenti di TAL (in inglese: Natural Language Processing, NLP), che processano testi e producono risultati (NLP Output). In particolare, l'output di uno specifico tipo di strumento di TAL (i cosidetti tokenizzatori) sono token, che a propria volta, entrano in input ad altri strumenti di TAL, come ad esempio i marcatori delle parti del discorso (Part Of Speech Tagger).

[9] https://nexuslinguarum.eu.

Nella Figura 1 è possibile vedere come le entrate lessicali, le occorrenze delle parole nei testi e gli output degli strumenti di TAL vengano resi interoperabili in LiLa attraverso il loro collegamento ai rispettivi lemmi, ovvero le forme convenzionali di citazione delle parole.

Passando attraverso i lemmi, è dunque possibile operare ricerche distribuite sulle risorse linguistiche collegate e rese interoperabili in LiLa. Ad esempio, si possono cercare tutte le occorrenze (i token) del medesimo lemma in più corpora testuali; oppure si possono estrarre da più corpora tutte le occorrenze di parole che hanno certe proprietà fornite da una o più risorse lessicali.

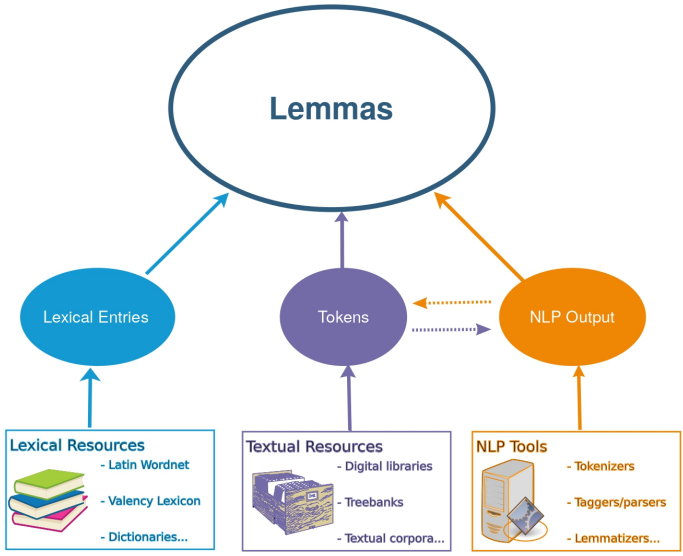

Figura 1. L'architettura di LiLa

Dato il ruolo centrale giocato dai lemmi nell'architettura di LiLa, una componente essenziale della Knowledge Base è una raccolta di forme convenzionali di citazione delle parole latine, chiamata Lemma Bank.

3.2 La Lemma Bank

La Lemma Bank di LiLa consiste in una raccolta di lemmi della lingua latina, ovvero forme di citazione lessicale adottate (più o meno convenzionalmente) nelle risorse linguistiche. Si tratta, cioè, dei nomi delle entrate nelle risorse lessi-

cali e delle forme scelte per raccogliere tutte le occorrenze di una medesima parola nelle risorse testuali. Come visto, la Lemma Bank ha un compito fondamentale nella Knowledge Base LiLa: attraverso essa vengono rese interoperabili le risorse linguistiche per il latino, rappresentando il punto di connessione tra le entrate delle diverse risorse lessicali e le occorrenze delle parole di quelle testuali.

Fondandosi sui principii del paradigma Linked Data, l'interoperabilità concettuale tra le risorse distribuite connesse in LiLa è realizzata attraverso l'applicazione di un vocabolario di descrizione della conoscenza condiviso non solo all'interno di LiLa ma, più ampiamente, nel mondo LLOD. Nello specifico della Lemma Bank, ciò consiste nel ricorso all'uso del vocabolario definito da OntoLex-Lemon (McCrae *et al.* 2017), una delle ontologie più adottate nel settore per fini di rappresentazione di risorse lessicali in Linked Data. La Figura 2 presenta il modello di OntoLex-Lemon.

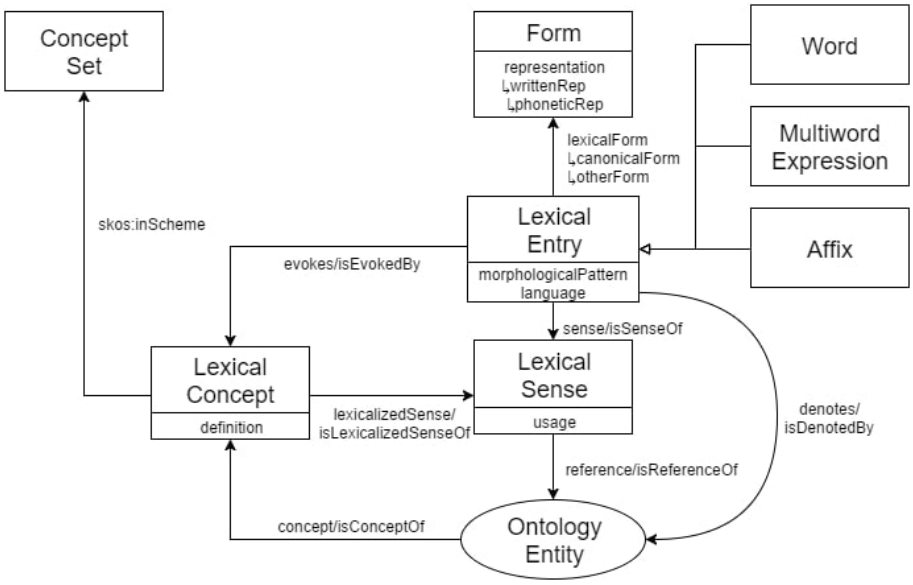

Figura 2. Il modello di OntoLex-Lemon

Nella Figura 2, le Classi di OntoLex-Lemon (ovvero, i "concetti" descritti dal modello) sono graficamente rappresentate entro rettangoli. Le relazioni tra le Classi sono frecce associate al nome della Proprietà (ovvero, il predicato) che collega tra loro due Classi.

La Classe principale di OntoLex-Lemon è la Lexical Entry, intesa come l'unità di analisi del lessico che raccoglie una o più forme (Lexical Form) e uno o più sensi (Lexical Sense) e concetti (Lexical Concept). I Lexical Sense sono sensi lessicalizzati, ovvero un senso appartiene esattamente a una Lexical Entry. Elementi semantici che possono essere espressi da più parole sono, invece, rappresentati attraverso i Lexical Concept, che dunque possono avere più di una lessicalizzazione. Un tipico esempio di Lexical Concept sono i synset di una risorsa come WordNet, che raggruppano più parole legate tra loro da un rapporto di sinonimia concettuale[10].

Le Lexical Form hanno una o più varianti grafiche (Written Representation) ed eventualmente fonetiche (Phonetic Representation). Una delle Lexical Form, Oggetto della Proprietà `canonicalForm`, è il lemma, cioè la forma che è scelta, più o meno convenzionalmente, per rappresentare l'intero insieme delle forme flesse di un'entrata lessicale. La Lemma Bank di LiLa è, dunque, una raccolta di Canonical Form di OntoLex-Lemon, svincolate da alcun rapporto con una Lexical Entry, in quanto la Lemma Bank non è una risorsa lessicale costituita da entrate lessicali, ma, appunto, un insieme di forme canoniche di citazione. Ciò rispecchia il ruolo della Lemma Bank in LiLa, che non è quello di una risorsa che connette risorse, ma di una raccolta di lemmi utilizzata per connettere tra loro risorse.

I lemmi che costituiscono la Lemma Bank sono stati tratti dalla base lessicale dell'analizzatore morfologico per il latino Lemlat (Passarotti *et al.* 2017), che consiste nella collazione di tre dizionari di latino classico (Georges & Georges 1913-1918; Glare 1982; Gradenwitz 1904), nell'intero Onomasticon del *Lexicon Totius Latinitatis* di Forcellini (1940) (Budassi & Passarotti 2016) e nel *Glossarium Mediae et Infimae Latinitatis* di du Cange *et al.* (1883-1887), per un totale di più di 130.000 parole, corrispondenti a circa 200.000 Canonical Form (Cecchini *et al.* 2018b)[11].

[10] WordNet (Fellbaum 2010) è una risorsa lessicale in cui nomi, verbi, aggettivi e avverbi sono raccolti in insiemi di sinonimi cognitivi (detti synset), ognuno dei quali esprime un concetto, descritto in una glossa. I synset sono connessi tra loro attraverso relazioni lessicali e semantico-concettuali, quali ad esempio, l'antinomia e l'ipo-/iperonimia. Per maggiori dettagli, si veda: https://wordnet.princeton.edu.

[11] Il numero delle parole registrate nella Lemma Bank è inferiore al numero delle Canonical Form, in quanto una parola può avere più di una Canonical Form (ad esempio, per la presenza di forme grafiche varianti). Per maggiori dettagli in proposito, si veda la Sezione 3.2 di Passarotti *et al*. (2020).

Le risorse testuali sono connesse alla Lemma Bank attraverso la Proprietà `hasLemma`[12], che collega un token in un corpus con il suo lemma nella Lemma Bank. Le risorse lessicali, invece, sono connesse alla Lemma Bank attraverso la Proprietà di OntoLex-Lemon `canonicalForm`[13], che collega una Lexical Entry della risorsa in questione con il corrispondente lemma, ovvero Canonical Form, nella Lemma Bank.

4. Le risorse linguistiche in LiLa

Questa sezione presenta le risorse lessicali e testuali per il latino attualmente rese interoperabili attraverso il loro allacciamento alla Knowledge Base LiLa. La sezione descrive il modo in cui ciascuna risorsa è modellizzata (ovvero rappresentata ontologicamente) e pubblicata come LLOD.

4.1 Risorse lessicali

Tutte le risorse lessicali allacciate a LiLa sono costituite di entrate lessicali, ovvero di istanze individuali della classe Lexical Entry di OntoLex-Lemon. Come detto, esse sono linkate alla Lemma Bank di LiLa attraverso la Proprietà `canonicalForm`. I dettagli della rappresentazione delle singole risorse lessicali sono riportati nelle sezioni che seguono.

4.1.1 *Word Formation Latin*

Word Formation Latin (WFL) (https://lila-erc.eu/data/lexicalResources/WFL/Lexicon) è un lessico latino le cui entrate (circa 30.000) sono messe in relazione tramite processi di formazione di parola (Litta *et al.* 2019).

WFL è modellizzato in LiLa tramite il ricorso a Classi e Proprietà (a) del modulo di OntoLex dedicato alla morfologia (morph) (Chiarcos *et al.* 2022)[14], (b) di un modello relativo alla rappresentazione di informazione lessicale concernente la variazione e la traduzione (vartrans) (Montiel-Ponsoda *et al.* 2015)[15]

[12] https://lila-erc.eu/ontologies/lila/hasLemma. Questa proprietà è definita nell'ontologia specifica di LiLa, in quanto OntoLex-Lemon non include alcuna proprietà atta a collegare una Lexical Form o una Lexical Entry con una sua occorrenza in un testo.
[13] http://www.w3.org/ns/lemon/ontolex#canonicalForm.
[14] https://www.w3.org/community/ontolex/wiki/Morphology.
[15] https://www.w3.org/ns/lemon/vartrans.

e (c) dell'ontologia sviluppata da LiLa specificamente per la rappresentazione di WFL[16].

La Figura 3 presenta la modellizzazione di due relazioni di derivazione connesse al verbo *deduco* in WFL. L'entrata lessicale di *deduco* in WFL, linkata al lemma *deduco* nella Lemma Bank di LiLa tramite la Proprietà `canonicalForm`, è l'output (vartrans:target) di una relazione di formazione di parola (istanza della Classe morph:wordFormationRelation), che connette *deduco* con il suo input (vartrans:source), ovvero il verbo *duco*. A propria volta, *deduco* è l'input di una relazione il cui output è l'aggettivo *deducticius*; tale relazione corrisponde a una regola di formazione di parola (di tipo suffissale)[17], che produce *deducticius* da *deduco* tramite il ricorso al suffisso *-ici* (Pellegrini *et al.* 2021).

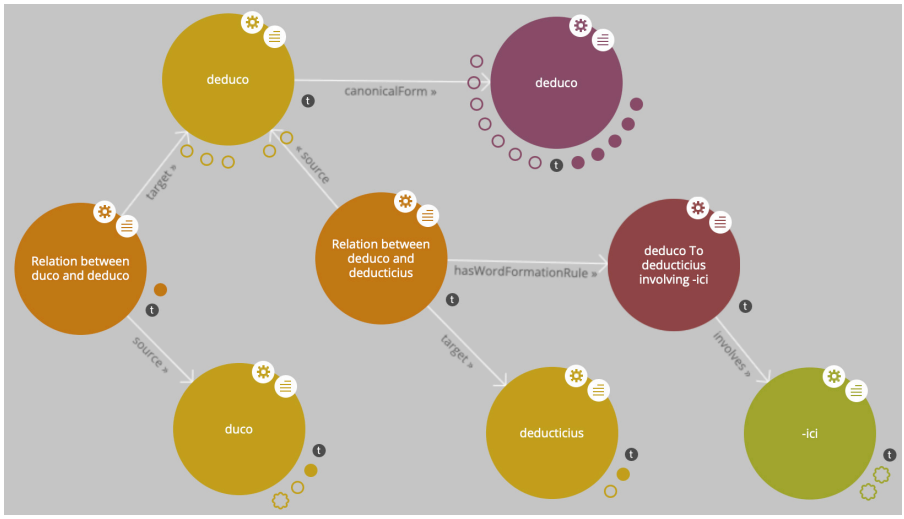

Figura 3. La modellizzazione di Word Formation Latin

4.1.2 *Dizionario etimologico*

L'*Etymological Dictionary of Latin & the Other Italic Languages* (https://lila-erc.eu/data/lexicalResources/BrillEDL/Lexicon) è un dizionario che include forme ricostruite proto-indoeuropee e proto-italiche per spiegare la storia etimologica di circa 1.400 forme latine (De Vaan 2008; Mambrini & Passarotti 2020).

[16] https://lila-erc.eu/lodview/ontologies/lila/wfl/

[17] https://lila-erc.eu/data/lexicalResources/WFL/rules/Derivation_Suffix_li_98210_To_li_98212

L'informazione etimologica registrata nel dizionario di De Vaan (2008) che è rappresentata in LiLa consiste nelle radici proto-indoeuropee e proto-italiche delle entrate lessicali che costituiscono la risorsa. Questa informazione è modellizzata tramite il ricorso all'estensione di OntoLex-Lemon lemonEty (Khan 2018)[18].

La Figura 4 mostra il modo in cui l'etimologia dell'aggettivo *mirus* riportata da De Vaan (2008) è rappresentata come LLOD in LiLa. Oltre al consueto link alla Lemma Bank, l'entrata lessicale di *mirus* è connessa a una etimologia (come definita da lemonEty[19]), che a propria volta è costituita da due etymon[20], ovvero le radici proto-indoeuropea (**smeiro-*) e proto-italica (**sméi-ro-*) di *mirus* riportate nel dizionario. L'estensione lemonEty consente, inoltre, di rappresentare la storia etimologica in termini di relazioni source-target tra i diversi etymon. Nella Figura 4 è visibile come, tramite un Etymology Link[21], la storia etimologica di *mirus* parta dalla radice proto-indoeuropea e passi attraverso la radice proto-italica per arrivare, infine, alla parola latina.

La medesima modellizzazione usata per il dizionario etimologico di De Vaan (2008) è adottata anche per rappresentare in LiLa l'*Index Graecorum Vocabulorum in Linguam Latinam Translatorum* (https://lila-erc.eu/data/lexicalResources/IGVLL/Lexicon), una lista di 1.763 prestiti latini dal greco antico pubblicata nel 1874 da Günther Alexander E.A. Saalfeld (Saalfeld 1884; Franzini *et al.* 2020). In questo caso, gli etymon consistono nelle parole greche da cui sono tratti i prestiti latini coperti dalla risorsa lessicale in questione.

[18] http://lari-datasets.ilc.cnr.it/lemonEty.
[19] http://lari-datasets.ilc.cnr.it/lemonEty#Etymology.
[20] http://lari-datasets.ilc.cnr.it/lemonEty#Etymon.
[21] http://lari-datasets.ilc.cnr.it/lemonEty#EtyLink.

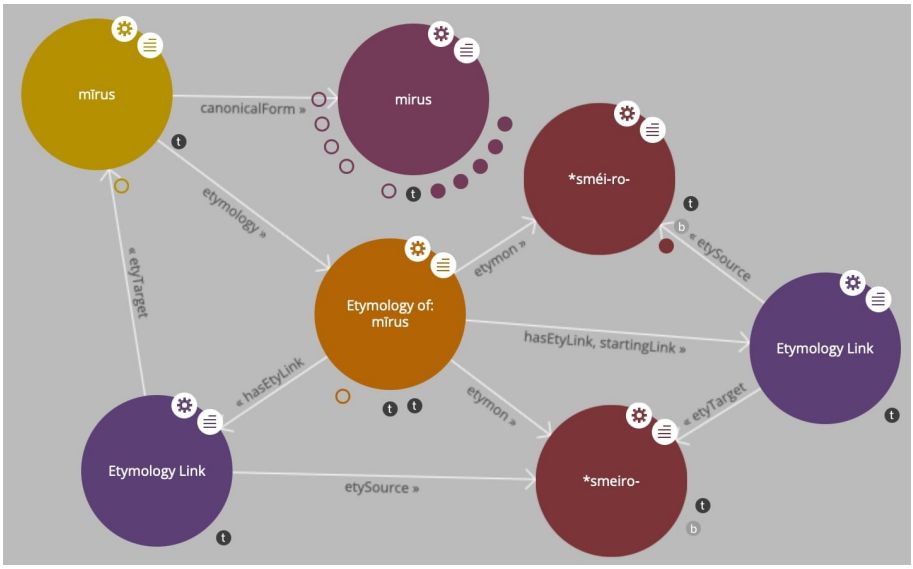

Figura 4. La modellizzazione del dizionario etimologico di De Vaan (2008)

4.1.3 *Latin Vallex e Latin WordNet*

Latin Vallex e Latin WordNet (https://lila-erc.eu/data/lexicalResources/LatinVallex/Lexicon; http://lila-erc.eu/data/lexicalResources/LatinWordNet/Lexicon) sono due risorse strettamente legate tra loro che raccolgono una porzione corretta manualmente di Latin WordNet, in cui a ogni senso di una parola, corrispondente a un synset di WordNet, è associato un frame valenziale (Mambrini *et al.* 2021).

La Figura 5 mostra con un esempio come l'informazione lessicale portata da Latin WordNet sia stata modellizzata in termini ontologici e, quindi, pubblicata come LLOD in LiLa. Nell'esempio della parola *genetrix* riportato in Figura 5, l'allacciamento alla Knowledge Base è garantito dal collegamento tra l'entrata lessicale del nome *genetrix* in Latin WordNet e il lemma *genetrix* (con variante grafica *genitrix*) nella Lemma Bank di LiLa. Quindi, applicando quanto stabilito da OntoLex-Lemon (McCrae *et al.* 2017), i synset di WordNet sono rappresentati come istanze della Classe Lexical Concept a cui le rispettive entrate lessicali sono connesse attraverso la Proprietà `evokes` (inversa: `isEvokedBy`). Ogni entrata lessicale di WordNet ha tanti Lexical Sense quanti sono i suoi sensi (Proprietà: `sense`; inversa: `isSenseOf`) e ogni Lexical Sense è collegato a un corrispondente synset tramite la Proprietà `isLexicalizedSenseOf` (inversa: `lexicalizedSense`).

Nella Figura 5, oltre a *genetrix*, sono visibili due entrate lessicali di Latin Wordnet che 'evocano' il medesimo Lexical Concept di *genetrix*, ovvero, in altre parole, appartengono al suo stesso synset: *institutor* e *plasmator*. Il synset evocato è quello identificato dal codice (di Princeton WordNet 3.1) 09614315-n, la cui glossa è "a person who grows or makes or invents things"[22].

Tramite questa modellizzazione e pubblicazione in LLOD di una risorsa come Latin WordNet, è possibile lanciare in LiLa una ricerca che trovi nei corpora linkati alla Knowledge Base le occorrenze di tutti i lemmi della Lemma Bank che, essendo connessi a una entrata lessicale di Latin WordNet, evocano uno specifico synset. Una query del genere dimostra il potenziale della interoperabilità tra risorse linguistiche, in quanto utilizza in contemporanea dati testuali tratti da più corpora e informazioni lessicali provenienti da una risorsa lessicale. Inoltre, dal momento che Latin WordNet adotta i medesimi synset di Princeton WordNet 3.1, la query si può estendere anche ad altre lingue oltre al latino.

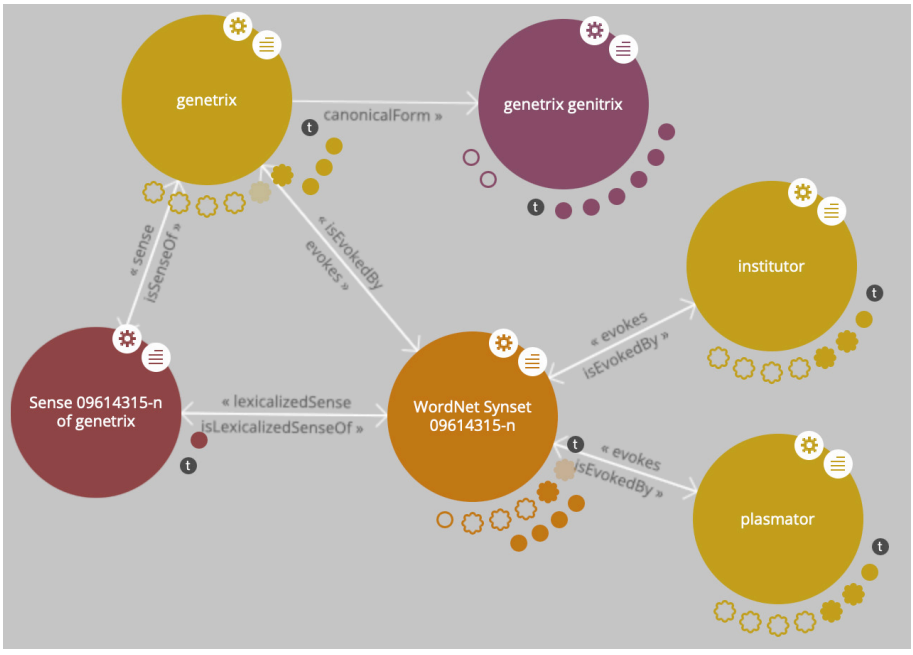

Figura 5. La modellizzazione di Latin WordNet

[22] http://wordnet-rdf.princeton.edu/pwn30/09614315-n.

La Figura 6 presenta un esempio di modellizzazione di un frame valenziale di Latin Vallex. La modellizzazione fa uso di componenti tratte dall'ontologia PreMOn (Rospocher *et al.* 2019), sviluppata per rappresentare classi e ruoli semantici, oltre che relazioni predicativo-argomentali[23]. Nella Figura 6, l'entrata lessicale di *amoveo*, condivisa tra Latin WordNet e Latin Vallex, evoca, oltre ai synset di WordNet cui appartiene, anche i frame valenziali ad essa assegnati dalla risorsa. Ciascun frame valenziale, che è una sottoclasse di Semantic Class di PreMOn[24], è connesso, attraverso la Proprietà di PreMOn `semRole`[25], ai propri ruoli semantici, che, nel caso in esempio, sono tre: Agente (ACT), Paziente (PAT) e Origine (ORIG). Le etichette dei ruoli semantici sono le medesime utilizzate dal lessico di valenza della lingua ceca VALLEX (Žabokrtský & Lopatková 2007) e adottate nel livello di annotazione tectogrammaticale della Prague Dependency Treebank[26].

Ciascun frame valenziale è, inoltre, connesso a uno dei synset di WordNet dell'entrata lessicale in questione attraverso un mapping (come stabilito da PreMOn[27]), istituendo così una corrispondenza tra un senso di una parola e il suo specifico frame valenziale.

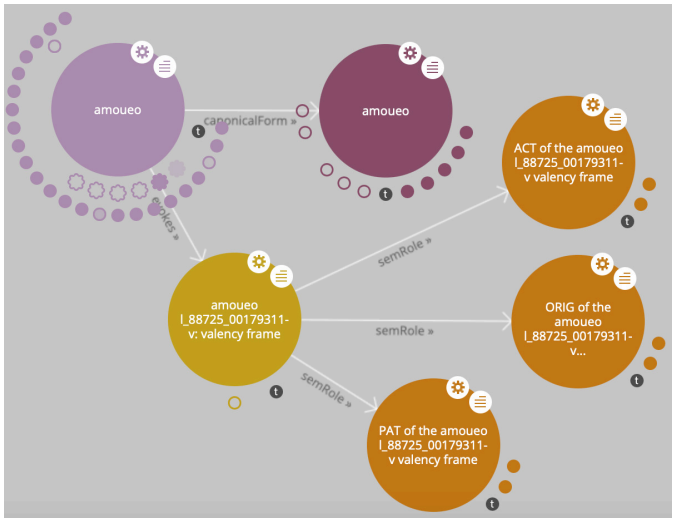

Figura 6. La modellizzazione di Latin Vallex

[23] https://premon.fbk.eu.

[24] https://lila-erc.eu/lodview/ontologies/latinVallex/ValencyFrame. http://premon.fbk.eu/ontology/core#SemanticClass.

[25] http://premon.fbk.eu/ontology/core#semRole.

[26] https://ufal.mff.cuni.cz/pdt2.0/doc/manuals/en/t-layer/html/index.html.

[27] http://premon.fbk.eu/ontology/core#Mapping.

4.1.4 *Latin Affectus*

Latin Affectus (https://lila-erc.eu/data/lexicalResources/LatinAffectus/Lexicon) è un lessico che assegna un valore di polarità di 'sentiment' propritario a più di 2.500 aggettivi e nomi latini (Sprugnoli *et al.* 2020). Tale valore propritario è identificato e assegnato alle entrate di Latin Affectus a livello lessicale, ovvero indipendentemente dai loro specifici contesti d'uso testuale.

Le entrate lessicali di Latin Affectus sono modellizzate attraverso l'ontologia Marl che è disegnata per annotare e descrivere opinioni soggettive (Buitelaar *et al*. 2013)[28]. La Figura 7 riporta l'entrata lessicale del nome *exitium* in Latin Affectus. Oltre al consueto link alla Lemma Bank tramite la Proprietà `canonicalForm`, l'entrata di *exitium* è connessa, attraverso la Proprietà `sense` di OntoLex-Lemon, al suo senso prioritario, che è un'istanza della Classe Lexical Sense. Tale senso prioritario è quindi il Soggetto di una tripla, che ha come Proprietà `hasPolarity`[29] e come Oggetto un'istanza della Classe Polarity[30] (che, in questo caso, ha polarità negativa[31]).

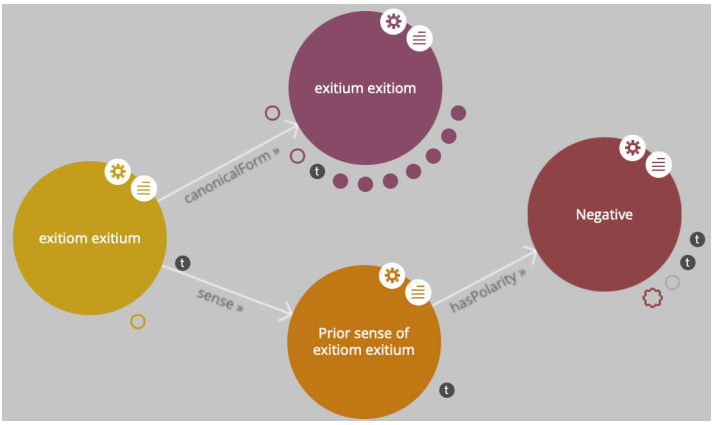

Figura 7. La modellizzazione di Latin Affectus

4.1.5 *Dizionario di Lewis & Short*

Il progetto LiLa ha incluso nella Knowledge Base il dizionario bilingue latino-inglese curato da Ch. T. Lewis e Ch. Short pubblicato nel 1879 (Lewis & Short 1879; Mambrini *et al.* 2021) (https://lila-

[28] http://www.gsi.upm.es:9080/ontologies/marl/.
[29] http://www.gsi.upm.es:9080/ontologies/marl/#hasPolarity.
[30] http://www.gsi.upm.es/ontologies/marl/ns#:Polarity.
[31] http://www.gsi.upm.es/ontologies/marl/ns#Negative.

erc.eu/data/lexicalResources/LewisShort/Lexicon). La modellizzazione del dizionario consiste in una rappresentazione dell'informazione lessicologico-semantica e lessicografica contenuta nelle entrate lessicali della risorsa. Il primo tipo di informazione è riprodotto ricorrendo a OntoLex-Lemon, mentre il secondo fa uso di una sua estensione, ovvero il modulo lessicografico lexicog (Bosque-Gil *et al*. 2017)[32].

La Figura 8 mostra la rappresentazione dell'entrata *blandus* del dizionario di Lewis & Short. Oltre al linking al corrispondente lemma nella Lemma Bank di LiLa, la Lexical Entry di *blandus* è connessa, tramite la Proprietà `entry` del vocabolario lime[33], a un'istanza della Classe Lexicon del medesimo vocabolario. Ciò rappresenta il fatto che il dizionario è un Lexicon e che *blandus* è una sua entrata.

La rappresentazione dell'informazione lessicologico-semantica contenuta nell'entrata di *blandus* consiste nel registrare le traduzioni in inglese dei singoli sensi della parola fornite dal dizionario di Lewis & Short come istanze della Classe Lexical Concept, cui corrisponde un senso lessicalizzato[34]. Nell'esempio, la Lexical Entry di *blandus* evoca, secondo il vocabolario di OntoLex-Lemon, il Lexical Concept consistente nella traduzione inglese "of a smooth tongue, flattering, fawning, caressing", che a propria volta ha una lessicalizzazione nello specifico senso di *blandus* identificato dal codice unico n5464.0 assegnato nel file XML da cui sono stati tratti i dati del dizionario[35].

La rappresentazione dell'informazione lessicografica dell'entrata di *blandus* consiste innanzitutto nella Lexicographic Entry[36] di *blandus*, che corrisponde alla parte dedicata a *blandus* nel dizionario modellizzato. Tale Lexicographic Entry descrive (secondo il vocabolario di lexicog) la Lexical Entry di *blandus* ed è costituita da uno o più Lexicographic Component[37], ciascuno dei quali descrive uno o più sensi della parola. In sostanza, i Component sono le diverse porzioni lessicografiche dell'entrata di *blandus* nel dizionario di Lewis & Short, mentre i Lexical Sense, che essi descrivono, corrispondono al contenuto lessicologico-semantico dei vari Component.

---

[32] https://www.w3.org/2019/09/lexicog/.
[33] http://www.w3.org/ns/lemon/lime#entry.
[34] Il modello vartrans non è stato qui utilizzato in quanto esso è mirato alla rappresentazione di traduzioni (connettendo tra loro Lexical Entry, o Lexical Sense), mentre nel caso in oggetto si tratta di rappresentare definizioni e liste di traduzioni.
[35] Il file XML del dizionario di Lewis & Short è disponibile presso la Perseus Digital Library sotto licenza CC BY SA 4.0: https://github.com/PerseusDL/lexica.
[36] http://www.w3.org/ns/lemon/lexicog#Entry.
[37] http://www.w3.org/ns/lemon/lexicog#LexicographicComponent.

Infine, la Lexicographic Entry di *blandus* è una entrata (entry) di quella specifica Lexicographic Resource[38] che è il dizionario di Lewis & Short.

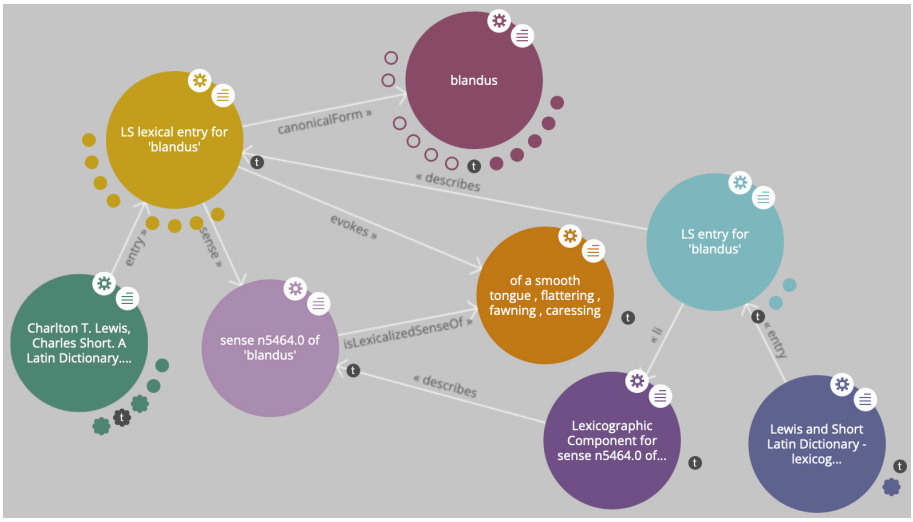

Figura 8. La modellizzazione del dizionario di Lewis & Short

4.2 Risorse testuali

Le risorse testuali attualmente allacciate a LiLa sono le seguenti:

- Index Thomisticus Treebank (https://lila-erc.eu/data/corpora/ITTB/id/corpus) (Passarotti 2019): il più esteso corpus annotato sintatticamente disponibile per la lingua latina. Raccoglie più di 350.000 occorrenze di parola tratte da testi di Tommaso d'Aquino (tra cui l'intera *Summa contra Gentiles*). La treebank è disponibile in due versioni: quella annotata secondo i criteri originali della risorsa (Bamman *et al.* 2008) e la sua conversione nello schema Universal Dependencies (Cecchini *et al.* 2018a);
- UDante (https://lila-erc.eu/data/corpora/UDante/id/corpus): corpus che raccoglie le opere latine di Dante Alighieri (circa 50.000 parole) arricchite con annotazione sintattica secondo lo schema Universal Dependencies (Cecchini *et al.* 2020);
- Querolus sive Aulularia (https://lila-erc.eu/data/corpora/Querolus/id/citationUnit/QuerolussiveAulularia): il testo

[38] http://www.w3.org/ns/lemon/lexicog#LexicographicResource.

di una commedia anonima della tarda antichità latina (circa 17.000 parole) (Gamba 2020);
- Liber Abbaci (https://lila-erc.eu/data/corpora/CorpusFibonacci/id/corpus/Liber%20Abbaci): un trattato di aritmetica scritto nel 1202 da Leonardo Fibonacci (VIII capitolo: circa 30.000 parole) (Grotto *et al.* 2021);
- LASLA (https://lila-erc.eu/data/corpora/Lasla/id/corpus): un corpus che raccoglie più di 130 testi di epoca classica e tarda, per un totale di circa 1.700.000 parole (Verkerk *et al.* 2020);
- Computational Historical Semantics (https://lila-erc.eu/lodview/data/corpora/CompHistSem/id/corpus): un corpus che include circa 4.000 testi latini scritti tra il II e il XV secolo. Al momento, è connessa a LiLa una sezione del corpus comprensiva di un totale di circa un milione di parole (Mehler *et al.* 2020);
- Confessiones (https://lila-erc.eu/data/corpora/CIRCSELatinLibrary/id/corpus/Confessiones): i 13 libri del testo digitale delle *Confessiones* di Agostino, tratto da The Latin Library[39], consistente in approssimativamente 92.000 parole.

In LiLa, ogni risorsa testuale è un oggetto di tipo corpus dell'ontologia POWLA[40] (Chiarcos 2012). Ogni corpus include uno, o più documenti (powla:Document)[41], che in LiLa sono i testi presenti in esso. Ogni documento è organizzato in tre livelli (*layer*), ovvero tre modi di arrivare dal documento a ciascuna occorrenza di parola (*token*) in esso:

- Document Layer[42]: raccoglie tutti i token del testo, senza ulteriori livelli intermedi;
- Sentence Layer: raccoglie le frasi del testo;
- Citation Layer: raccoglie i livelli di citazione del testo[43].

Ad esempio, la Figura 9 mostra come il corpus LASLA (nodo "Lasla Corpus") abbia tra i propri documenti le *Satire* di Orazio (nodo "Hor Sermones"), a cui

[39] http://www.thelatinlibrary.com/august.html.
[40] http://purl.org/powla/powla.owl#Corpus.
[41] http://purl.org/powla/powla.owl#Document.
[42] http://purl.org/powla/powla.owl#DocumentLayer.
[43] Il Sentence e il Citation Layer sono oggetti di tipo Citation Structure (https://lila-erc.eu/ontologies/lila_corpora/CitationStructure), definito nell'ontologia di LiLa come una sottoclasse della Classe DocumentLayer di POWLA.

sono associati i tre layer menzionati e, attraverso la Proprietà di Dublin Core `creator`[44], il nome dell'autore dell'opera.

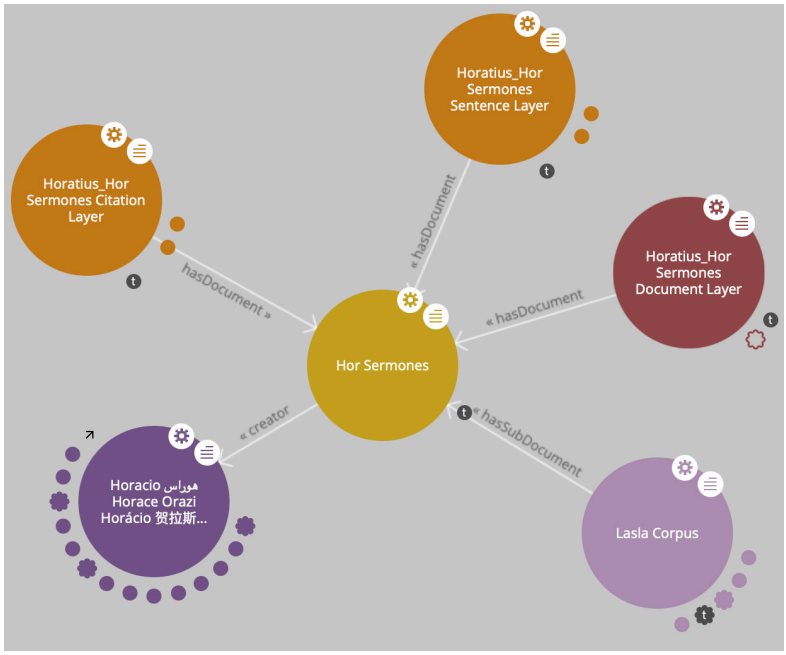

Figura 9. I layer di un documento del corpus LASLA

La Figura 10 presenta due token tratti dalle *Satire* di Orazio con il proprio Citation Layer espanso e, per uno di essi, il link alla Lemma Bank. Nella Figura è visibile il percorso di citazione dei token *demens* e *genetricem*: entrambi compaiono al verso 133 ("Versus 133"), della terza satira ("Poema 3") del secondo libro ("Liber 2") delle *Satire* di Orazio. La Proprietà `hasCitSubUnit`[45], definita da lilaCorpora così come `isLayer`[46], è una sotto-Proprietà di `hasChild` di POWLA[47], che connette l'ultimo livello di citazione (il verso) ai singoli token, che sono istanze della classe Terminal di POWLA[48]. Infine, i token sono connessi alla LemmaBank tramite la Proprietà di LiLa `hasLemma`.

[44] https://www.dublincore.org/specifications/dublin-core/dcmi-terms/#creator.
[45] https://lila-erc.eu/lodview/ontologies/lila_corpora/hasCitSubUnit.
[46] https://lila-erc.eu/lodview/ontologies/lila_corpora/isLayer
[47] http://purl.org/powla/powla.owl#hasChild.
[48] http://purl.org/powla/powla.owl#Terminal.

Figura 10. Il Citation Layer di due token delle *Satire* di Orazio

Le parole di due risorse testuali connesse a LiLa, la Index Thomisticus Treebank e UDante, sono associate anche (1) a un Morphology Layer, che raccoglie i tratti morfologici assegnati ai singoli token e (2) a un Annotation Layer, che registra il livello di annotazione sintattica e sostituisce il Sentence Layer, in quanto include la divisione del testo in frasi. La Figura 11 presenta l'organizzazione di questi layer a partire da due token della Index Thomisticus Treebank. I token d'esempio sono *gradatim* e *ascendens*, tratti dalla seguente frase della *Summa contra Gentiles* (libro 4, capitolo 1, numero 3): «ut scilicet, quia omnes rerum perfectiones quodam ordine a summo rerum vertice Deo descendunt, ipse, ab inferioribus incipiens et gradatim ascendens, in Dei cognitionem proficiat;».

La sequenza testuale, secondo cui *gradatim* precede immediatamente *ascendens*, è registrata attraverso una relazione tra i due token marcata con la Proprietà `next` (inversa: `previous`) di POWLA[49]. Attraverso la Proprietà `hasTerminal` di POWLA[50] è rappresentato il fatto che il token *ascendens* appartiene alla frase 17.726 dell'UD Annotation Layer dell'Index Thomisticus Treebank (`isLayer`). Tale layer raccoglie l'annotazione sintattica del testo della *Summa contra Gentiles* secondo lo schema di Universal Dependencies (UD). Come è visibile nella Figura 11, il documento nominato *Summa contra Gentiles* è con-

[49] http://purl.org/powla/powla.owl#next. http://purl.org/powla/powla.owl#previous.
[50] http://purl.org/powla/powla.owl#hasTerminal.

nesso anche a un altro Annotation Layer (attraverso la Proprietà di POWLA `hasDocument`[51]): questo layer raccoglie l'annotazione sintattica del testo secondo i criteri originali della risorsa. In tal modo vengono, cioè, rappresentate le due diverse versioni dell'annotazione sintattica della Index Thomisticus Treebank.

La Figura 11 mostra la relazione sintattica che intercorre tra *ascendens* e *gradatim* nell'albero a dipendenze della Index Thomisticus Treebank relativo alla frase in cui occorrono questi due token: in tale albero, la parola *gradatim* dipende da *ascendens* e la loro dipendenza è marcata con la relazione sintattica di UD che indica i modificatori avverbiali (advmod). In LiLa, la dipendenza di *gradatim* da *ascendens* è registrata tramite la rappresentazione della relazione advmod tra i due token come un'istanza[52] della Classe delle relazioni di dipendenza definita da lilaCorpora[53]. Tale istanza connette il token testa della relazione (*ascendens*) con il token dipendente (*gradatim*), rispettivamente tramite le Proprietà `hasHead`[54] e `hasDep`[55] di LiLaCorpora.

La Figura 11 presenta altresì l'annotazione del tratto morfologico del Numero del token *ascendens* nel cosiddetto Morphology Layer. I tratti morfologici sono rappresentati ricorrendo a Classi e Proprietà della Web Annotation Ontology[56]. Nel caso di *ascendens*, il token *ascendens* è l'Oggetto di una tripla il cui Soggetto è un'istanza della Classe Annotation[57], che rappresenta i tratti morfologici di *ascendens* secondo l'insieme delle etichette di UD, mentre la Proprietà che lega il Soggetto all'Oggetto è `hasTarget`[58]. Nella Figura 11, l'istanza di Classe Annotation è il Soggetto di altre due triple. La prima tripla la connette al Numero Singolare tramite la Proprietà `hasBody`[59]; la seconda la connette al Morphology Layer tramite la proprietà `hasLayer` di POWLA. Il Numero Sin-

[51] http://purl.org/powla/powla.owl#hasDocument.
[52] https://lila-erc.eu/lodview/data/corpora/ITTB/depAnnotation/UD/005.SCG*LB4.CP--++1.N.-3.7-2.11-1W23.
[53] https://lila-erc.eu/lodview/ontologies/lila_corpora/DependencyRel. La Classe DependencyRel di lilaCorpora è una sottoclasse di powla:Relation (http://purl.org/powla/powla.owl#Relation) ed è stata introdotta per specificare il tipo di relazione da rappresentare, ovvero una relazione di tipo sintattico registrata come annotazione in un corpus.
[54] https://lila-erc.eu/lodview/ontologies/lila_corpora/hasHead.
[55] https://lila-erc.eu/lodview/ontologies/lila_corpora/hasDep.
[56] https://www.w3.org/ns/oa.
[57] https://www.w3.org/ns/oa#Annotation.
[58] https://www.w3.org/ns/oa#hasTarget.
[59] https://www.w3.org/ns/oa#hasBody.

golare è un'istanza della Classe del tratto Numero Singolare di UD[60] e rappresenta, in particolare, la versione per il latino di questo tratto.

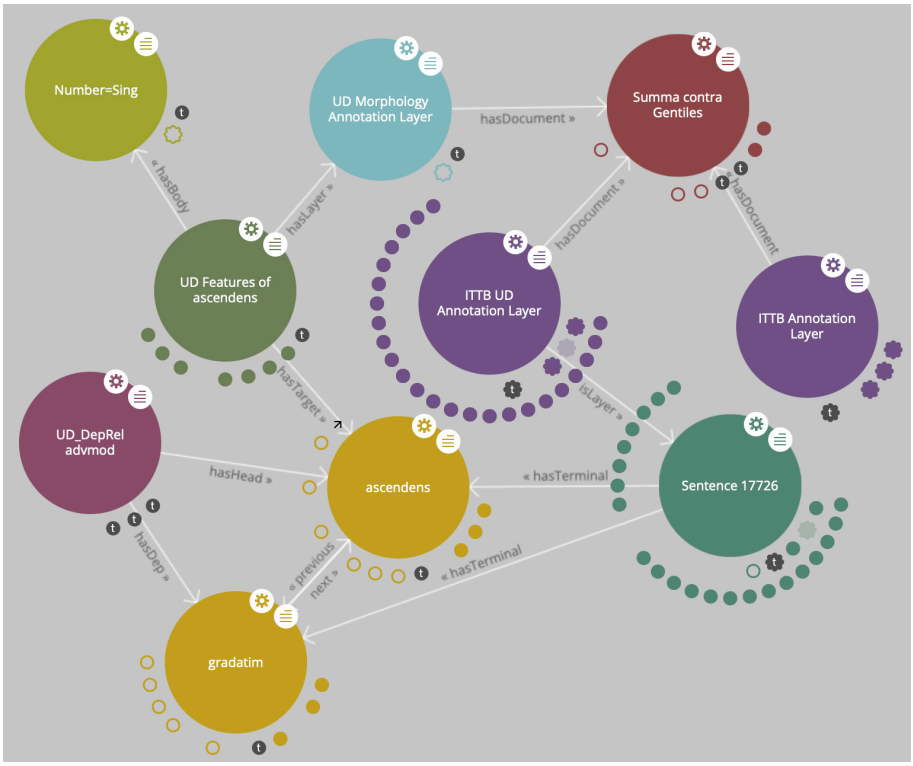

Figura 11. Il Morphology Layer e l'Annotation layer di due token della Index Thomisticus Treebank

5. Prospettive e conclusioni

Il progetto che ha permesso la realizzazione della Knowledge Base LiLa terminerà a giugno 2023; tuttavia, coerentemente con l'approccio LLOD, LiLa ha una natura "aperta". Ciò significa che, oltre al linking delle singole risorse, il risultato più importante del progetto, ovvero la Lemma Bank, che consente l'interoperabilità tra le risorse linguistiche della lingua latina, è destinato a rimanere attivo e disponibile anche dopo la conclusione del progetto stesso. La Lemma Bank, infatti, consiste di forme canoniche di citazione di parole latine

[60] https://universaldependencies.org/u/feat/Number#Sing.

cui è stato assegnato un identificativo unico e persistente e che sono state pubblicate sul Web. Proprio questi due aspetti rendono la Knowledge Base LiLa aperta a nuove inclusioni e connessioni: chiunque, infatti, può produrre triple che puntino ai lemmi della Lemma Bank di LiLa, al fine di fare interagire risorse linguistiche per il latino che il progetto Lila non ha trattato.

In tal senso, una delle sfide che l'ultima fase di LiLa e, quindi, il suo mantenimento successivo devono affrontare consiste nel rendere la Knowledge Base uno dei luoghi di pubblicazione (se non, il luogo di pubblicazione per eccellenza) delle risorse linguistiche latine. Questo obiettivo passa per almeno tre colli di bottiglia, che rappresentano ulteriori questioni aperte per il futuro più prossimo di LiLa. Il primo riguarda l'inclusione in LiLa di risorse proprietarie, che porterebbe mutui vantaggi sia alla Knowledge Base (espandendone la dimensione) sia alle case editrici che pubblicano i dati (rendendoli interoperabili con quelli di altre risorse). Il secondo è di ordine pedagogico: è necessario formare una nuova generazione di classicisti, e più ampiamente di umanisti, che abbia competenze di risorse linguistiche e, nello specifico, di LLOD. Il terzo, in vero strettamente legato al precedente, consiste nelle auspicabili conseguenze positive che la disponibilità della Knowledge Base LiLa può comportare per una delle sue comunità di riferimento: appunto, quella dei classicisti. Da sempre (e per necessità) abituati a ricorrere all'evidenza empirica testuale nei propri studi, i classicisti rappresentano una comunità che trarrebbe grande giovamento dalla possibilità di avere a disposizione un ambiente Web dove interrogare simultaneamente più corpora, lessici e dizionari; tuttavia, proprio anche a causa dell'ancora limitata offerta di formazione in ambito linguistico-computazionale nei curricula delle nuove leve dei classicisti, l'accesso e la conoscenza stessa delle risorse digitali disponibili e della possibilità di farle interagire sono ancora lasciati all'intraprendenza del singolo, se non addirittura al caso.

Anche al fine di colmare questa mancanza, il progetto LiLa sta sviluppando una interfaccia d'interrogazione delle risorse rese interoperabili dalla Knowledge Base. Tale interfaccia consentirà agli utenti di comporre e operare query sui (meta)dati delle risorse di LiLa in un ambiente grafico che, sulla base delle selezioni degli utenti, comporrà automaticamente le corrispondenti query in linguaggio SPARQL, evitando così che la conoscenza di SPARQL sia condizione necessaria all'interrogazione della Knowledge Base.

Guardando oltre ai confini delle risorse linguistiche per il latino, è auspicabile che la semplice, ma solida ed efficace architettura di LiLa, basata sul ruolo della Lemma Bank, possa essere adottata a supporto dell'interoperabilità tra risorse anche di altre lingue. A tal proposito, è in fase di avvio un progetto esplorativo (finanziato da CLARIN e diretto da Francesco Mambrini) che mira a in-

dagare il grado di applicabilità del 'modello LiLa' ad altre lingue, utilizzando come base empirica i (meta)dati delle risorse linguistiche fornite da CLARIN, con l'obiettivo di medio termine di realizzare nell'infrastruttura una *resource family* dedicata proprio all'interoperabilità in LLOD tra le risorse e, a lungo termine, con l'aspirazione di fare in modo che le risorse raccolte in CLARIN interagiscano secondo i principii del paradigma Linked Data a livello altamente granulare e, auspicabilmente, multilinguistico.

Ringraziamenti

Il progetto LiLa: Linking Latin è stato finanziato dallo European Research Council (ERC) nell'ambito del programma di ricerca e innovazione European Union's Horizon 2020 – Grant Agreement No. 769994.

Bibliografia

Bamman, D., Passarotti, M., 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 N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis & D. Tapias (eds), *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC 2008)*. Marrakech, Morocco: European Language Resources Association (ELRA), 71-76.

Berners-Lee, T., Hendler, J. & Lassila, O. 2011. The Semantic Web. *Scientific American* 284(5): 34-43.

Bosque-Gil, J., Gracia, J. & Montiel-Ponsoda, E. 2017. Towards a Module for Lexicography in OntoLex. In J.P. McCrae, F. Bond, P. Buitelaar, P. Cimiano, T. Declerck, J. Gracia, I. Kernerman, E. Montiel-Ponsoda, N. Ordan & M. Piasecki (eds), *Proceedings of the LDK 2017 Workshops: 1st Workshop on the OntoLex Model (OntoLex-2017), Shared Task on Translation Inference Across Dictionaries & Challenges for Wordnets co-located with 1st Conference on Language, Data and Knowledge (LDK 2017).* Galway, Ireland: CEUR, 74-84.

Broeder, D., Windhouwer, M., Van Uytvanck, D., Goosen, T. & Trippel, T. 2022. CMDI: a component metadata infrastructure. In V. Arranz, D. Broeder, B. Gaiffe, M. Gavrilidou, M. Monachini & T. Trippel (eds), *Proceedings of the workshop "Describing language resources with metadata: towards flexibility and interoperability in the documentation of language resources"*. Istanbul, Turkey: European Language Resources Association (ELRA), 1-4.

Budassi, M. & Passarotti, M. 2016. Nomen Omen. Enhancing the Latin Morphological Analyser Lemlat with an Onomasticon. In N. Reiter, B. Alex & K.A. Zervanou (eds), *Proceedings of the 10th SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH 2016)*. Berlin, Germany: The Association for Computational Linguistics, 90-94.

Buitelaar, P., Arcan, M., Iglesias, C.A., Sánchez-Rada, J.F. & Strapparava, C. 2013. Linguistic linked data for sentiment analysis. In C. Chiarcos, P. Cimiano, T. Declerck & J.P. McCrae (eds), *Proceedings of the 2nd Workshop on Linked Data in Linguistics (LDL-2013): Representing and linking lexicons, terminologies and other language data*. Pisa, Italy: Association for Computational Linguistics, 1-8.

Cecchini, F.M., Passarotti, M., Marongiu, P. & Zeman, D. 2018a. Challenges in Converting the Index Thomisticus Treebank into Universal Dependencies. In M.C. De Marneffe, T. Lynn & S. Schuster (eds), *Proceedings of the Second Workshop on Universal Dependencies (UDW 2018)*. Bruxelles, Belgium: The Association for Computational Linguistics, 27-36.

Cecchini, F.M., Passarotti, M., Testori, M., Ruffolo, P., Draetta, L., Fieromonte, M., Liano, A., Marini, C. & Piantanida, G. 2018b. Enhancing the Latin Morphological Analyser LEMLAT with a Medieval Latin Glossary. In E. Cabrio, A. Mazzei & F. Tamburini (eds), *Proceedings of the Fifth Italian Conference on Computational Linguistics*. Torino, Italy: aAccademia university press, 87-92.

Cecchini, F.M., Sprugnoli, R., Moretti, G. & Passarotti, M. 2020. UDante: First Steps Towards the Universal Dependencies Treebank of Dante's Latin Works. In J. Monti, F. Dell'Orletta & F. Tamburini (eds), *Proceedings of the Seventh Italian Conference on Computational Linguistics*. Bologna, Italy: CEUR Workshop Proceedings, 1-7.

Chiarcos, C. 2012. Interoperability of corpora and annotations. In C. Chiarcos, S. Nordhoff & S. Hellmann (eds), *Linked Data in Linguistics*. Berlin-Heidelberg: Springer, 161-179.

Chiarcos, C., Moran, S., Mendes, P.N., Nordhoff, S. & Littauer, R. 2013. Building a Linked Open Data Cloud of Linguistic Resources: Motivations and Developments. In I. Gurevych & J. Kim (eds), *The People's Web Meets NLP*. Berlin-Heidelberg: Springer-Verlag, 315-348.

Chiarcos, C., Gkirtzou, K., Khan, A.F., Labropoulou, P., Passarotti, M. & Pellegrini, M. 2022. Computational Morphology with OntoLex-Morph. In T. Declerck, J.P. McCrae, E. Montiel, C. Chiarcos & M. Ionov (eds), *Proceedings of the 8th Workshop on Linked Data in Linguistics: Revisiting a Decade of Linguistic Linked Open Data*. Marseille, France: European Language Resources Association (ELRA), 78-86.

De Vaan, M. 2008. *Etymological dictionary of Latin and the other Italic languages*. Leiden-Boston: Brill.

du Cange, C., Bénédictins de Saint-Maur, Carpentier, P., Henschel, L. & Favre, L. 1883-1887. *Glossarium Mediae et Infimae Latinitatis*. Niort: L. Favre.

Fellbaum, C. 2010. WordNet. In R. Poli, M. Healy & A. Kameas (eds), *Theory and Applications of Ontology: Computer Applications*. Dordrecht: Springer, 231-243.

Forcellini, E. 1940, *Lexicon Totius Latinitatis / ad Aeg. Forcellini lucubratum, dein a Jos. Furlanetto emendatum et auctum; nunc demum Fr. Corradini et Jos. Perin curantibus emendatius et auctius meloremque in formam redactum adjecto altera quasi parte Onomastico totius latinitatis opera et studio ejusdem Jos. Perin*. Padova: Typis Seminarii.

Franzini, G., Zampedri, F., Passarotti, M., Mambrini, F. & Moretti, G. 2020, Græcissare: Ancient Greek Loanwords in the LiLa Knowledge Base of Linguistic Resources for Latin. In J. Monti, F. Dell'Orletta & F. Tamburini (eds), *Proceedings of the Seventh Italian Conference on Computational Linguistics*. Bologna, Italy: CEUR Workshop Proceedings, 1-6.

Gamba, F. 2020. *Including a New Textual Resource into the LiLa Knowledge Base. Lemmatization, PoS Tagging and Linking of Querolus*. MA diss., Università di Pavia.

Georges, K.E. & Georges, H. 1913-1918. *Ausführliches Lateinisch-Deutsches Handwörterbuch*. Hannover: Hahn.

Glare, P.G.W. 1982. *Oxford Latin Dictionary*. Oxford: Oxford University Press.

Gradenwitz, O. 1904. *Laterculi Vocum Latinarum*. Leipzig: Hirzel.

Grotto, F., Sprugnoli, R., Fantoli, M., Simi, M., Cecchini, F.M. & Passarotti, M. 2021. The Annotation of Liber Abbaci, a Domain-Specific Latin Resource. In E. Fersini, M. Passarotti & V. Patti (eds), *Proceedings of the Eighth Italian Conference on Computational Linguistics*. Milan, Italy: CEUR Workshop Proceedings, 2021, 1-8.

Hinrichs, M., Zastrow, T. & Hinrichs, E.W. 2010. WebLicht: Web-based LRT Services in a Distributed eScience Infrastructure. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC 2010)*. La Valletta, Malta: European Language Resources Association (ELRA), 489-493.

Ide, N. & Pustejovsky, J. 2010. What does interoperability mean, anyway? toward an operational definition of interoperability for language technology. In *Proceedings of the Second International Conference on Global Interoperability for Language Resources*. Hong Kong, China.

Khan, A.F. 2018. Towards the Representation of Etymological Data on the Semantic Web. *Information* 9(12): 304.

Lassila, O., Swick, R.R. & World Wide and Web Consortium. 1998. *Resource description framework (RDF) model and syntax specification*. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.44.6030 (accessed March 3rd, 2023).

Lewis, C.T. & Short, C. 1879. *A Latin Dictionary. Founded on Andrews' edition of Freund's Latin dictionary*. Oxford: Clarendon Press.

Litta, E., Passarotti, M. & Mambrini, F. 2019. The treatment of word formation in the LiLa knowledge base of linguistic resources for Latin. In Z. Žabokrtský, M. Ševčíková, E. Litta & M. Passarotti (eds), *Proceedings of the Second International Workshop on Resources and Tools for Derivational Morphology (DeriMo 2019)*. Prague, Czech Republic: Charles University, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics, 35-43.

Mambrini, F. & Passarotti, M. 2020. Representing etymology in the LiLa knowledge base of linguistic resources for Latin. In I. Kernerman, S. Krek, J.P. McCrae, J. Gracia, S. Ahmadi & B. Kabashi (eds), *Proceedings of the Globalex Workshop on Linked Lexicography. LREC 2020 Workshop*. Paris, France: European Language Resources Association (ELRA), 20-28.

Mambrini, F., Passarotti, M., Litta, E. & Moretti, G. 2021. Interlinking Valency Frames and WordNet Synsets in the LiLa Knowledge Base of Linguistic Resources for Latin. In M. Alam, P. Groth, V. de Boer, T. Pellegrini, H.J. Pandit, E. Montiel, V. Rodríguez Doncel, B. McGillivray & A. Meroño-Peñuela (eds), *Further with Knowledge Graphs. Proceedings of the 17th International Conference on Semantic Systems*. Series: *Studies on the Semantic Web,* Volume 53. Amsterdam, The Netherlands: IOS Press, 16-28.

Mambrini, F., Litta, E., Passarotti, M. & Ruffolo, P. 2021. Linking the Lewis & Short Dictionary to the LiLa Knowledge Base of Interoperable Linguistic Resources for Latin. In E. Fersini, M. Passarotti & V. Patti (eds), *Proceedings of the Eighth Italian*

Conference on Computational Linguistics (CLiC-it 2021). Milan, Italy: CEUR Workshop Proceedings, 1-7.

McCrae, J.P., Bosque-Gil, J., Gracia, J., Buitelaar, P. & Cimiano, P. 2017. The Ontolex-Lemon model: development and applications. In I. Kosem, C. Tiberius, M. Jakubíček, J. Kallas, S. Krek & V. Baisa (eds), *Proceedings of eLex 2017 conference*. Brno, Czech Republic: Lexical Computing, 19-21.

Mehler, A., Jussen, B., Geelhaar, T., Hnlein, A., Abrami, G., Baumartz, D., Uslu, T. & Hemati, W. 2020. The Frankfurt Latin Lexicon: From morphological expansion and word embeddings to semiographs. *Studi e saggi linguistici* LVIII(1): 121-155.

Montiel-Ponsoda, E., Bosque-Gil, J., Gracia, J., de Cea, G.A. & Vila-Suero, D. 2015. Towards the Integration of Multilingual Terminologies: an Example of a Linked Data Prototype. In *Proceedings of the conference Terminology and Artificial Intelligence 2015*. Granada, Spain: CEUR, 205-206.

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

Passarotti, M., Budassi, M., Litta, E. & Ruffolo, P. 2017. The Lemlat 3.0 Package for Morphological Analysis of Latin. In G. Bouma & Y. Adesam (eds), *Proceedings of the NoDaLiDa 2017 Workshop on Processing Historical Language*. Gothenburg: Linköping University Electronic Press, 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. *Studi e Saggi Linguistici* LVIII(1): 177-212.

Passarotti, M. & Mambrini, F. 2021. Linking Latin: Interoperable Lexical Resources in the LiLa Project. In E. Biagetti, C. Zanchi & S. Luraghi (eds), *Building new resources for historical linguistics*. Pavia, Italy: Pavia University Press, 103-124.

Pellegrini, M., Litta, E., Passarotti, M., Mambrini, F. & Moretti, G. 2021. The Two Approaches to Word Formation in the LiLa Knowledge Base of Latin Resources. In F. Namer, N. Hathout, S. Lignon, M. Ševčíková & Z. Žabokrtský (eds), *Proceedings of the Third International Workshop on Resources and Tools for Derivational Morphology (DeriMo 2021)*. Nancy, France: ATILF & CLLE, 105-113.

Rospocher, M., Corcoglioniti, F. & Palmero Aprosio, A. 2019. PreMOn: LODifing linguistic predicate models. *Language Resources and Evaluation* 53(3): 499-524.

Saalfeld, G.A.E.A. 1884. *Tensaurus Italograecus: Ausführliches historisch-kritisches Wörterbuch der griechischen Lehn- und Fremdwörter im Lateinischen*. Wien, Austria: Carl Gerold's Sohn.

Schuurman, I., Windhouwer, M., Ohren, O. & Zeman, D. 2016, CLARIN concept registry: the new semantic registry. In *Selected Papers from the CLARIN Annual Conference 2015*. Wroclaw, Poland: Linköping University Electronic Press, 62-70.

Sprugnoli, R., Passarotti, M., Corbetta, D. & Peverelli, A. 2020. Odi et Amo. Creating, Evaluating and Extending Sentiment Lexicons for Latin. 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 International Conference on Language Resources and Evaluation (LREC 2020)*. Paris, France: European Language Resources Association (ELRA), 3078-3086.

Verkerk, P., Ouvrard, Y., Fantoli, M. & Longrée, D. 2020. LASLA and Collatinus: a convergence in lexica. *Studi e Saggi Linguistici* LVIII(1): 95-120.

Žabokrtský, Z. & Lopatková, M. 2007. Valency information in vallex 2.0. *The Prague Bulletin of Mathematical Linguistics* 87: 41-60.

Zinn, C. 2018. The Language Resource Switchboard. *Computational Linguistics* 44(4): 631-639.