Umanistica Digitale - ISSN:2532-8816 - n.8, 2020
F. Mambrini, F. M. Cecchini, G. Franzini, E. Litta, M. C. Passarotti, P. Ruffolo – LiLa: Linking Latin. Risorse linguistiche per il latino nel Semantic Web
DOI: http://doi.org/10.6092/issn.2532-8816/9975

LiLa: Linking Latin
Risorse linguistiche per il latino nel Semantic Web

[1]Francesco Mambrini, [2]Flavio Massimiliano Cecchini, [3]Greta Franzini, [4]Eleonora Litta, [5]Marco Carlo Passarotti, [6]Paolo Ruffolo

Università Cattolica del Sacro Cuore, Milano, Italia

[1]francesco.mambrini@unicatt.it
[2]flavio.cecchini@unicatt.it
[3]greta.franzini@unicatt.it
[4]eleonoramaria.litta@unicatt.it
[5]marco.passarotti@unicatt.it
[6]paolo.ruffolo@posteo.eu

Abstract

Il progetto LiLa: Linking Latin intende creare una Knowledge Base di risorse linguistiche (corpora, lessici digitali e strumenti di trattamento automatico del linguaggio) per lo studio del latino secondo il modello dei Linked Open Data. Questo articolo descrive gli obiettivi e le motivazioni del progetto, soffermandosi in particolare sulla centralità del lemma come forma che consente di creare la rete di informazioni linguistiche. L'articolo discute le strategie impiegate per creare una banca dati di lemmi latini che possa supportare tale modello, e i primi esperimenti volti a connettere risorse testuali (treebank del latino) alla raccolta dei lemmi.

The LiLa: Linking Latin project aims to build a Knowledge Base of language resources for the study of Latin (corpora, digital lexicons, natural-language-processing tools), based on the Linked Open Data paradigm. In this paper, we discuss the goals and motivation of the project. In particular, we focus on the role played by the lemma as a hub node that holds together the network of linguistic information. The architecture of LiLa is therefore based on lemmas and their morphological properties. The paper illustrates the strategies used to build a collection of Latin lemmas and the first experiments to link them to a set of textual resources (Latin treebanks).

Introduzione

Corpora, lessici, software

Nel corso degli ultimi due decenni, la comunità dei classicisti e dei linguisti ha assistito ad una

vera e propria esplosione delle risorse digitali a disposizione degli utenti per lo studio della lingua latina. Le collezioni digitali (*digital libraries*) i cui testi sono liberamente accessibili, e spesso scaricabili, da internet raggiungono oramai un'estensione che supera abbondantemente le centinaia di milioni di parole,[1] coprendo un variegato spettro di generi, epoche e autori. A titolo di esempio, possiamo citare la Perseus Digital Library, dedicata in particolare al latino classico e agli autori più tradizionalmente "scolastici"; in ambito italiano, basterà menzionare il database di testi poetici latini Musisque Deoque, l'Archivio della Latinità Italiana del Medioevo (ALIM) e la Digital library of late-antique Latin texts (digilibLT).[2]

Un'ulteriore classe di risorse è composta da quei progetti che, accanto alla serie spesso notevole di metadati filologici reperibile, ad esempio, presso i siti menzionati sopra, offrono anche annotazione linguistica, sia essa morfologica e/o sintattica. Il corpus sviluppato dal Laboratoire d'Analyse Statistique des Langues Anciennes di Liegi (LASLA) raccoglie testi dei maggiori autori classici latini arricchiti con analisi morfologica e lemmatizzazione, per un totale di più di 1,6 milioni di parole. Infine, ben quattro progetti offrono un'annotazione morfosintattica completa di testi latini, che vanno da autori classici come Cesare, la *Vulgata* di Girolamo o Virgilio raccolti nell'Ancient Greek and Latin Dependency Treebank (AGLDT) di Perseus e nel corpus PROIEL ([10]), a testi tardo-antichi come il trattato *De Agricoltura* di Palladio (PROIEL), fino alle *cartulae* notarili toscane di età carolingia nel corpus LLCT ([13]) e ai trattati latini di Tommaso d'Aquino nella *Index Thomisticus* Treebank ([22]).[3]

Accanto ai corpora digitali, linguisticamente annotati o meno, è attualmente disponibile un gran numero di risorse lessicali e di software per l'analisi automatizzata dei dati. Tra i primi, oltre a versioni digitalizzate di noti dizionari a stampa, possiamo ricordare il Latin WordNet ([19]), parte del MultiWordNet Project ([25]), che raccoglie una serie di lessici di numerose lingue allineate ai *synset* del Princeton WordNet e che, per il latino, comprende un numero abbastanza limitato di lemmi (9.124) e *synset* (8.973). Il lessico Word Formation Latin (WFL: [16]), invece, arricchisce i lemmi dell'analizzatore morfologico LEMLAT (vedi più sotto) con informazioni relative alla morfologia derivazionale delle parole.

Infine, un ulteriore prodotto dello sforzo di coniugare le tecnologie del linguaggio allo studio del latino è rappresentato dallo sviluppo di strumenti e modelli di trattamento automatico del linguaggio (TAL) per supportare l'analisi dei testi. Oltre ad alcuni "storici" analizzatori morfologici come Morpheus ([9]), TreeTagger ([28]) o LEMLAT ([23]), possiamo ricordare la ricca *suite* di librerie in Python per l'analisi di numerose lingue antiche, tra cui il latino, raccolte dal progetto Classical Language Toolkit (CLTK), la collezione di documenti e strumenti nel

1 Per non citare che un esempio, il *(meta-)repository* Corpus Corporum, che raccoglie in un'unica piattaforma testi di altre collezioni digitali, ha superato i 160 milioni di parole (con l'ultimo aggiornamento del 22 marzo 2019).

2 Per una panoramica di altre risorse (aggiornata, però, solo fino al 2011) e un'ottima introduzione si veda anche [1]:18-73.

3 PROIEL, Ancient Greek and Latin Dependency Treebank e *Index Thomisticus* Treebank sono disponibili in Universal Dependencies ([21]) in una versione convertita allo stile di annotazione del progetto.

progetto Computational Historical Semantics ([8]) o i modelli addestrati sui testi latini pubblicati nella raccolta di treebank Universal Dependencies disponibili per la *pipeline* di annotazione morfosintattica UDPipe ([29]).

E pluribus unum? Il problema dell'interoperabilità

La disponibilità di queste risorse digitali (e molte altre ancora) ha indubbiamente ampliato le potenzialità per lo studio della lingua latina e degli autori antichi. E tuttavia, benché l'oggetto cui tali risorse e strumenti si riferiscono e su cui operano sia intuitivamente il medesimo (gli autori, i testi e le parole della lingua latina), le informazioni che esse veicolano sono, al momento, destinate a restare sparse e a non interagire. Corpora, lessici e software di analisi linguistica, infatti, sono per lo più sviluppati autonomamente e la loro mancanza di reciproca interoperabilità ne limita fortemente l'utilità nello studio di fenomeni altamente complessi come la lingua, la letteratura o la storia delle civiltà e dei testi.

Il problema dell'interoperabilità è, in effetti, cruciale nell'attuale panorama delle risorse linguistiche digitali ([12]). Esso si articola in una duplice dimensione pratica: da un lato, molte risorse e strumenti vivono in isolamento, pubblicate su siti internet dedicati e fruibili solo attraverso piattaforme di ricerca indipendenti. Dall'altro, un limite ancora più serio è rappresentato dalla pluralità dei vocabolari per descrivere i concetti trattati: molteplici progetti che si occupano del medesimo fenomeno (ad esempio, l'annotazione morfosintattica) adottano spesso tagset e definizioni idiosincratiche; in mancanza di opportune formalizzazioni, è impossibile stabilire se, ad esempio, le etichette "Sb" e "SBJ" indichino la medesima funzione sintattica o, a maggior ragione, se la definizione di "soggetto" cui i due tag citati rimandano sia identica nei corpora in cui sono adottati, se la si intenda secondo le medesime regole di annotazione, o se piuttosto rinvii a teorizzazioni diverse della sintassi.

Il primo problema è stato parzialmente affrontato creando grandi infrastrutture che fungono da portale e consentono un accesso unificato alle risorse linguistiche, fra cui possiamo citare CLARIN ([11]), DARIAH ([27]) o META-SHARE. Per quanto riguarda il secondo scoglio, una soluzione di successo è quella di favorire l'uso di linee guida e tagset uniformati che possano essere condivisi fra progetti diversi; il già citato Universal Dependencies (vedi nota 3) raccoglie più di 100 treebank annotati usando il medesimo schema e linee guida unificate.[4]

Seppure tali lodevoli iniziative accrescano il livello di fruibilità generale dei diversi progetti, esse affrontano solo una dimensione pratica e superficiale del problema dell'integrazione. Una definizione più ambiziosa dell'interconnessione fra le risorse evoca, infatti, l'idea di una rete di conoscenza condivisa, dove è sempre possibile accedere simultaneamente alle informazioni che si riferiscono alla medesima entità linguistica. Risorse realmente interoperabili nel senso appena detto sono capaci di produrre un *network effect* tale da rendere il tutto maggiore della somma delle parti ([4]: iii). Una vera rete di risorse linguistiche consentirebbe all'utente di seguire i collegamenti fra, ad esempio, un'istanza di una parola, o di una classe di parole (per esempio, i

4 Si noti che, tra i progetti di annotazione menzionati, l'AGLDT e l'*Index Thomisticus* Treebank già dalla loro fondazione si sono basati su linee guida condivise ([2]).

sostantivi), all'interno di un corpus verso altri esempi d'uso in tutte le collezioni di testi disponibili e verso tutti i lessici (inclusi WordNet, WFL e altro) dove tali parole sono analizzate.

È questo il modello che il progetto LiLa: Linking Latin (LiLa) (2018-2023) intende adottare per connettere le risorse linguistiche del latino. LiLa, infatti, nasce con lo scopo di costruire una Knowledge Base basata sul paradigma dei Linked Open Data (LOD) (vedi sotto). Con questa definizione intendiamo una raccolta interoperabile di banche dati in cui le diverse risorse siano collegate tra di loro, descritte ed armonizzate usando i medesimi vocabolari.

Nelle rimanenti sezioni descriveremo in dettaglio l'architettura della Knowledge Base e illustreremo i primi tentativi di popolare la struttura ideata con le prime risorse linguistiche.

L'architettura di LiLa

Il lemma come fulcro

Uno degli obiettivi centrali del paradigma dei Linked Data è la creazione di un *Web of Data*, una rete che connetta, anziché documenti in HTML, le entità del mondo esterno che nei documenti pubblicati nel web sono discusse e menzionate ([3]). Per perseguire questa integrazione in relazione alle risorse linguistiche, LiLa ha adottato fin dal principio una prospettiva incentrata sul lessico: le "entità del mondo", per adottare la terminologia dei LOD, su cui corpora, lessici digitali e strumenti di TAL predicano informazioni sono le parole del lessico di una determinata lingua.

Nel trattamento di una lingua dalla complessa morfologia flessiva come il latino, la riduzione delle diverse forme attestate ad un lemma canonico è da sempre stata identificata come un'operazione fondamentale per favorire l'accesso ai testi. Diverse biblioteche digitali e software di ricerca, infatti, offrono agli utenti la possibilità di svolgere ricerche nelle opere a partire dal lemma; tali servizi sono basati sull'output di lemmatizzatori automatici, o più raramente sull'annotazione manuale.[5]

Oltre che utile in sé per le applicazioni alla ricerca lessicale appena viste, la lemmatizzazione è anche un'operazione fondamentale di *preprocessing* per il TAL di testi latini. Lemmatizzazione e Part-of-Speech (POS) tagging sono, infatti, operazioni molto vicine e spesso inestricabili, in

5 La ricerca implementata dalla Perseus Digital Library si basa sull'output del già menzionato Morpheus, incorporato anche nel popolare software Diogenes utilizzato per interrogare in locale le collezioni del Thesaurus Linguae Graecae, e dei corpora del Packard Humanities Institute PHI5 e PHI7. La lemmatizzazione offerta dal servizio di aggregazione di corpora Corpus Corporum è invece condotta utilizzando il già citato TreeTagger. Infine, il sito Perseus under PhiloLogic è basato su dati lemmatizzati manualmente. È da notare che Morpheus, contrariamente a TreeTagger e ovviamente all'annotazione manuale, non esegue alcuna forma di disambiguazione su base statistica, o normativa: in caso di molteplici lemmatizzazioni ammesse, tutte le analisi possibili sono fornite in output agli utenti. Per tale ragione, le statistiche lessicali della Perseus Digital Library indicano un range compreso tra un numero massimo di occorrenze per lemma (nel caso tutte le forme ambigue appartengano alla parola cercata) e un numero minimo (che include solo i casi non ambigui).

quanto molte forme ambigue comportano analisi diverse sia in termini di parte del discorso sia di lemma. La forma *rosam*, ad esempio, può essere interpretata tanto come nome (lemma *rosa*) quanto come participio del verbo *rodo*; informazioni sul lemma possono quindi risultare decisive per un tagging corretto, e viceversa. Ugualmente, lemmatizzazione e POS tagging costituiscono infine livelli di analisi spesso usati come portatori d'informazioni fondamentali da parte di software che eseguono analisi più avanzate, quali il parsing sintattico.

Per tali ragioni, il lemma è il primo e più produttivo livello di incontro tra risorse lessicali, che attraverso le forme citazionali canoniche indicizzano le entrate, i testi dei corpora e gli strumenti di TAL. In questa prospettiva, il primo compito fondamentale identificato per il successo di LiLa è quello di creare una raccolta di lemmi sufficientemente comprensiva da includere quelle forme canoniche che possono essere attestate nei dizionari, e dunque utilizzate per lemmatizzare manualmente i testi latini o fornite in output da software e pipeline di TAL. L'interconnessione basata sul paradigma dei LOD può dunque concretizzarsi nel progetto di collegare fra loro le risorse che predicano proprietà relative al medesimo lemma.

Lemmatizzazione del latino: criteri, strategie e problemi aperti

Fare interagire le risorse linguistiche per il latino attraverso l'uso del lemma quale mezzo per una loro interconnessione in LOD presenta delle difficoltà di non poco conto.

Benché la lemmatizzazione, definita come l'operazione di ricondurre le forme del paradigma flessionale di una parola ad una forma canonica, possa apparire un'attività concettualmente semplice, esistono casi ambigui, dove sussistono ampi margini per interpretazioni contrastanti tra i diversi annotatori e dove le soluzioni adottate nei vari progetti possono divergere. Alcune di queste divergenze sono dovute alla storia e a caratteri propri della lingua latina; altre sono legate a difficoltà più generali nel definire in cosa consista il lessico di una lingua.

Il latino ha attestazioni scritte che coprono un arco di circa venticinque secoli e continuano ancora al giorno d'oggi; oltre a una copiosa letteratura dell'epoca repubblicana e imperiale di Roma, il latino è stato per molti secoli la lingua franca della cultura, della scienza e dell'amministrazione in vaste aree del mondo ed è tuttora la lingua ufficiale della Chiesa Cattolica. Nonostante la lingua dotta abbia conosciuto una forte standardizzazione sul latino degli autori classici, il livello di variazione ortografica e morfologica che si riscontra rimane notevole. Come dovrà comportarsi un lemmatizzatore nel caso di coppie come *libido/lubido,* o in quello del latino classico *condicio* che sviluppa la grafia *conditio* nel latino tardo e medievale?[6] E come trattare verbi attestati sia come deponenti (privi cioè della diatesi attiva), sia all'attivo, come *abominor/abomino, dignor/digno o sequor/sequo*? Quale delle due forme, entrambe "canoniche" a loro modo, si dovrà adottare?

6 Sull'alternanza delle grafie in alcune espressioni giuridiche latine comunemente in uso anche in italiano (come *conditio/condicio sine qua non, sub condicione/conditione* o *par condicio*) si possono leggere, in risposta ai dubbi dei lettori, le note di Mussomeli ([20]) sul sito web dell'Accademia della Crusca.

Più in generale, inoltre, e indipendentemente dalle vicende proprie del latino, stilare un catalogo del lessico di una lingua, distinguendo le parole degne di un'entrata nel dizionario dalle ordinarie forme flesse, non è sempre un'operazione semplice o priva di zone grigie. Come considerare, ad esempio, gli avverbi deaggettivali come *celeriter*? Saranno da intendere come forme derivate dall'aggettivo (*celer*)? E si dovranno intendere i comparativi e superlativi politematici (ad esempio *melior*, usato come comparativo dell'aggettivo *bonus*) come forme derivate da lemmatizzare sotto il grado positivo, o come forme canoniche a sé stanti? Un caso anche più complesso è rappresentato dai participi, come ad esempio *doctus*, grammaticalmente forma del participio passato del verbo *doceo*, che può essere analizzato tanto come forma del paradigma verbale, da ricondurre quindi alla prima persona del presente indicativo, quanto come aggettivo o persino nome a sé stante (come nell'espressione (*homines*) *docti*, "i saggi").

I lemmi latini: per un'ontologia

Nel trattamento dei casi simili a quelli appena discussi le soluzioni adottate dai dizionari differiscono sensibilmente, così come possono differire le pratiche di lemmatizzazione adottate dai corpora o implementate dai lemmatizzatori. Ciascun progetto definisce le proprie linee guida per risolvere casi ambigui come quelli visti, ma è opportuno ricordare che l'approccio di LiLa non è quello di stilare un elenco di *best practices* o raccomandare una soluzione rispetto alle altre. Piuttosto, LiLa mira a integrare i diversi prodotti del lavoro di lemmatizzazione svolto dai vari progetti e le diverse entrate nei lessici digitali disponibili nel web. Per tale ragione, la sua architettura deve essere sufficientemente solida da poter, da un lato, rappresentare la totalità delle forme canoniche utilizzate per lemmatizzare un testo latino nella pluralità dei progetti esistenti e da essere in grado, dall'altro, di esprimere in modo corretto le relazioni possibili fra le diverse forme.

Il primo e più importante requisito che la modellizzazione delle forme canoniche in LiLa deve possedere, tuttavia, è la capacità di descrivere il livello delle forme in modo tale che questo piano morfologico possa poi essere messo in relazione con gli altri strati dell'analisi linguistica e, in particolare, con il livello lessicale. Per tale ragione, ci è parso naturale pensare alla nostra formalizzazione delle proprietà dei lemmi in LiLa come a una estensione di quella ontologia che è divenuta uno standard *de facto* per la rappresentazione delle risorse lessicali, ovvero Ontolex ([18]).

Il modello di Ontolex è primariamente pensato per rappresentare la lessicalizzazione di una data ontologia, ovvero il collegamento semantico fra i concetti rappresentati con il lessico di una lingua. Ontolex è stato già utilizzato per descrivere i *synset* di WordNet,[7] ed è stato esteso in molteplici direzioni per rappresentare altri fenomeni linguistici, tutti estremamente rilevanti per LiLa, come la morfologia derivazionale o l'etimologia ([14]). L'elemento primario di Ontolex è l'entrata lessicale (*Lexical Entry*), il cui collegamento alla dimensione concettuale può essere espresso secondo molteplici approcci; degli elementi lessicali, tuttavia, l'ontologia prevede classi per descrivere anche le diverse forme flesse e, attraverso la relazione di "forma canonica" (*canonical form*), anche il legame privilegiato fra il lemma e l'elemento lessicale.

7 https://www.w3.org/2016/05/ontolex/#conceptualization-set.

Il primo passo, dunque, consiste nel definire la classe primaria dell'ontologia di LiLa (il lemma, usato come forma canonica) come forma (*Form*) di Ontolex. Allo stesso tempo, in quanto forme, i lemmi possiedono alcune caratteristiche morfologiche che vogliamo descrivere. Innanzitutto, essi appartengono ad una (e non più di una)[8] parte del discorso e possono essere analizzati per i tratti tipici della morfologia nominale (genere, numero, caso), aggettivale (numero, genere, caso, grado) o verbale (tempo, modo, persona, numero, diatesi); inoltre, i lemmi appartengono ad una classe flessionale (le declinazioni e coniugazioni verbali della grammatica tradizionale). Nell'ontologia, sotto la super-classe dell'annotazione linguistica che si applica ad una forma, tali proprietà sono definite con le opportune restrizioni, di modo tale che non si possa, ad esempio, predicare il genere di un verbo. Per assicurare una compatibilità con i principali tagset in uso per il latino, le categorie impiegate in LiLa per rappresentare le parti del discorso, derivate dal tagset adottato per le POS in Universal Dependencies ([24]), sono allineate all'ontologia OLiA ([6]): per fare un esempio, la classe "Adjective" di LiLa è una sottoclasse di "Adjective" di OLiA, in cui è ricompreso anche l'insieme dei tag morfologici di nove caratteri dell'AGLDT che cominciano con la lettera "a". Usando questo modello, ci proponiamo di creare un framework di allineamento dei tagset in uso per l'annotazione linguistica del latino nelle varie risorse che integreremo in LiLa.

Una forma, secondo la definizione di Ontolex, è caratterizzata da una o più rappresentazioni scritte (*written representation*), ovvero stringhe che rappresentano la realizzazione della forma. Variazioni ortografiche, varianti dialettali, o grafie alternative rispetto a quella standardizzata possono essere registrate come *written representation* alternative; è questo il caso, per richiamare un esempio già menzionato, di *libido/lubido*, ma anche dei prestiti dal greco che alternano una forma completamente latinizzata (*hexagonum*) ad una che presenta la desinenza originale greca (*hexagonon*).

La variabilità fra strategie di lemmatizzazione nei casi di ambiguità lessicale o morfologica citati sopra (participi, comparativi, alternanza fra paradigma flessionale attivo e passivo) si rivela più complessa da modellizzare. Anzitutto, è da notare che il caso di *sequor/sequo* si presenta diverso da quello di *libido/lubido*: laddove in quest'ultimo la variazione grafica o fonetica non coinvolge l'interpretazione morfologica, i membri della prima coppia differiscono per almeno un tratto, segnatamente quello della diatesi (attiva in *sequo*, passiva in *sequor*). Poiché non possiamo escludere a priori che progetti che si trovassero ad annotare testi in cui sono attestate forme (pur se molto più rare) da ricondurre all'attivo possano scegliere di lemmatizzare proprio sotto *sequo* anziché sotto il più comune *sequor*, abbiamo deciso di considerare entrambi i membri di siffatte coppie come lemmi e di legarli in LiLa attraverso la speciale relazione simmetrica *lemma variant*.

Il caso dei participi è ancora diverso. In questa fattispecie, infatti, abbiamo a che fare non con

8 La definizione di "Lexical Entry" data nell'ontologia prescrive esplicitamente che ogni entrata lessicale abbia una singola parte del discorso ([7]). Così, la stessa restrizione si applica anche alle forme canoniche: se, ad esempio, la forma *supra* può corrispondere ad un avverbio e ad una preposizione, nel nostro modello ciò significherà che abbiamo a che fare con due lemmi, collegati come *canonical form* a due entrate lessicali differenti.

due lemmi alternativi, bensì con una forma specifica del paradigma flessionale che può essere "elevata" a forma canonica in determinate circostanze. Esempi simili sono definibili come una sottoclasse della famiglia dei lemmi, che abbiamo chiamato "ipolemmi" (*Hypolemma*). Un ipolemma è definito come una forma del paradigma flessionale del suo iperlemma, che può essere utilizzata come forma canonica di lemmatizzazione di alcuni token (occorrenza) nei corpora, o per indicizzare una voce specifica di un dizionario. Iperlemma e ipolemma sono connessi nella Knowledge Base attraverso le relazioni inverse *has hypolemma*/*is hypolemma*, definite nell'ontologia di LiLa. Tale relazione è stata estesa anche alla classe degli avverbi deaggettivali, i cui iperlemmi sono proprio gli aggettivi da cui essi derivano.

Infine, oltre che per i tratti di morfologia flessionale di cui abbiamo parlato, i lemmi sono analizzabili sulla base dei diversi morfemi utilizzati nei processi di derivazione. Questo livello di analisi morfologica ci consente di collegare la nostra banca dati di lemmi ai dati registrati nel già citato lessico WFL. Benché siano in corso iniziative per creare un modulo specifico di Ontolex dedicato alla morfologia derivazionale, tale estensione non è ancora a disposizione della comunità. Al momento, nonostante sia possibile definire i morfemi attraverso Ontolex come regolari entrate lessicali, dotate di una forma, ciascuna delle quali connessa a una o più rappresentazioni grafiche, abbiamo deciso di optare per un'estensione minimale della nostra ontologia. I morfemi di LiLa sono definiti come una classe a sé stante, suddivisa in affissi (a loro volta comprendenti suffissi e prefissi) e basi. Definiamo "Base" la classe aperta di quei morfemi lessicali che rimangono dopo che una parola è stata scomposta nei suoi morfemi derivazionali costitutivi e che risulta comune ad un'intera famiglia derivazionale, intesa come un insieme di lemmi che condividono la medesima base, come ad esempio *adduco*, *adductio*, *duco*, *productivus* e *produco*.[9]

Risorse linguistiche in LiLa: un primo abbozzo

Lemmi e morfemi: verso una Knowledge Base

L'ontologia delineata nella sezione precedente consente di descrivere gli aspetti morfologici (tanto flessionali quanto derivazionali) dei lemmi latini allo scopo di connetterli ai corpora lemmatizzati e alle entrate delle risorse lessicali.

Il passo successivo consiste nel creare una collezione la più esaustiva possibile di tali lemmi rappresentati secondo l'ontologia citata. Una risorsa estremamente preziosa per tale scopo è costituita dal già citato LEMLAT. Questo analizzatore morfologico è, infatti, fondato su una

9 A causa dell'elevato allomorfismo delle basi lessicali e dei complessi fenomeni fonetici che possono interessare i composti "base + affissi", non è semplice attribuire alla base una rappresentazione scritta; la natura stessa di questa "etichetta" può essere complessa da definire teoricamente. Proprio per questo, per non prendere una posizione sulla questione *se* una base lessicale debba necessariamente avere una *written representation* e come definirla, non abbiamo al momento allineato i morfemi di LiLa ai morfemi di Ontolex (i quali, come detto, in quanto *lexical entries* dovrebbero avere una forma canonica necessariamente dotata di una rappresentazione scritta).

base lessicale costituita da 43.433 lemmi tratti da dizionari del latino classico, recentemente arricchita di ulteriori 82.556 lemmi di latino medievale e 26.250 nomi propri ([23]), raccolti in un database relazionale; per ognuno degli oltre 150.000 lemmi, LEMLAT registra una serie di informazioni (quali genere, numero e, soprattutto, categoria flessionale) che sono utilizzate per produrre l'analisi dell'input.

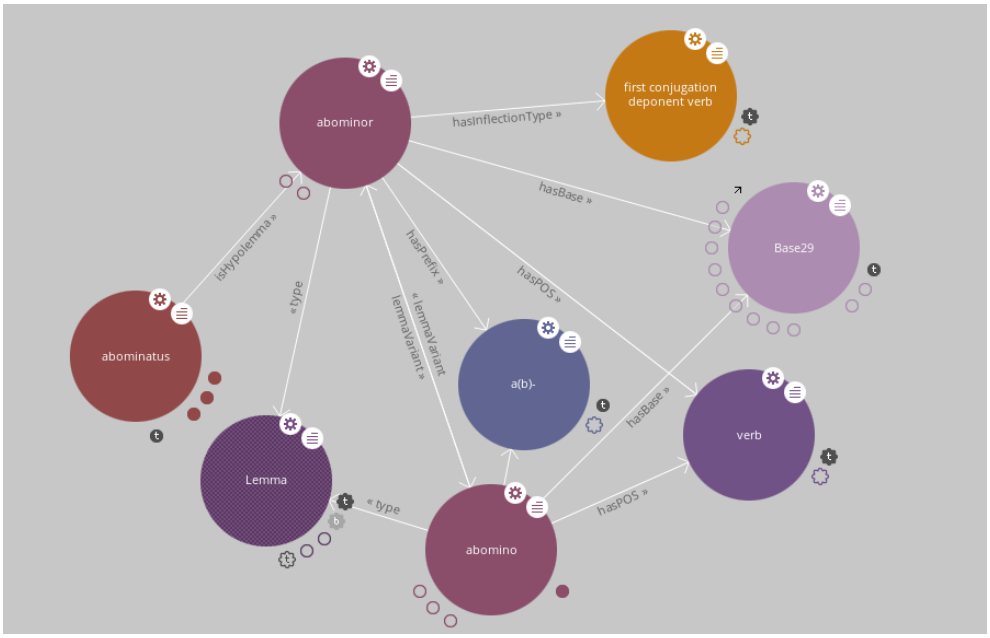

Figura 1: Il lemma *abominor* nella Knowledge Base LiLa

Il cuore della Knowledge Base di LiLa è, dunque, rappresentato da questo insieme di lemmi, in cui ad ognuna delle entrate del database di LEMLAT è assegnato un identificatore univoco, le opportune rappresentazioni grafiche e le informazioni morfologiche appropriate definite nell'ontologia. Inoltre, ciascun lemma è collegato ai propri ipolemmi e alle varianti, secondo i criteri illustrati più sopra; poiché gli ipolemmi non sono inclusi tra i lemmi usati da LEMLAT, abbiamo proceduto a generare le forme participiali (presenti, passate e future) di tutti i verbi e gli avverbi deaggettivali.

Le relazioni fra lemmi, classi e proprietà morfologiche sono espresse utilizzando il *Resource Description Framework* (RDF) ([15]) in forma di triple e sono salvate in un triplestore, che può essere interrogato utilizzando il linguaggio di query SPARQL ([26]). Al momento, la base lessicale di LiLa comprende 130.925 lemmi, 92.947 ipolemmi, 292.657 *written representation* relative a lemmi e ipolemmi, 59.945 relazioni tra lemmi e ipolemmi e 6.120 relazioni di *lemma variant*.[10] Infine, la Knowledge Base comprende 118 suffissi, 39 prefissi e 3.990 basi lessicali.

10 Si noti che, dato che la lista dei lemmi di latino medievale di LEMLAT si sovrappone in alcuni casi a quella di lemmi classici, è stata condotta una revisione dei dati per verificare la presenza di doppioni e,

La Figura 1 mostra l'esempio della rete di relazioni intrattenute dal lemma *abominor* nella Knowledge Base LiLa. Il verbo è connesso alla sua classe flessionale (verbo deponente della prima coniugazione) e alla parte del discorso; inoltre, esso è legato da una relazione reciproca (si notino le frecce bidirezionali) alla sua variante attiva *abomino,* secondo quell'alternanza non infrequente attivo/deponente che abbiamo citato a mo' di esempio; infine, esso ha una serie di ipolemmi, tra cui solo il participio passato *abominatus* è riportato in figura. Le due varianti sono poi entrambe connesse ai due morfemi che ne spiegano la derivazione nella famiglia cui appartiene anche il verbo *ominor*, ovvero il prefisso *a(b)-* e la base lessicale (contrassegnata dall'identificativo numerico "29") condivisa anche dal sostantivo *omen*. Espandendo quest'ultima entità, la "Base 29", è possibile raggiungere tutti gli altri lemmi connessi alla famiglia derivazionale e mostrare le informazioni morfologiche e le connessioni di ciascuno. Seguendo, invece, il link rappresentato dalla categoria flessionale si possono elencare tutti i verbi deponenti della prima coniugazione registrati in LiLa, espandendo significativamente la rete d'interazioni mostrate.

Le informazioni relative ai lemmi e alle loro proprietà sono accessibili al pubblico sia attraverso uno SPARQL endpoint sia attraverso un'interfaccia di ricerca semplificata;[11] la collezione può essere altresì esplorata attraverso una piattaforma che produce visualizzazioni analoghe a quella riportata nella Figura 1.[12]

La pagina da cui è possibile interrogare lo SPARQL endpoint include alcune ricerche già predisposte a titolo esemplificativo. In una di esse, ad esempio, è possibile richiedere la lista di lemmi connessi alla base lessicale a cui appartiene il lemma *classis* "classe, flotta"; i nove risultati includono lemmi come *classicus* "appartenente ad una classe / alla flotta" o *classarii* "soldati della flotta". L'interfaccia di ricerca consente di combinare le proprietà linguistiche descritte nell'ontologia (come ad esempio la parte del discorso, il genere, il collegamento ad una base o ad un affisso) ed ottenere l'elenco dei lemmi che soddisfano i criteri selezionati dall'utente.

Verso l'integrazione dei corpora testuali

La più immediata applicazione della Knowledge Base consiste nel collegare i lemmi alle parole attestate nei corpora che possono essere lemmatizzate sotto di essi. Nel *Bellum Gallicum* di Cesare, ad esempio, *incolunt* (1.1.1 e 1.1.4), *incoluerant* (1.5.3) e *incolant* (2.3.4) sono forme (tra le tante altre) del verbo *incolo* e ciascuna di queste occorrenze può essere collegata al nodo di tale lemma in LiLa tramite un'opportuna proprietà dell'ontologia.

Il collegamento fra token e lemma sotto cui la forma può essere lemmatizzata viene esteso alla totalità di un corpus testuale. Naturalmente, un prerequisito imprescindibile è che il corpus che si vuole collegare abbia informazioni granulari sulla lemmatizzazione delle parole. Poiché, come si è accennato, la parte del discorso può contribuire a disambiguare forme e persino lemmi

nel caso, rimuoverli. Tale processo è ancora in corso e alcuni lemmi ripetuti potranno venire fusi in una sola forma a breve.

11 Raggiungibili rispettivamente agli indirizzi: https://lila-erc.eu/sparql/ e https://lila-erc.eu/query/.

12 La visualizzazione può essere generata a partire dalla pagina: https://lila-erc.eu/lodlive/.

omografi (come ad esempio nel caso di *artus*, aggettivo, "stretto" e *artus*, sostantivo, "arto, articolazione"), è ragionevole includere anche il POS tagging tra i requisiti. Una risorsa testuale, dunque, deve essere annotata come minimo con questi due tipi di informazione per poter essere connessa alla rete delle risorse di LiLa.

I treebank del latino, che come si è accennato includono le informazioni richieste,[13] sono un punto di partenza ottimale per valutare operativamente il procedimento di interconnessione fra le risorse attraverso LiLa.

Un primo passo verso tale integrazione è la conversione dei treebank, che sono distribuiti dai singoli sviluppatori in una varietà di formati, in una semplice rappresentazione RDF in cui frasi e token divengono nodi identificati tramite identificatori unici (URI: Uniform Resource Identifier). In un primo stadio di conversione, la maggior parte delle annotazioni viene registrata come un semplice attributo del nodo in forma di stringa (*data property*, in RDF), mentre le relazioni sintattiche di dipendenza o le relazioni lineari di sequenza nel testo possono essere rappresentate come link tra nodi.[14]

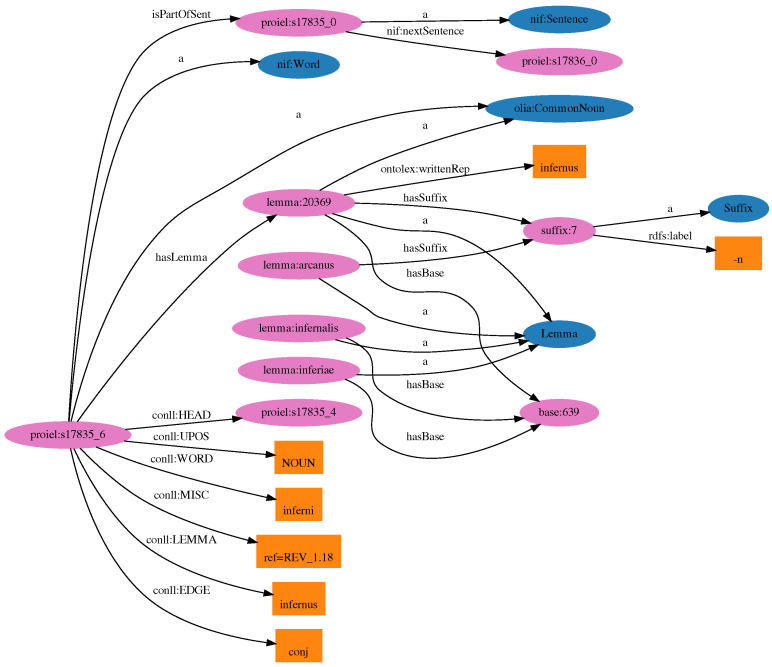

Figura 2: LiLa e un token del corpus PROIEL (Vg. Ap. 1,18)

A questo punto, un'elementare forma di match può essere effettuata confrontando ciascuna delle rappresentazioni scritte dei lemmi alla proprietà del token che registra la stringa usata per

13 Vedi la sezione Corpora, lessici, software.

14 Per questa prima conversione dei treebank a RDF abbiamo utilizzato il software conll-rdf ([5]).

lemmatizzare la parola nel corpus. Nel caso in cui il corpus adotti una stringa di lemmatizzazione ambigua come *artus*, che corrisponde a più lemmi di LiLa, la parte del discorso può essere usata per consentire la disambiguazione. Alcuni casi di effettiva omografia fra parole che hanno la medesima POS, come i verbi *uolo*, "volare", e *uolo*, "volere", oppure i sostantivi *tempus*, "tempo" e *tempus,* "tempia" (che condividono persino il paradigma flessionale), sono tuttavia destinati a rimanere insoluti.

La Figura 2 raffigura un esempio dell'esito di un primo esperimento di collegamento operato sul corpus PROIEL. Il nodo posto alla sinistra nella figura (proiel:s17835_6) rappresenta il token corrispondente alla parola *inferni* tratta da una frase del corpus.[15] Tra le proprietà che corrispondono alle annotazioni disponibili nella risorsa (i rettangoli arancioni in basso a sinistra) si nota la stringa che rappresenta il lemma: *infernus*. Tale stringa corrisponde alla rappresentazione scritta di due lemmi di LiLa, l'aggettivo *infernus*, "sotterraneo, infernale" e il sostantivo corrispondente all'italiano "inferno". Solo quest'ultimo, tuttavia, condivide la POS con il token di PROIEL, cosicché la disambiguazione basata sulla parte del discorso lascia un solo candidato possibile.

La creazione di un link fra *inferni* in PROIEL e il lemma appropriato in LiLa inserisce il token della *Apocalisse* in una fitta rete di informazioni linguistiche. Il lemma, infatti, è legato a una base lessicale cui appartengono altre parole della medesima famiglia derivazionale, tra cui l'aggettivo *inferus*, "inferiore, infernale", e il sostantivo *inferiae*, "offerte ai morti"; il nome *infernus* è poi formato tramite il suffisso *-n*, lo stesso utilizzato, ad esempio, nella derivazione (che non è mostrata nella figura, ma è immediatamente recuperabile dai dati di WFL connessi a LiLa) di *arcanus*, "segreto, nascosto", da *arca*, "cassa, scrigno".

Conclusioni

L'esempio di *inferni* permette già di comprendere quanto l'interconnessione di informazioni linguistiche tramite le connessioni ai lemmi di LiLa possa ampliare le possibilità di ricerca e, più in generale, di applicazione delle risorse per il latino.

Grazie all'infrastruttura di LiLa, diviene possibile immaginare di estrarre dati da un'unica interfaccia per rispondere a domande come: quali sono le coppie verbo-soggetto in cui il soggetto sia un nome composto con il prefisso *(-)tor* ([17])? Oppure, quali sono i prefissi, suffissi e le basi lessicali più frequenti nei diversi autori i cui testi sono connessi a LiLa?

Simili esempi sono, come si vede, limitati al livello della sola morfologia derivazionale, per cui i dati collegati ai lemmi della Knowledge Base sono già oggi integrati. Tuttavia, per realizzare appieno le potenzialità dell'architettura il primo requisito è ovviamente quello di espandere il numero e la copertura dei corpora e delle risorse connesse.

Laddove alcune di esse, come i treebank che abbiamo citato, sono già mature a sufficienza per

15 La frase intera è tratta dalla *Apocalisse* (1,18) di Giovanni nella versione della *Vulgata* di Girolamo: *et habeo claves mortis et inferni*, "e ho le chiavi della morte e degli inferi". La versione di PROIEL che è stata convertita ad RDF e collegata a LiLa è quella inclusa in Universal Dependencies (versione 2.3).

essere collegate ai lemmi di LiLa, molte altre necessitano di un supplemento di lavoro editoriale o di annotazione. Il Latin WordNet, come si è detto, ha una copertura ancora relativamente limitata di parole, le cui connessioni ai *synset*, per di più, devono in alcuni casi essere ancora validate da esperti.

Soprattutto, poi, la maggioranza delle edizioni digitali di testi latini disponibili sul web non soddisfa i prerequisiti cui abbiamo accennato. Milioni di parole potenzialmente integrabili alle risorse di LiLa non sono né lemmatizzate né annotate per quel che riguarda la parte del discorso. In questo, un aiuto fondamentale può venire dagli strumenti di TAL come tagger e lemmatizzatori. È indispensabile, tuttavia, che il grado di precisione di questi software sia innanzitutto valutato accuratamente e, nel caso, portato ad uno standard accettabile per la comunità degli utenti di LiLa.

Gli esempi discussi e le risorse che sono potenzialmente integrabili in LiLa illustrano bene, come ci auguriamo, quanto la connessione dell'informazione linguistica renda il tutto maggiore della somma delle componenti. La nostra speranza è che LiLa possa contribuire ad alimentare in modo incisivo lo sfruttamento delle diverse collezioni di edizioni digitali, dizionari e software per lo studio del latino già oggi disponibili sul web, indipendentemente dalle dimensioni dei progetti e dal loro grado di notorietà fra gli addetti ai lavori. Grazie alla sua natura aperta e *open-ended*, LiLa può certamente ambire a diventare un punto di riferimento per chiunque voglia pubblicare risorse linguistiche, edizioni o archivi digitali di testi che abbiano a che fare con la lingua latina, nel senso più lato possibile.

Ringraziamenti

Il progetto è finanziato dallo European Research Council (ERC) nel quadro del programma di ricerca e innovazione Horizon 2020 – Grant Agreement No. 769994.

References

[1] Babeu, Alison. 2011. *Rome Wasn't Digitized In A Day. Building a Cyberinfrastructure for Digital Classicists*. Washington, DC: Council on Library and Information Resources. http://www.clir.org/pubs/reports/pub150.

[2] Bamman, David, Marco Passarotti, Gregory Crane, e Savina Raynaud. 2007. «A collaborative model of treebank development». In *Proceedings of the Sixth International Workshop on Treebanks and Linguistic Theories*, 1–6. Bergen: Northern European Association for Language Technology (NEALT). http://tlt07.uib.no/papers/4.pdf.

[3] Bizer, Christian, Tom Heath, e Tim Berners-Lee. 2009. «Linked data: The story so far». *Journal on Semantic Web and Information Systems* 5 (3): 1–22. https://doi.org/10.4018/jswis.2009081901.

[4] Chiarcos, Christian, Sebastian Hellmann, e Sebastian Nordhoff. 2012. «Introduction and Overview». In *Linked Data in Linguistics: Representing and Connecting Language Data and Language Metadata*, a cura di Christian Chiarcos, Sebastian Nordhoff, e Sebastian Hellmann, 1–12. Berlin, Heidelberg: Springer. https://doi.org/10.1007/978-3-642-28249-2_1.

[5] Chiarcos, Christian, and Christian Fäth. 2017. "CoNLL-RDF: Linked Corpora Done in an NLP-Friendly Way." In *International Conference on Language, Data and Knowledge*, edited by Jorge Gracia, Francis Bond, John McCrae, Paul Buitelaar, Christian Chiarcos, and Sebastian Hellmann, 74–88. Berlin: Springer. https://doi.org/doi.org/10.1007/978-3-319-59888-8_6.

[6] Chiarcos, Christian, e Maria Sukhareva. 2015. «Olia – ontologies of linguistic annotation». *Semantic Web* 6 (4): 379–386. http://www.semantic-web-journal.net/system/files/swj518_0.pdf.

[7] Cimiano, Philipp, John P. McCrae, e Paul Buitelaar. 2016. «Lexicon Model for Ontologies: Community Report, 10 May 2016». 2016. https://www.w3.org/2016/05/ontolex/.

[8] Cimino, Roberta, Tim Geelhaar, e Silke Schwandt. 2015. «Digital Approaches to Historical Semantics: new research directions at Frankfurt University». *Storicamente* 11. http://dx.doi.org/10.12977/stor594.

[9] Crane, Gregory. 1991. «Generating and Parsing Classical Greek». *Literary and Linguist Computing* 6 (4): 243–45. https://doi.org/10.1093/llc/6.4.243.

[10] Haug, Dag Trygve Truslew, e Marius Larsen Jøhndal. 2008. «Creating a Parallel Treebank of the Old Indo-European Bible Translations». In *Proceedings of the Second Workshop on Language Technology for Cultural Heritage Data (LaTeCH 2008)*, 27–34. Marrakech, Morocco: European Language Resources Association (ELRA). https://www.hf.uio.no/ifikk/english/research/projects/proiel/Activities/proiel/publications/marrakech.pdf.

[11] Hinrichs, Erhard, e Steven Krauwer. 2014. «The CLARIN research infrastructure: resources and tools for e-humanities scholars». In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)*, 1525–1531. Reykjavik, Iceland: European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2014/pdf/415_Paper.pdf.

[12] Ide, Nancy, e James Pustejovsky. 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. https://www.cs.vassar.edu/~ide/papers/ICGL10.pdf.

[13] Korkiakangas, Timo, e Marco Passarotti. 2011. «Challenges in annotating medieval Latin charters». *Annotation of Corpora for Research in the Humanities*, 105.

https://jlcl.org/content/2-allissues/12-Heft2-2011/16.pdf.

[14] Khan, Anas Fahad. 2018. «Towards the Representation of Etymological Data on the Semantic Web». *Information* 9 (12): 304. https://doi.org/doi:10.3390/info9120304.

[15] Lassila, Ora, and Ralph R. Swick. 1998. «Resource Description Framework (RDF) Model and Syntax Specification.» 1998. https://www.w3.org/TR/1999/REC-rdf-syntax-19990222/.

[16] Litta, Eleonora, Marco Passarotti, e Chris Culy. 2016. «Formatio formosa est. Building a Word Formation Lexicon for Latin». In *Proceedings of the third italian conference on computational linguistics (clic–it 2016)*, 185–189. Naples: aAccademia University Press. http://ceur-ws.org/Vol-1749/paper32.pdf.

[17] Mambrini, Francesco, e Marco Passarotti. 2019. «Linked Open Treebanks. Interlinking Syntactically Annotated Corpora in the LiLa Knowledge Base of Linguistic Resources for Latin». In *Proceedings of the 18th International Workshop on Treebanks and Linguistic Theories (TLT, SyntaxFest 2019)*, 74–81. Paris: Association for Computational Linguistics. https://www.aclweb.org/anthology/W19-7808.pdf.

[18] McCrae, John P., Julia Bosque-Gil, Jorge Gracia, Paul Buitelaar, e Philipp Cimiano. 2017. «The OntoLex-Lemon Model: development and applications». In *Proceedings of eLex 2017*, 587–97. https://elex.link/elex2017/wp-content/uploads/2017/09/paper36.pdf.

[19] Minozzi, Stefano. 2017. «Latin WordNet, una rete di conoscenza semantica per il latino e alcune ipotesi di utilizzo nel campo dell'information retrieval». In *Strumenti digitali e collaborativi per le Scienze dell'Antichità*, a cura di Paolo Mastrandrea, 123–33. Università Ca' Foscari Venezia, Italia. https://doi.org/10.14277/6969-182-9/ANT-14-10.

[20] Mussomeli, Chiara. 2009. «Condicio o conditio sine qua non?» Accademia della Crusca. 11 dicembre 2009. http://www.accademiadellacrusca.it/it/lingua-italiana/consulenza-linguistica/domande-risposte/condicio-conditio-sine-qua.

[21] Nivre, Joakim, Marie-Catherine De Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic, Christopher D Manning, Ryan McDonald, et al. 2016. «Universal dependencies v1: A multilingual treebank collection». In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, 1659–1666. Portorož, Slovenia: European Language Resources Association. http://www.lrec-conf.org/proceedings/lrec2016/pdf/348_Paper.pdf.

[22] Passarotti, Marco. 2009. «Theory and Practice of Corpus Annotation in the "Index Thomisticus Treebank"». *Lexis* 27: 5–24. http://hdl.handle.net/10807/1403.

[23] Passarotti, Marco, Marco Budassi, Eleonora Litta, e Paolo Ruffolo. 2017. «The Lemlat 3.0 Package for Morphological Analysis of Latin». In *Proceedings of the NoDaLiDa 2017 Workshop on Processing Historical Language*, 24–31. Linköping University

Electronic Press. http://www.ep.liu.se/ecp/133/006/ecp17133006.pdf.

[24] Petrov, Slav, Dipanjan Das, and Ryan McDonald. 2011. «A Universal Part-of-Speech Tagset.» *ArXiv Preprint* https://arxiv.org/abs/1104.2086.

[25] Pianta, Emanuele, Luisa Bentivogli, e Christian Girardi. 2002. «MultiWordNet: developing an aligned multilingual database». In *Proceedings of the First International Conference on Global WordNet*. Vol. 152. Mysore, India. http://multiwordnet.fbk.eu/paper/MWN-India-published.pdf.

[26] Prud'hommeaux, Eric, and Andy Seaborne. 2008. «SPARQL Query Language for RDF.» 2008. https://www.w3.org/TR/rdf-sparql-query/.

[27] Romary, Laurent, e Jennifer Edmond. 2019. «A Tangential View on Impact for the Arts and Humanities through the Lens of the DARIAH-ERIC». In *Stay Tuned To The Future - Impact of the Research Infrastructures for Social Sciences and Humanities*, a cura di Bente Maegaard e Riccardo Pozzo, 149–58. Florence: Olschki. https://hal.inria.fr/hal-02094713.

[28] Schmid, Helmut. 1999. «Improvements in part-of-speech tagging with an application to German». In *Natural language processing using very large corpora*, 13–25. Springer. https://link.springer.com/chapter/10.1007%2F978-94-017-2390-9_2.

[29] Straka, Milan, e Jana Straková. 2017. «Tokenizing, POS Tagging, Lemmatizing and Parsing UD 2.0 with UDPipe». In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, 88–99. Vancouver, Canada: Association for Computational Linguistics. http://www.aclweb.org/anthology/K/K17/K17-3009.pdf.