LE OPERE LATINE DI DANTE TRA ANNOTAZIONE LINGUISTICA E WEB SEMANTICO

MARCO PASSAROTTI · GIULIA PEDONESE
RACHELE SPRUGNOLI [1]

RIASSUNTO · In questo articolo viene descritto il processo di realizzazione di UDante, una risorsa linguistica che raccoglie i testi latini di Dante Alighieri annotati linguisticamente a diversi livelli. I testi sono tokenizzati, divisi in frasi, lemmatizzati, annotati in base alle parti del discorso e arricchiti con un'annotazione sintattica a dipendenze in base ai criteri dell'iniziativa Universal Dependencies. Inoltre, viene presentata la procedura di allacciamento di UDante alla Base di Conoscenza di LiLa, che raccoglie e connette tra loro risorse lessicali e testuali secondo il paradigma Linked Data rendendole interoperabili. Le potenzialità di questo allacciamento e della creazione di risorse interoperabili per svolgere studi sui testi danteschi sono mostrate attraverso esempi di interrogazione della Base di Conoscenza: in particolare, sono descritti casi di analisi di tipo lessicale e sintattico.

PAROLE CHIAVE · Dante, latino, Linguistic Linked Data, annotazione linguistica, sintassi.

ABSTRACT · *Dante's Latin works between linguistic annotation and semantic Web* · This paper describes the development of UDante, a linguistic re-

marco.passarotti@unicatt.it, Università Cattolica del Sacro Cuore, Milano, Italia.

giulia.pedonese@unicatt.it, Università Cattolica del Sacro Cuore, Milano, Italia.

rachele.sprugnoli@unicatt.it, Università Cattolica del Sacro Cuore, Milano, Italia.

[1] Questo articolo è il risultato della collaborazione tra i tre autori. Per il sistema di attribuzione accademico italiano si specifica che: Marco Passarotti è responsabile delle sezioni 1 e 5, Giulia Pedonese della sezione 2.2 e Rachele Sprugnoli della sezione 2.1. Le sezioni 3 e 4 sono state scritte collaborativamente da Giulia Pedonese e da Rachele Sprugnoli.

HTTPS://DOI.ORG/DOI 10.19272/202101602003 · «LINGUISTICA E LETTERATURA», XLVI, 1-2, 2021

RECEIVED: 30.10.2021 · REVISED: 18.12.2021 · ACCEPTED: 23.12.2021

source that collects Dante Alighieri's Latin works annotated linguistically at different levels. The texts are tokenized, splitted into sentences, lemmatized, tagged with parts of speech and enriched with a layer of syntactic annotation based on the Universal Dependencies style. Furthermore, the procedure for including UDante into the LiLa Knowledge Base is presented: this Knowledge Base collects and connects lexical and textual resources according to the Linked Data paradigm, making them interoperable. The potential of this connection and of the creation of interoperable resources for enhancing the studies on Dante's texts are shown through examples of queries on the Knowledge Base: in particular, cases of lexical and syntactic analysis are described.

KEYWORDS · Dante, Latin, Linguistic Linked Data, Linguistic annotation, Syntax.

1. INTRODUZIONE

Lo studio linguistico, letterario e filologico dei testi antichi è strettamente legato all'esistenza di corpora che li raccolgono e consentono d'interrogarne i contenuti. Non è un caso che una delle prime risorse linguistiche registrate su supporto elettronico sia stato l'*Index Thomisticus* di padre Roberto Busa, che raccoglie l'opera omnia in latino di Tommaso d'Aquino.[1] E, ancora, una delle prime biblioteche digitali realizzate è stata la *Perseus Digital Library*: avviata nel 1987, con il fine di raccogliere materiali a supporto dello studio dell'antica Grecia (e dei testi in greco antico), nei decenni è divenuta un riferimento nell'area delle lingue classiche grazie al suo ampio catalogo di testi, dizionari e lessici digitali per il greco e il latino.[2]

Per quanto riguarda l'italiano antico, l'istituto dell'*Opera del Vocabolario Italiano* (OVI-CNR) pubblica fin dal 1997 le voci del *Tesoro della Lingua Italiana delle Origini* (*TLIO*), un dizionario storico redatto sulla base dei dati testuali forniti da un ampio corpus (il

[1] ROBERTO BUSA, *Index Thomisticus Sancti Thomae Aquinatis Operum Omnium Indices et Concordantiae in Quibus Verborum Omnium et Singulorum Formae et Lemmata cum Suis Frequentiis et Contextibus Variis Modis Referuntur*, Stuttgart-Bad Cannstatt, Frommann-Holzboog, 1974-1980.

[2] http://www.perseus.tufts.edu/hopper/.

Corpus TLIO *per il vocabolario*, da ora *Corpus* TLIO) che include poco meno di 3.000 testi (dalle origini alla fine del XIV secolo) per un totale di più di 23 milioni di occorrenze, di cui circa 4 milioni sono lemmatizzate.[1]

Tra i testi registrati nel *Corpus* TLIO sono presenti anche le opere volgari di Dante Alighieri, altresì disponibili nel corpus *DanteSearch,* che ne fornisce la lemmatizzazione e l'annotazione dei tratti morfologici e delle relazioni sintattiche tra proposizioni principali e subordinate.[2] Oltre alle opere in volgare, *DanteSearch* include anche i testi in latino di Dante: *De vulgari eloquentia* (DVE), *Egloghe*, *Epistole*, *Monarchia*, *Quaestio de aqua et terra*. Questi ultimi sono stati oggetto di due attività condotte nell'ambito del progetto ERC-Consolidator LiLa: Linking Latin[3] presso il centro di ricerca CIRCSE dell'Università Cattolica di Milano, in collaborazione con il Dipartimento di Filologia, Letteratura e Linguistica dell'Università di Pisa:

(1) l'annotazione morfologica e sintattica dei testi in accordo con uno standard de facto oggi applicato a dati in più di 100 lingue. In particolare, l'annotazione morfologica è stata prodotta attraverso un processo di conversione dai codici morfologici di *DanteSearch*, mentre quella sintattica è stata realizzata manualmente. Questa attività ha portato alla creazione di una nuova risorsa linguistica, denominata UDante, che consiste nella *treebank* (un corpus annotato sintatticamente) dei testi latini di Dante Alighieri;

(2) il collegamento a una Base di Conoscenza (in inglese,

[1] Il *Corpus* TLIO è consultabile online attraverso la versione web del programma GATTO: http://gattoweb.ovi.cnr.it/(S(rduxldwcoqud4bjhvhaobmao))/CatForm01.aspx. Il dizionario storico (TLIO) è invece disponibile al sito: http://tlio.ovi.cnr.it/TLIO/.

[2] https://dantesearch.dantenetwork.it/, vedi Mirko Tavoni, *DanteSearch: il corpus delle opere volgari e latine di Dante lemmatizzate con marcatura grammaticale e sintattica*, in *Lectura Dantis 2002-2009. Omaggio a Vincenzo Placella per i suoi settanta anni*, II: *Lectura Dantis 2004 e 2005*, a cura di Anna Cerbo, con la collaborazione di Roberto Mondola, Aleksandra Žabjek, Ciro Di Fiore, Napoli, Università degli Studi di Napoli L'Orientale, Il Torcoliere, pp. 583-608.

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

Knowledge Base) che consente l'interazione via web tra i testi danteschi in latino e altri testi, dizionari e lessici latini.

La motivazione di fondo che ha dato avvio allo sviluppo di queste attività è consistita nella necessità di innestare i testi latini di Dante raccolti in *DanteSearch* nello stato dell'arte del settore delle risorse linguistiche, in particolare accordandoli a standard condivisi (sia di formato che di annotazione) il cui rispetto è oggi richiesto al fine di valorizzare una specifica risorsa attraverso l'interazione con le molte altre sviluppate nel corso dei decenni.

Infatti, pur definiti e applicati con criteri scientificamente validi, i criteri di annotazione sintattica dei testi adottati in *DanteSearch* si discostano dalle attuali pratiche di annotazione dei corpora sintattici. Se da un lato ciò può essere considerato un valore aggiunto di *DanteSearch*, in quanto il suo stile di annotazione fornisce una chiave di accesso originale alla sintassi dei testi danteschi, dall'altro è un limite, dal momento che impedisce di fare ricerche sui testi che interroghino corpora sintattici diversi da *DanteSearch* stesso. Con l'obiettivo di armonizzare gli stili di annotazione delle *treebank* rese disponibili per molte lingue nel corso degli anni, dal 2015 il progetto *Universal Dependencies* (UD)[1] pubblica regolarmente una raccolta di *treebank* a dipendenze annotate secondo uno schema condiviso, che nel tempo si è imposto come uno standard de facto di annotazione sintattica. L'ultima versione di UD (2.8; 15 maggio 2021) include 202 *treebank*, che coprono 114 lingue, tra cui il latino, che è presente con 5 *treebank*. Oltre a UDante,[2] si segnalano: l'*Index Thomisticus Treebank* (ITTB) contenente testi di Tommaso D'Aquino,[3] la *treebank* PROIEL con testi sia di epoca classica che tarda nonché la traduzione della Bibbia realizzata da

[1] https://universaldependencies.org/, vedi Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, Daniel Zeman, *Universal dependencies*, «Computational Linguistics», XLVII, 2, 2021, pp. 255-308.

[2] https://github.com/UniversalDependencies/UD_Latin-UDante.

[3] Marco Passarotti, *The Project of the Index Thomisticus Treebank*, in *Digital Classical Philology. Ancient Greek and Latin in the Digital Revolution*, edited by Monica Berti, Berlin-Boston, de Gruyter Saur, 2019 («Age of Access? Grundfragen der Informationsgesellschaft», 1), pp. 299-320.

San Girolamo,[1] la *treebank Perseus* con testi solo di epoca classica[2] e la *Late Latin Charter Treebank* (LLCT) comprendente testi legali toscani scritti tra l'VIII e il IX secolo.[3]

Se lo sviluppo di una *treebank* secondo lo schema di UD consente all'opera omnia dantesca in lingua latina di condividere il medesimo formato e gli stessi criteri ed etichette di annotazione con centinaia di altri corpora sintattici, la seconda attività presentata in questo articolo ha l'obiettivo di rendere i testi latini di Dante (e le loro annotazioni metalinguistiche) interoperabili con altre risorse linguistiche per il latino, quali ad esempio corpora, lessici e dizionari, attraverso l'allacciamento a una Base di Conoscenza condivisa chiamata LiLa costruita in base ai principî del paradigma *Linked Data*. Ciò significa che tutte le risorse del latino connesse a LiLa possono interagire tra loro attraverso l'utilizzo di rappresentazioni formali condivise delle classi di (meta)dati in questione (come entrate lessicali, occorrenze testuali, etichette di parti del discorso e lemmi) e di ontologie comuni per esprimerne le reciproche relazioni. Attraverso questo standard, le risorse testuali e lessicali per il latino allacciate a LiLa possono essere interrogate attraverso un punto di accesso comune, che permette ricerche sui loro (meta)dati. Ciò consente a UDante di andare oltre sé stessa, diventando parte di un ecosistema sul web, dove le risorse linguistiche sono valorizzate al meglio proprio in virtù della possibilità di essere utilizzate in modalità interoperabile secondo principî e modelli di dati condivisi.

[1] Hanne Eckhoff *et alii*, *The PROIEL treebank family: a standard for early attestations of Indo-European languages*, «Language Resources and Evaluation», LII, 1, 2018, pp. 29-65.

[2] David Bamman, Gregory Crane, *The Latin Dependency Treebank in a cultural heritage digital library*, in *Proceedings of the Workshop on Language Technology for Cultural Heritage Data (LaTeCH 2007)*, edited by Caroline Sporleder, Antal van den Bosch, Claire Grover, New York, Association for Computational Linguistics (ACL), 2007 (June), pp. 33-40.

[3] Flavio Massimiliano Cecchini, Timo Korkiakangas, Marco Passarotti, *A New Latin Treebank for Universal Dependencies: Charters between Ancient Latin and Romance Languages*, in *Proceedings of the 12th Language Resources and Evaluation Conference*, edited by Nicoletta Calzolari *et alii*, European Language Resources Association (ELRA), 2020 (May), pp. 933-942.

Questo articolo descrive i processi condotti per realizzare le due suddette attività. Nello specifico, la Sezione 2 descriverà la Base di Conoscenza di LiLa, introducendo i principî alla base della sua architettura e le risorse linguistiche attualmente allacciate, per poi presentare dettagliatamente UDante. La Sezione 3 sarà invece dedicata alla procedura di allacciamento di UDante a LiLa dando particolare attenzione al lessico dantesco. La Sezione 4 mostrerà esempi di ricerca nella Base di Conoscenza evidenziando come l'adozione del paradigma *Linked Data* possa facilitare l'uso proficuo di risorse linguistiche di tipo diverso. Infine, la Sezione 5 conterrà le conclusioni.

2. La Knowledge base di LiLa e UDante

2. 1. *LiLa*

Grazie a numerose attività di ricerca e a progetti svolti nel corso dei passati decenni a livello internazionale, sono attualmente disponibili molte risorse linguistiche e strumenti di Trattamento Automatico del Linguaggio (TAL) per il latino. Tuttavia, tali risorse e strumenti sono isolati gli uni dagli altri in quanto raccolti su siti diversi e basati su schemi di annotazione differenti: questa condizione impedisce loro di essere veramente d'aiuto ad una vasta comunità di storici, filologi, archeologi e letterati che, a vario titolo, si occupano della lingua latina. Il progetto LiLa è nato proprio per risolvere questo problema riducendo l'isolamento di risorse e strumenti per il latino attraverso la costruzione di una Base di Conoscenza secondo il paradigma *Linked Data*. La Base di Conoscenza di LiLa è fortemente lessicalizzata: questa scelta si basa sulla constatazione che i corpora sono formati da occorrenze di parole, le risorse lessicali descrivono proprietà di parole e gli strumenti TAL elaborano automaticamente parole. Vale a dire che tutte le fonti di (meta)dati linguistici hanno a che fare con parole che possono essere lemmatizzate. Sulla base di questo assunto, LiLa usa i lemmi come nodo centrale, come punto di connessione tra risorse lessicali, corpora annotati e strumenti TAL. Dal punto di vista operativo, il nucleo della Base di Conoscenza è costituito da un'ampia raccolta di lemmi latini chiamata

Lemma Bank: l'interoperabilità si ottiene collegando ai lemmi della *Lemma Bank* tutte le entrate delle risorse lessicali, le parole dei corpora e degli output di strumenti TAL che puntano allo stesso lemma. La lemmatizzazione, quindi, assume un ruolo centrale: le risorse, se non contengono già informazioni sui lemmi, vengono lemmatizzate prima di essere allacciate alla Base di Conoscenza.[1] Tali risorse sono interrogabili in due modi: ricerche sui singoli lemmi facenti parte della *Lemma Bank* sono possibili attraverso un'interfaccia grafica,[2] mentre ricerche più complesse che coinvolgono le risorse interoperabili sono effettuabili usando il linguaggio di interrogazione SPARQL.[3]

Al momento attuale, le risorse connesse in LiLa coprono vari aspetti linguistici (dalla morfologia alla sintassi e alla semantica) e periodi temporali diversi (dall'epoca classica al medioevo). Nello specifico, le risorse lessicali sono:

- un lessico di valenza, chiamato *Latin Vallex*,[4] in cui di ogni entrata si elencano i ruoli semantici;

[1] I sistemi che si occupano di lemmatizzazione automatica raggiungono alti livelli di accuratezza, superiori al 95%. Tuttavia, le prestazioni dei sistemi TAL basati su algoritmi di apprendimento automatico dipendono molto dai dati di addestramento usati per cui l'accuratezza tende a diminuire quando i sistemi vengono applicati a testi di un genere o di un periodo temporale diverso da quello dei testi usati in fase di addestramento. Questo problema è particolarmente rilevante in latino data la sua ampia variabilità diacronica, diatopica e di genere. I risultati della campagna di valutazione EvaLatin 2020 forniscono un'aggiornata panoramica dello stato dell'arte. Si veda Rachele Sprugnoli, Marco Passarotti, Flavio Massimiliano Cecchini, Matteo Pellegrini, *Overview of the EvaLatin 2020 evaluation campaign*, in *Proceedings of LT4HALA 2020-1st Workshop on Language Technologies for Historical and Ancient Languages*, European Language Resources Association (ELRA), 2020 (May), pp. 105-110.

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

[3] https://lila-erc.eu/sparql/.

[4] Francesco Mambrini, Marco Passarotti, Eleonora Litta Modignani Picozzi, Giovanni Moretti, *Interlinking Valency Frames and WordNet Synsets in the LiLa Knowledge Base of Linguistic Resources for Latin*, in *Further with Knowledge Graphs. Proceedings of the 17th International Conference on Semantic Systems, 6-9 September 2021*, Amsterdam, IOS Press, 2021, pp. 16-28. Marco Passarotti, Berta Gonzalez Saavedra, Cristophe Ledoux Onambele Manga, *Latin vallex. A treebank-based semantic valency lexicon for Latin*, in *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, European Language Resources Association (ELRA), 2016 (May), pp. 2599-2606.

- un lessico, *LatinAffectus*,[1] in cui ogni entrata è associata ad una polarità positiva, negativa o neutra al di là del suo contesto di uso;

- una raccolta di prestiti dal greco antico tratti dall'*Index Graecorum Vocabulorum in Linguam Latinam Translatorum*;[2]

- una base di dati che raccoglie informazioni derivazionali denominata *Word Formation Latin*[3] in cui ogni lemma è analizzato nei suoi componenti formativi spiegandone le regole di formazione;

- una raccolta di forme ricostruite del proto-italico e proto-indoeuropeo estratte dall'*Etymological Dictionary of Latin and the other Italic Languages*;[4]

- il dizionario bilingue latino-inglese curato da Ch. T. Lewis e Ch. Short[5] e pubblicato nel 1879;

[1] Rachele Sprugnoli, Marco Passarotti, Daniela Corbetta, Andrea Peverelli, *Odi et Amo. Creating, Evaluating and Extending Sentiment Lexicons for Latin*, in *Proceedings of the 12th Language Resources and Evaluation Conference*, 2020 (May), European Language Resources Association (elra), pp. 3078-3086.

[2] Greta Franzini, Federica Zampedri, Marco Passarotti, Francesco Mambrini, Giovanni Moretti, *Græcissâre: Ancient Greek Loanwords in the LiLa Knowledge Base of Linguistic Resources for Latin*, in *Seventh Italian Conference on Computational Linguistics*, Torino, Accademia University Press, 2020, pp. 1-6. Günther Alexander Ernst Adolf Saalfeld, *Index Graecorum vocabulorum in linguam Latinam translatorum quaestiunculis* auctus apud F. Berggold, Berlin, 1874.

[3] Matteo Pellegrini, Eleonora Litta Modignani Picozzi, Marco Passarotti, Francesco Mambrini, Giovanni Moretti, *The Two Approaches to Word Formation in the LiLa Knowledge Base of Latin Resources*, in *Proceedings of the Third International Workshop on Resources and Tools for Derivational Morphology (DeriMo 2021)*, Nancy, France, atilf (cnrs-Université de Lorraine), 2021, pp. 101-109.

[4] Francesco Mambrini, Marco Passarotti, *Representing etymology in the LiLa knowledge base of linguistic resources for Latin*, in *Proceedings of the 2020 Globalex Workshop on Linked Lexicography*, European Language Resources Association (elra), 2020 (May), pp. 20-28. Michiel De Vaan, *Etymological Dictionary of Latin and the other Italic languages*, vol. 7, Leiden-Boston, Brill, 2008.

[5] Francesco Mambrini, Eleonora Litta Modignani Picozzi, Marco Passarotti, Paolo Ruffolo, *Linking the Lewis & Short Dictionary to the LiLa Knowledge Base of Interoperable Linguistic Resources for Latin*, in *Eighth Italian Conference on Computational Linguistics*, Torino, Accademia University Press, 2021. Charlton T. Lewis, Charles Short, *A Latin Dictionary. Founded on Andrews' edition of Freund's Latin dictionary*, Oxford, Clarendon Press, 1879.

- un gruppo di entrate di *Latin WordNet*[1] manualmente ricontrollate in cui i lemmi sono collegati tra di loro in relazioni semantiche e riuniti in gruppi sinonimici denominati *synset*.

Le risorse testuali allacciate sono, invece, l'*Index Thomisticus Treebank* (ITTB) contenente il testo della *Summa contra Gentiles* di Tommaso D'Aquino, la commedia di autore ignoto *Querolus sive Aulularia*, l'ottavo capitolo del *Liber Abaci*, trattato matematico di Fibonacci, e UDante. Tutti questi corpora sono annotati seguendo lo schema UD: tra queste, però, solo l'ITTB e UDante contengono anche l'annotazione sintattica, oltre alla lemmatizzazione e all'attribuzione delle parti del discorso. La prossima sezione descriverà la creazione di UDante fornendo esempi di annotazione dei testi danteschi.

2. 2. *UDante*

UDante è una risorsa liberamente disponibile che contiene i testi latini di Dante Alighieri annotati a vari livelli da quattro annotatrici. Ogni opera è stata divisa in token e in frasi, a ogni token è stata aggiunta informazione riguardante il lemma, la parte del discorso, i tratti morfologici e le relazioni di dipendenza sintattica rispetto agli altri token componenti ciascuna frase. Nella sezione 4 si propongono due esempi di interrogazione che fanno uso dell'annotazione fornita da UDante e del suo allacciamento a LiLa, al fine di mostrarne il potenziale nell'estrazione di informazione linguistica dai testi del corpus.

I testi facenti parte di UDante sono stati tratti da *DanteSearch*, corpus delle opere volgari e latine di Dante Alighieri lemmatizzate e annotate grammaticalmente dall'Università di Pisa e distri-

[1] Greta Franzini, Andrea Peverelli, Paolo Ruffolo, Marco Passarotti, Helena Sanna, Edoardo Signoroni, Federica Zampedri, *Nunc Est Aestimandum: Towards an Evaluation of the Latin WordNet*, in *Sixth Italian Conference on Computational Linguistics (CLiC-it 2019)*, Torino, Accademia University Press, 2019, pp. 1-8. Stefano Minozzi, *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 Mastandrea, Venezia, Edizioni Ca' Foscari, 2017 («Antichistica», 14), pp. 123-134.

buite in formato TEI-XML. I file XML sono stati automaticamente trasformati nel formato standard dell'iniziativa UD denominato CoNLL-U convertendo, allo stesso tempo, i codici di annotazione usati in *DanteSearch* in quelli utilizzati da UD. Per quanto riguarda le parti del discorso, la mappatura tra i due schemi di annotazione non è stata biunivoca in quanto a uno stesso codice di *DanteSearch* può corrispondere più di un'etichetta di UD. Ad esempio, il codice "a" usato per gli aggettivi corrisponde ad ADJ (*adjective*) nel caso di aggettivi attributivi e indeclinabili, a NUM (*numeral*) per gli aggettivi numerali e DET (*determiner*) per quelli pronominali, possessivi, determinativi, indefiniti e interrogativi. Anche la mappatura dei codici relativi ai tratti morfologici non si è rivelata immediata perché l'annotazione di UD tende a essere più granulare rispetto a quella originale. Per esempio, il codice morfologico del participio presente "pp" corrisponde a cinque tratti morfologici in UD rappresentati dalla stringa che segue:

· Aspect=Imp|Degree=Pos|Tense=Pres|VerbForm=Part|Voice=Act

In altre parole, in UD viene specificato che l'aspetto è imperfetto, il grado è positivo, il tempo è presente, la forma verbale è quella del participio e la diatesi è attiva. Data la complessità della conversione, il risultato della trasformazione automatica è stato controllato manualmente dalle annotatrici durante l'annotazione sintattica.

In base all'approccio UD, la struttura sintattica delle frasi in UDante ha uno stile di annotazione a dipendenze. Ciò significa che ciascuna frase è rappresentata da grafi aciclici detti 'alberi' i cui nodi si orientano a partire da una radice (*root*). La gerarchia è stabilita in base a un approccio che pone al centro il predicato, da cui dipendono i complementi senza distinzione fra argomenti e aggiunti (informazione che, come si vedrà, è recuperabile a livello di sottorelazioni). I nodi di ciascun albero sono costituiti da parole o segni di interpunzione con una differenza sostanziale tra parole semanticamente piene o categorematiche e parole sincategorematiche o funzionali che tendono a occupare nodi 'foglia' ovvero che non sono testa di alcuna relazione sintattica. A ciascu-

na relazione di dipendenza è assegnata un'etichetta che ne descrive la funzione sintattica (*dependency relation*: *deprel*). Nell'esempio in Fig. 1 è riportato l'albero sintattico di *De vulgari eloquentia* (d'ora in poi DVE) I IV 1: «Soli homini datum fuit ut loqueretur, ut ex premissis manifestum est.».[1]

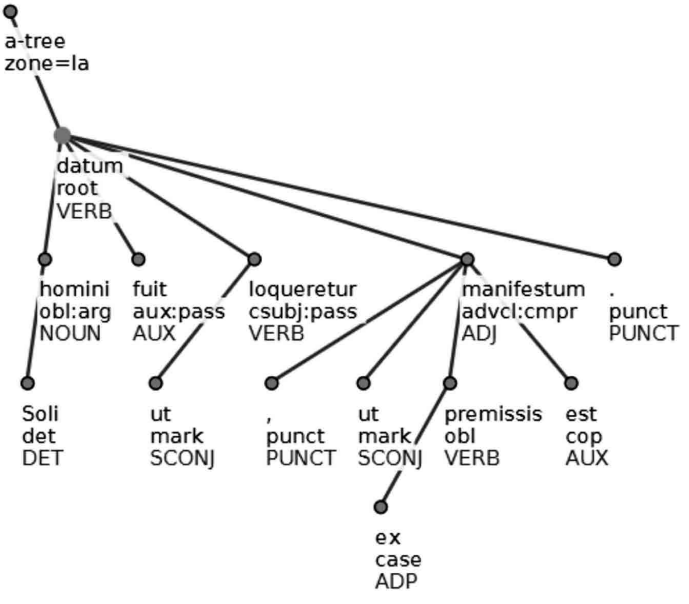

Fig. 1. Albero sintattico della frase «Soli homini datum fuit ut loqueretur, ut ex premissis manifestum est.».

Dopo un nodo puramente formale che identifica la frase nella *treebank* entro l'area dedicata al latino nel sistema di visualizzazione (a-tree zone = la[tin]), l'albero inizia con un predicato verbale formato da un perfetto passivo di tipo perifrastico, comune nel latino tardo (*datum fuit*): il participio perfetto del verbo *do* costituisce la radice (*root*) del grafo, mentre il perfetto indicativo

[1] «Solo all'uomo fu dato di parlare, come è chiaro per le ragioni dette sopra». Le frasi del *De vulgari eloquentia* negli esempi in Fig. 1 e Fig. 2 con relativa traduzione sono tratte dall'edizione di riferimento di *DanteSearch*, ovvero Dante Alighieri, *De Vulgari Eloquentia*, a cura di Mirko Tavoni, in Idem, *Opere*, 1, direzione di Marco Santagata, Milano, Mondadori, 2011 («I Meridiani»), pp. 1067-1547, in questo caso alle pp. 1154-1155.

del verbo *sum* dipende da esso (come nodo 'foglia') ed è annotato con la *deprel* aux:pass (ausiliare passivo).

Il complemento di termine *homini*, a sua volta testa di un sintagma nominale a cui si lega il pronome *soli* come modificatore di tipo determinante (det), è espresso dalla relazione obl (*oblique*). Tale relazione è assegnata ai complementi obliqui e, in questo caso, è specificata dalla sottorelazione :arg che esprime la natura argomentale del complemento *homini soli*.

Il soggetto del predicato *datum fuit* è una proposizione soggettiva espressa tramite la *deprel* csubj (*clausal subject*, ovvero soggetto frasale) con testa *loqueretur* e precisata a sua volta da :pass in quanto soggetto di una frase passiva. La soggettiva è introdotta dalla congiunzione subordinativa *ut* che dipende dalla testa verbale con la *deprel* mark.

Allo stesso modo, una proposizione extranucleare (la comparativa «ut ex premissis manifestum est») si connette alla *root* come modificatore avverbiale frasale (advcl, *adverbial clause modifier*) con la specifica :cmpr (*comparative*). La testa della proposizione subordinata è costituita dalla testa del predicato nominale, che in questo caso è un aggettivo (*manifestum*) da cui dipende *est* in funzione di copula (cop), ed è introdotta dalla congiunzione subordinativa *ut* con la relazione mark.

Da *manifestum* dipende anche un complemento obliquo con testa *premissis* (che, non essendo un argomento, non è corredato della specifica :arg) ed è introdotto da *ex*, nodo dipendente da *premissis* attraverso la relazione case. Infine, si nota che la punteggiatura (punct) si lega al nodo di grado più alto possibile senza creare casi di proiettività, ovvero di incrocio fra gli archi di un albero.

Un secondo esempio è DVE I XI 5: «Post quos Mediolanenses atque Pergameos eorumque finitimos eruncemus, in quorum etiam improperium quendam cecinisse recolimus Enter l'ora del vesper, ciò fu del mes d'ochiover.».[1]

[1] «Dopo questi sradichiamo i milanesi e i bergamaschi e i loro vicini; anche su di loro ricordiamo che un tale ha composto versi di scherno: 'Enter l'ora del vesper, ciò fu del mes d'ochiover'», vedi ALIGHIERI, *De vulgari eloquentia*, cit., p. 1259.

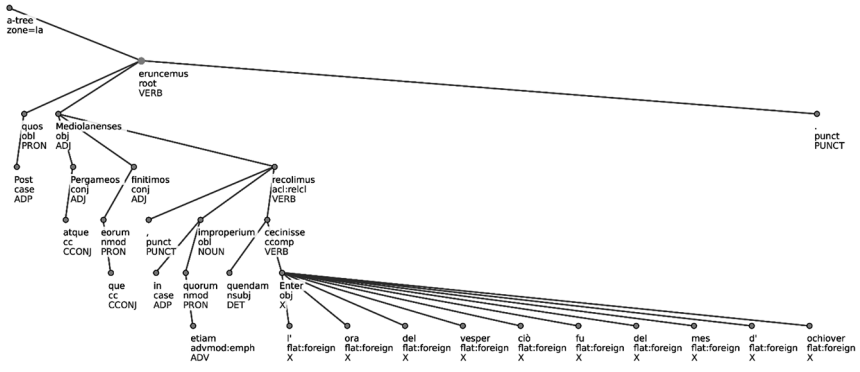

Fig. 2. Albero sintattico della frase «Post quos Mediolanenses atque Pergameos eorumque finitimos eruncemus, in quorum etiam improperium quendam cecinisse recolimus Enter l'ora del vesper, ciò fu del mes d'ochiover.».

L'albero si sviluppa da *eruncemus* (root) che regge il complemento indiretto «post quos» e tre oggetti diretti (relazione obj) fra loro coordinati e dipendenti dal primo in ordine di apparizione (*Mediolanenses*). Le congiunzioni coordinative (*atque* e l'enclitica *-que*) dipendono attraverso la relazione cc (*coordinative conjunction*) dagli elementi che introducono, cioè, rispettivamente, da *Pergameos* e da *eorum*, modificatore nominale (nmod) del complemento oggetto *finitimos*.

La subordinata con testa *recolimus* è una proposizione relativa che dipende direttamente dall'elemento nominale da essa modificato (in questo caso *Mediolanenses*) attraverso la relazione acl:relcl (*adnominal clause: relative clause*). Da questa discendono: (a) un complemento indiretto «in quorum etiam improperium» con testa nominale (*improperium*) da cui dipendono una preposizione (*in*) e un modificatore nominale (*quorum*) rafforzato da un avverbio con funzione enfatica (relazione advmod:emph); (b) una proposizione oggettiva con soggetto indipendente (*quendam*) governata dal predicato *cecinisse* cui è associata la relazione ccomp (*clausal complement*).

In accordo con le norme UD,[1] l'oggettiva con testa *cecinisse*

[1] https://universaldependencies.org/u/dep/flat.html.

regge a sua volta un oggetto diretto (obj) che consiste in un'intera citazione in volgare rappresentata nell'albero in modo 'piatto' ovvero non gerarchico, con il primo elemento in ordine di apparizione che funge da testa mentre il resto delle parole straniere sono ad esso legate dalla relazione flat:foreign.

La FIG. 3 riporta la rappresentazione della frase in questione nel formato CoNLL-U.[1]

```
# sent_id = DVE-124
# text = Post quos Mediolanenses atque Pergameos eorumque finitimos eruncemus, in quorum etiam improperium quendam cecinisse recolimus Enter l' ora del vesper ciò
fu del mes d' ochiover.
# citation_hierarchy = Liber_Primus,xi,Paragraphus_5
1 Post          post          ADP   e       AdpType=Prep                                                                                  2  case        _ _
2 quos          qui           PRON  prepma  Case=Acc|Gender=Masc|InflClass=LatPron|Number=Plur|PronType=Rel                               9  obl         _ _
3 Mediolanenses mediolanensis ADJ   Smp3a   Case=Acc|Gender=Masc|InflClass=IndEurI|NameType=Nat|Number=Plur                               9  obj         _ _
4 atque         atque         CCONJ co      Emphatic=Yes                                                                                  5  cc          _ _
5 Pergameos     pergameus     ADJ   Smp2a   Case=Acc|Gender=Masc|InflClass=IndEurO|NameType=Nat|Number=Plur                               3  conj        _ _
6-7 eorumque    _             _     _       _                                                                                             _  _           _ _
6 eorum         is            PRON  ddepmg  Case=Gen|Gender=Masc|InflClass=LatPron|Number=Plur|Person=3|PronType=Prs                     8  nmod        _ _
7 que           que           CCONJ co9     Clitic=Yes                                                                                    6  cc          _ _
8 finitimos     finitimus     ADJ   smp2a   Case=Acc|Gender=Masc|InflClass=IndEurO|Number=Plur                                           3  conj        _ _
9 eruncemus     erunco        VERB  va1cpp1 Aspect=Imp|InflClass=LatA|Mood=Sub|Number=Plur|Person=1|Tense=Pres|VerbForm=Fin|Voice=Act 0  root        _ SpaceAfter=No
10 ,            ,             PUNCT Pu      _                                                                                             17 punct       _ _
11 in           in            ADP   e       AdpType=Prep                                                                                  14 case        _ _
12 quorum       qui           PRON  prepmg  Case=Gen|Gender=Masc|InflClass=LatPron|Number=Plur|PronType=Rel                               14 nmod        _ _
13 etiam        etiam         ADV   co      Compound=Yes                                                                                  12 advmod:emph _ _
14 improperium  improperium   NOUN  sns2a   Case=Acc|Gender=Neut|InflClass=IndEurO|Number=Sing                                           17 obl         _ _
15 quendam      quidam        DET   dinsma  Case=Acc|Gender=Masc|InflClass=LatPron|Number=Sing|PronType=Ind                               16 nsubj       _ _
16 cecinisse    cano          VERB  va3fr   Aspect=Perf|InflClass=LatX|InflClass[noun]=Ind|Tense=Past|VerbForm=Inf|Voice=Act            17 ccomp       _ _
17 recolimus    recolo        VERB  va3ipp1 Aspect=Imp|InflClass=LatX|Mood=Ind|Number=Plur|Person=1|Tense=Pres|VerbForm=Fin|Voice=Act 3  acl:relcl    _ _
18 Enter        enter         X     zi      Foreign=Yes                                                                                   16 obj         _ _
19 l'           l             X     zi      Foreign=Yes                                                                                   18 flat:foreign _ _
20 ora          ora           X     zi      Foreign=Yes                                                                                   18 flat:foreign _ _
21 del          del           X     zi      Foreign=Yes                                                                                   18 flat:foreign _ _
22 vesper       uesper        X     zi      Foreign=Yes                                                                                   18 flat:foreign _ _
23 ciò          cio           X     zi      Foreign=Yes                                                                                   18 flat:foreign _ _
24 fu           fu            X     zi      Foreign=Yes                                                                                   18 flat:foreign _ _
25 del          del           X     zi      Foreign=Yes                                                                                   18 flat:foreign _ _
26 mes          mes           X     zi      Foreign=Yes                                                                                   18 flat:foreign _ _
27 d'           d             X     zi      Foreign=Yes                                                                                   18 flat:foreign _ _
28 ochiover     ochiouer      X     zi      Foreign=Yes                                                                                   18 flat:foreign _ SpaceAfter=No
29 .            .             PUNCT Pu      _                                                                                             9  punct       _ _
```

FIG. 3. Rappresentazione in formato CoNLL-U della frase «Post quos Mediolanenses atque Pergameos eorumque finitimos eruncemus, in quorum etiam improperium quendam cecinisse recolimus Enter l'ora del vesper, ciò fu del mes d'ochiover.».

Le prime tre linee, introdotte dal cancelletto (#), sono di commento: la prima a partire dall'alto (#sent_id = DVE-124) identifica la posizione della frase nell'opera dantesca (in questo caso DVE per indicare il *De vulgari eloquentia*). La seconda riga (#text) riproduce il testo così come esso appare nell'edizione di riferimento di *DanteSearch*, mentre la terza e ultima riga di commento (# citation_hierarchy) fornisce i riferimenti interni all'opera, indicando in questo caso libro, capitolo e paragrafo.

[1] Si tratta di una rielaborazione del formato CONLL-X, vedi SABINE BUCHHOLZ, ERWIN MARSI, *CoNLL-X shared task on Multilingual Dependency Parsing*, in *Proceedings of the Tenth Conference on Computational Natural Language Learning CoNLL-X* (New York, 8-9 giugno 2006), a cura di L. Márquez, D. Klein, New York, Association for Computational Linguistics (ACL), 2006, pp. 149-164.

Ciascuna riga non introdotta dal cancelletto è dedicata a un *token*, ovvero l'elemento lessicale sintattico corrispondente a un nodo nell'albero a dipendenze.[1] Nelle colonne, separate da tabulazioni, da sinistra a destra si leggono: 1) il numero che indica la posizione del *token* all'interno della frase; 2) il singolo *token*; 3) il lemma; 4) la parte del discorso indicata con una serie di etichette (le *Universal POS tags*)[2] secondo le norme di UD; 5) i tratti morfologici attribuiti al token in *DanteSearch*; 6) i tratti morfologici; 7) il numero identificativo della testa sintattica da cui il *token* dipende (che è 0 nel caso della *root*); 8) la *deprel*, ovvero l'etichetta che descrive la relazione del *token* con la sua testa sintattica; 9) eventuali dipendenze estese dedicate all'annotazione, ad esempio, delle strutture ellittiche, e delle coreferenze (campo sempre vuoto) e 10) uno spazio per osservazioni miscellanee.

Fa eccezione la riga dedicata alla parola *eorumque*, che nell'albero corrisponde a due nodi (cioè a due *token*): *eorum* e *que*. Il valore *range* 6-7 indica che nel testo *eorumque* corrisponde alla somma dei *token* registrati nel file CoNNL-U rispettivamente con il numero della posizione 6 e 7.

3. L'allacciamento di UDante a LiLa

Il primo passo per l'allacciamento dei lemmi di UDante alla Base di Conoscenza di LiLa è stato di tipo completamente automatico: è stata cercata una corrispondenza perfetta di stringhe tra i lemmi dei testi e quelli nella *Lemma Bank* considerando anche la parte del discorso. Per esempio, la prima parola della prima epistola è *Reverendissimo* il cui lemma è *reverendus* a cui è attribuita la parte del discorso ADJ (aggettivo). Nella *Lemma Bank* ci sono però due lemmi con la stessa rappresentazione grafica (*reverendus*): uno è un aggettivo (https://lila-erc.eu/data/id/lem-

[1] Oltre ai nodi lessicali, negli alberi UD sono presenti anche nodi per i segni d'interpunzione, che sono dunque registrati come token nel formato CoNLL-U.

[2] Slav Petrov, Dipanjan Das, Ryan McDonald, *A Universal Part-of-Speech Tagset*, in *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, European Language Resources Association (ELRA), 2012, pp. 2089-2096.

ma/76728) e l'altro è un nome proprio (https://lila-erc.eu/data/id/lemma/20112). Tenendo in considerazione la parte del discorso annotata in UDante, il lemma del testo viene connesso senza ambiguità al primo e non al secondo. Questo approccio ha permesso di allacciare in modo diretto la maggioranza dei lemmi: la percentuale media di allacciamento è stata dell'80% ma con alcune variazioni tra le opere (dal 77% del *DVE* all'84% delle *Epistole*).

I lemmi restanti ricadevano in due classi: alcuni erano lemmi ambigui, cioè collegabili a più di un lemma nella *Lemma Bank*, mentre altri erano lemmi non presenti nella Base di Conoscenza.

In media, la percentuale di lemmi ambigui nelle opere dantesche di UDante è stata dell'8% ma con una punta del 12% registrata nelle *Egloghe*. La maggior parte di questi lemmi sono nomi comuni e verbi: alcuni sono stati disambiguati controllando le caratteristiche morfologiche mentre per altri è stato necessario analizzare lo specifico contesto d'uso. Nel primo gruppo ricadono, ad esempio, verbi come *dico* che hanno due possibili corrispondenze nella *Lemma Bank* aventi due categorie flessionali diverse: *dico*, infatti, può essere un verbo di prima o di terza coniugazione. In questo caso la disambiguazione è avvenuta confrontando l'annotazione delle caratteristiche morfologiche in UDante con la categoria flessionale dei lemmi nella *Lemma Bank*. Lemmi come *frons* hanno, invece, avuto bisogno di essere controllati nelle loro singole occorrenze all'interno del contesto. *Frons*, infatti, può essere connesso a due lemmi aventi la stessa parte del discorso (nome comune) e la stessa categoria flessionale (terza declinazione con genitivo plurale in *-um*, *-ium*) ma con significato diverso: 'fronte' o 'fronda'. Nel *DVE* le occorrenze si collegano tutte al primo lemma (https://lila-erc.eu/data/id/lemma/103845) facendo riferimento alla struttura delle stanze nelle canzoni, mentre nelle *Egloghe*, componimento bucolico, sono allacciate al secondo (https://lila-erc.eu/data/id/lemma/103844).

Per quanto riguarda i lemmi non presenti nella *Lemma Bank*, la loro percentuale media è dell'11% sul totale dei lemmi. L'analisi di questo insieme di lemmi dialoga direttamente con gli studi sul lessico latino di Dante, recentemente aggiornati dall'avvio dei

lavori per il V*ocabolario Dantesco Latino* (VDL) parallelo al vocabolario volgare (VD) patrocinato dall'Accademia della Crusca,[1] e si svolge al netto delle questioni di progettazione specifiche della *Lemma Bank*. Infatti, alcune tipologie di lemmi formalmente introdotti in LiLa da UDante sono dovute a scelte di modellizzazione in situazioni di ambiguità lessicale e morfologica: è il caso delle *lemma variant*, ovvero dei lemmi che differiscono per almeno un tratto morfologico e che si è scelto di considerare come forme rappresentative poste sullo stesso piano. Un esempio di questa tipologia è dato dall'oscillazione fra la diatesi attiva e passiva in verbi come *caupono* (*Monarchia*), *haereditor* (DVE) e *peregrino* (*Epistole*), oppure da variazioni nella flessione nominale in sostantivi come *endecadis*, *grandinis*, *uibices* (DVE).

Un caso analogo è dato dalle *written representation*, ovvero le varianti grafico-fonetiche di lemmi già presenti in LiLa come *prosaicus*, *prosaycus* e *gramatice*, *grammatica* e *gramatica*: a prescindere dalla situazione testuale delle opere dantesche, che non consente di fare affermazioni sulla veste linguistica originaria, queste varianti sono il risultato di scelte lessicografiche e non sono significative dal punto di vista lessicale. Infine, la grande quantità di avverbi immessi da UDante in LiLa è un dato da filtrare considerando che per ragioni di carattere pratico la Lemma Bank include tutti gli avverbi deaggettivali in *-e* e in *-(it)er*, ma non quelli prodotti con altre strategie (ad esempio gli avverbi in *-o* e in *-um* del tipo *multo* e *tantum*). Questi meccanismi derivazionali non sono generalizzabili, per cui si è deciso di introdurre i singoli avverbi su base *corpus-driven*, ovvero alla loro prima occorrenza in un corpus allacciato a LiLa, includendoli come ipolemmi avverbiali dell'aggettivo corrispondente (es. *tantum* è ipolemma di *tantus*).[2]

[1] Online all'indirizzo http://www.vocabolariodantescolatino.it/ (VDL) e http://www.vocabolariodantesco.it/ (VD).

[2] Per le definizioni di *lemma variant*, *written representation* e *hypolemma*, vedi Francesco Mambrini, Flavio Maissimiliano Cecchini, Greta Franzini, Eleonora Litta Modignani Picozzi, Marco Carlo Passarotti, Paolo Ruffolo, *LiLa: Linking Latin. Risorse linguistiche per il latino nel Semantic Web*, «Umanistica Digitale», VIII, 2020, pp. 63-78: 69-70.

Per quanto riguarda i restanti lemmi introdotti nella *Lemma Bank*, tra le categorie più folte ci sono quelle dei sostantivi (182) e degli aggettivi (129), di cui la maggior parte si distribuisce nel DVE e nella *Monarchia*. La metà esatta dei sostantivi (91) è costituita da nomi propri, con una suddivisione piuttosto netta fra nomi di poeti e letterati nel DVE (*Brunectus*, *Caualcantis*, *Guinizelli*, *Guitto*, *Sordellus*) e nomi di personaggi della storia romana (*Clelia*, *Curiatii*, *Fabritius*, *Pirrus*, *Porsenna*) e di autorità in campo scientifico (*Galienus*, *Ptolomeus*) nella *Monarchia*.

Dal punto di vista morfologico derivazionale, si ha una grande quantità di femminili astratti della terza declinazione, specialmente in *-tio*, *-tionis* come *congremiatio* (DVE), *depauperatio* (*Monarchia*), *repatriatio* (*Epistole*) e in *-tas*, *-tatis* come *curialitas* (DVE), *ecentricitas* (*Questio*), *inequitas* (*Monarchia*), mentre passano in secondo piano i sostantivi della prima declinazione. Al secondo posto i neutri, spesso indicanti elementi tipici della metrica in volgare o parti di componimento come *decasillabum*, *endecasillabum*, *eptasillabum*, *neasillabum*, *parysillabum*, *preludium* e *proemium* nel DVE, oppure aggettivi sostantivati tratti dal lessico aristotelico-scolastico quali *differenziale*, *directivum*, *logicale*, *minerale* e *perfectivum* nella *Monarchia*.

Fra gli aggettivi prevalgono quelli a tre uscite della seconda declinazione e fra gli aggettivi della terza spiccano quelli in *-alis* di estrazione scientifico-scolastica, come ad esempio *emisperialis* (*Questio*), *equinoctialis* (*Monarchia*) e *punctalis* (*Epistole*). Sono inoltre estremamente frequenti i nomi etnici, la cui categoria grammaticale è stata modificata in fase di annotazione della *treebank* in accordo con le norme di UD.[1] Sul totale di 50 nomi di popolazioni, 40 si trovano nel DVE e precisamente nel I libro, che sviluppa la ricerca del volgare illustre attraverso la geografia linguistica dell'Italia. Infatti, si tratta per la maggior parte di suddivisioni geografiche interne al territorio italiano eccetto: *anglici*, *hispani*,

[1] FLAVIO MASSIMILIANO CECCHINI, MARCO PASSAROTTI, RACHELE SPRUGNOLI, *UDante: First Steps Towards the Universal Dependencies Treebank of Dante's Latin Works*, in *Proceedings of the Seventh Italian Conference on Computational Linguistics, CLiC-it 2020* (Bologna, March 1-3, 2021), «CEUR Workshop Proceedings», 2021, pp. 1-7.

sclavones, *troiani* e *ungari*. La restante parte dei nomi di popoli si distribuisce fra *Monarchia* ed *Epistole*, dove in proporzione i popoli stranieri costituiscono la maggioranza.

Più in generale, al di fuori di questa tipologia specifica, si nota lo scarso apporto della *Monarchia*, che introduce 12 aggettivi a fronte dei 13 derivanti dalle *Egloghe*, opera di estensione molto più ridotta, mentre il DVE continua sulla scia dei nomi etnici con voci come *cremonensis*, *estensis*, *florentinus*, *lombardus*, *longobardus*, *paduanus*, *romandiolus*, *sicilianus*, *trivisanus* ed etichette relative al volgare italiano oggetto della ricerca quali *latius* e *semilatius*. Quest'ultimo in particolare è neoformazione dantesca e hapax a partire da *semi-*, prefisso di origine greca, e *latius*, che prosegue l'uso poetico latino di indicare per sineddoche l'Italia attraverso il Lazio, e riunisce i volgari della parte sinistra della penisola dividendola lungo l'Appennino.

Un'altra categoria degna di nota sono proprio i composti aggettivali e nominali che costituiscono anche neologismi, soprattutto nel DVE e nelle *Egloghe*. Nel primo caso spiccano i composti di *loquor* come *turpiloquium*, neosemia che stigmatizza i volgari toscani, e *tristiloquium*, neologismo hapax modellato sul primo e riferito al volgare romano. Entrambi denunciano la forma linguistica oscena delle parlate non degne di essere elevate al volgare comune di tutta Italia e si distaccano dalla tradizione cristiana, che riferisce il turpiloquio al contenuto del discorso. L'altrettanto raro *primiloquium* indica il primo atto di parola concesso da Dio ad Adamo e, sebbene possa essere stato formalmente modellato sui primi due, se ne distacca diventando una parola chiave della ricerca dantesca in senso positivo: visto che il tratto semantico pregnante di questo neologismo è l'anteriorità temporale, può essere rilevante l'esempio di *antiloquium*, che Uguccione da Pisa definisce «prima locutio», locuzione peraltro utilizzata da Dante a DVE I V 3 come sinonimo di *primiloquium*.[1]

Ulteriori neoformazioni composizionali del DVE sono l'hapax *benegenitus*, riferito a Manfredi a DVE I XII 4 in stretta correlazio-

[1] UGUCCIONE DA PISA, *Derivationes*, edizione critica princeps a cura di Enzo Cecchini *et alii*, Tavernuzze, SISMEL-Edizioni del Galluzzo, 2004, L 97, 5.

ne con il rarissimo sostantivo *heros*, che scopre la linea politica del trattato con una lode sperticata dell'Imperatore,[1] e l'aggettivo *astripetus*, ripreso nel sostantivo *astricola* nella prima egloga dantesca. Le *Egloghe* sono infatti la seconda fonte principale di composti considerando che *laetifluus* è in realtà neoformazione delvirgiliana dovuta alla prima epistola metrica inviata a Dante. Pienamente danteschi sono invece i due composti aggettivali *vatificus* e *vatisonus*, entrambi connessi a un nodo concettuale bilingue che in Dante è volto a definire il ruolo del poeta in volgare.[2] Entrano così in LiLa i neologismi del latino dantesco, una tipologia che comprende avverbi (*abmotim*, *contatim*, *contemptive*, *superficietenus*, tutti nel DVE), aggettivi (*abvius*, *adiuvalis*, *comicomus* e *cantionarius*, quest'ultimo dovuto a una probabile corruzione del testo di DVE II VI 1),[3] sostantivi (*conticentia*, *fastigiositas*, *nequitatrix*, *scatescentia*, *sinistratio*, *trisillabitas* e *metaphorismus*) e soprattutto verbi, spesso tratti direttamente dalla lessicografia medievale come *avieo*, *confecto*, *prosayco* e *returgeo*. Tali hapax, non vitali in latino ma trasparenti quanto al loro significato etimologico, in Dante possono diventare parole chiave dell'argomentazione proprio in virtù della loro natura lessicografica. È il caso di *avieo* e *prosayco*, legati nello stretto giro di frase di DVE II I 1 in relazione al diverso ruolo degli scrittori di poesia e di prosa nella costruzione della lingua letteraria.[4]

Dal punto di vista storico linguistico sono inoltre significativi i grecismi, fra cui *athlotes*, *coathleta*, *epyikia*, *eubulia* e *gignasium*, tutti dalla *Monarchia,* e i verbi in *-izo armonizo* (DVE), *athletizo*, *politizo* (*Monarchia*), e i volgarismi di stampo retorico del DVE come *ballata*, *sirma* e, con tutta probabilità, anche *montaninus*, possibile neoformazione per diminuzione da *montanus* per cui tuttavia non si può escludere il calco da *montanino*, che solo in volgare è attestato nel significato dantesco 'di montagna'.[5]

[1] MIRKO TAVONI, *Qualche idea su Dante*, Bologna, il Mulino, 2015, pp. 39-40.

[2] IDEM, *Il nome di poeta in Dante*, in *Studi offerti a Luigi Blasucci dai colleghi e dagli allievi pisani*, a cura di Lucio Lugnani, Marco Santagata, Alfredo Stussi, Lucca, Pacini Fazzi, 1996, pp. 545-577, ora in IDEM, *Qualche idea su Dante*, cit., pp. 295-327.

[3] ALIGHIERI, *De vulgari eloquentia*, cit., pp. 1435.

[4] Vedi *avieo* in VDL.

[5] Vedi *montaninus* in VDL.

Grazie all'allacciamento alla Base di Conoscenza, i lemmi delle opere dantesche si trovano inserite in un ecosistema di risorse che fornisce loro informazioni linguistiche aggiuntive. La Fig. 4 mostra in alto la sequenza di token che aprono l'*Epistola* VI (*Dantes Alagherii florentinus et exul inmeritus*...); ognuno di questi token è collegato al corrispettivo lemma nella *Lemma Bank* il quale è descritto da varie proprietà morfo-sintattiche. Ad esempio, *exul* (https://lila-erc.eu/data/id/lemma/102480) è un nome comune della terza declinazione di genere maschile e femminile che appartiene alla stessa famiglia derivazionale di altri 9 lemmi tra cui l'aggettivo *exulaticius*. I lemmi appartenenti alla stessa famiglia derivazionale sono connessi tra di loro attraverso un nodo specifico (in Fig. 4 ha l'etichetta *Base of exul*). Sia *exul* che *exulaticius* sono registrati nella Base di Conoscenza con una rappresentazione grafica alternativa: *exsul* e *exsulaticius*. Il lemma *exul* è a sua volta collegato alle entrate lessicali di quattro risorse. L'*Etymological Dictionary of Latin and the other Italic Languages* ci restituisce la ricostruzione della forma protoitalica di *exul* (*ek()s()Vl); *LatinAffectus* assegna una polarità negativa al lemma; il lessico di morfologia derivazionale, *Word Formation Latin,* descrive come da *exul* derivino altri tre lemmi (*exilium*, *exularis* e *exulo*/*exulor*); il dizionario bilingue *Lewis & Short* indica i tre concetti evocati dal lemma dandone una definizione.

4. Ricerche tra risorse interoperabili

In questa sezione proponiamo due esempi di ricerca nella Base di Conoscenza di LiLa per mostrare l'utilità dell'interoperabilità tra risorse linguistiche.[1] Il primo caso d'uso interroga contemporaneamente la *Lemma Bank* e UDante allo scopo di estrarre il lessico usato da Dante solo nel DVE e non nelle altre sue opere latine.

[1] Le interrogazioni sono espresse con il linguaggio SPARQL e sono state effettuate usando l'endpoint di LiLa. Per permettere la replicabilità delle ricerche riportate in questa sezione, le interrogazioni sono state caricate nel repository disponibile all'indirizzo https://github.com/CIRCSE/SPARQL-queries: si tratta di distinctivelammas-DVE.rq e di UDante-amon-nmod-loquor.rq.

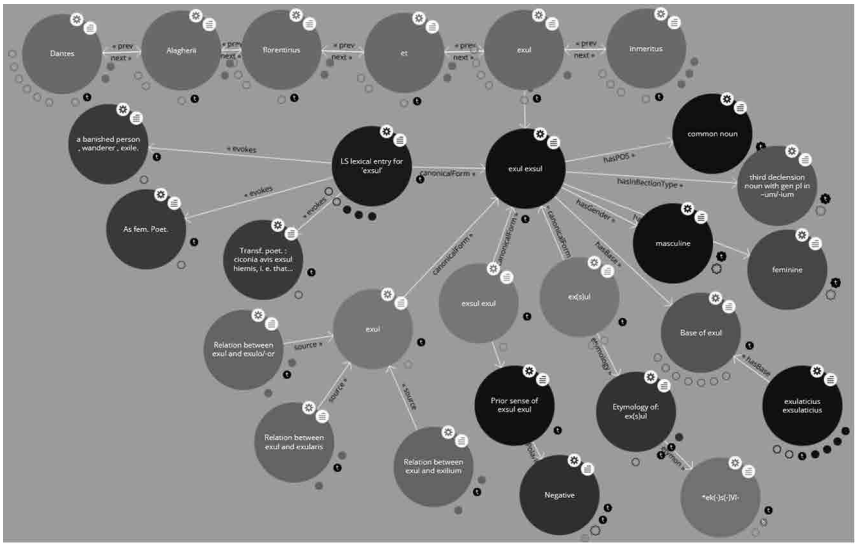

Fig. 4. Allacciamento di una occorrenza testuale a LiLa.

LEMMA	FREQ	LEMMA	FREQ
Cantio	65	*eptasillabum*	15
Stantia	30	*constructio*	14
Hendecasyllabum	18	*loquela*	14
Quare	18	*latius*	12
Syllaba	18	*desinentia*	11
Habitudo	17	*profero*	11
Poetor	16	*curialis*	10

Tab. 1. Lemmi propri del solo DVE (frequenza ≥ 10).

Come si vede nella Tab. 1, quasi tutti i 14 lemmi con frequenza pari o superiore a 10 sono parole del lessico linguistico e retorico. Infatti, nel DVE Dante si pone in dialogo con la tradizione medievale dell'*Ars Dictandi*, innovandola con l'applicazione alla retorica in volgare dei termini tecnici normalmente riservati al latino e per mezzo di slittamenti semantici e pretti volgarismi.

L'hapax relativo più diffuso nel DVE è *cantio*, che Dante intende

in due modi: o come *nomen actionis* che identifica l'atto del comporre poesia, e dunque il componimento poetico a prescindere dalla sua forma, oppure la specifica «forma lirica strofica composta di stanze aventi la stessa formula sillabica e lo stesso schema di rime comune alla poesia provenzale e italiana».[1] In questo senso, maggioritario nel secondo libro del trattato, la parola costituisce una risemantizzazione dal volgare che nella definizione dantesca esclude, ma al tempo stesso riconosce altre possibili accezioni del latino *cantio*, ad esempio in quanto 'componimento cantato' e come sinonimo di *modulatio*.

La *stantia* è l'«unità metrica della canzone e della ballata, definita dal numero dei versi, dalla misura del verso che occupa ogni posizione e dallo schema delle rime»[2] e nel DVE viene ulteriormente precisata come volgarismo tecnico a partire dal latino medievale *stantia*, già volgarismo nel significato di 'camera'.[3]

Le voci *hendecasyllabum* ed *eptasillabum* indicano i due tipi di verso propri della versificazione in stile alto, la cui diversa disposizione (*habitudo*) determina la partizione della melodia verbale all'interno della *stantia*. La frequenza della voce *syllaba* e la discussione sulla *desinentia* delle parole in rima nel capitolo II XIII è un'ulteriore spia del fatto che l'opera introduce una sistemazione organica della metrica accentuativa in volgare. Di qui discende anche l'altissima frequenza di *poetor*, verbo altrimenti molto raro che Dante usa solo in riferimento ai poeti volgari in lingua di sì e in lingua d'hoc tranne che a DVE II IV 3, dove questi vengono affiancati ai poeti latini. Nell'uso dantesco si può inoltre individuare una gerarchia fra i poeti in volgare illustre e in latino, entrambi

[1] Vedi TLIO s.v. canzone 1.

[2] TLIO s.v. stanza (1) 4. Il significato metrico è raro anche in volgare, vedi TLIO s.v. stanza (1) 0.6, ma trova riscontro nel volgare dantesco, ad esempio in *Vita Nova* XIX 9: «Poscia quando dico: Canzone, io so che tu, aggiungo una stanza quasi come ancella de l'altre».

[3] DU CANGE *et alii*, *Glossarium mediae et infimae latinitatis*, Niort, L. Favre, 1883-1887 s.v. *stantia* 1 e FRANCESCO ARNALDI, PASQUALE SMIRAGLIA, *Latinitatis Italicae Medii Aevi Lexicon (saec. vex. - saec. XIIn.)*. Editio altera aucta addendis quae confecerunt L. Celentano, A. De Prisco, A. V. Nazzaro, I. Polara, P. Smiraglia, M. Turriani, Firenze, SISMEL-Edizioni del Galluzzo, 2001, s.v. *stantia*.

soggetti del verbo *poetor*, e i rimatori volgari in generale, soggetti del verbo *versificor*, che non ha in sé la radice tecnica della poesia definita a DVE II IV 2 come «fictio rethorica musicaque poita».[1]

La seconda interrogazione va oltre l'ambito puramente lessicale per sfruttare l'annotazione sintattica presente sia in UDante che nell'ITTB. Nello specifico, si interrogano contemporaneamente le due risorse testuali e la *Lemma Bank* per estrarre le relazioni di dipendenza amod (*adjectival modifier*) e nmod (*nominal modifier*) che coinvolgono i nomi comuni appartenenti alla famiglia derivazionale di *loquor* fungendone da testa o da dipendenza. Per esempio, in *Monarchia* III IV 11: «Nam quanquam scribe divini eloquii multi sint» *divini* è modificatore aggettivale di *eloquii* e quest'ultimo è modificatore nominale di *scribe*. La ricerca sulla Base di Conoscenza permette di estrarre le coppie come *divini-eloquii* e *eloquii-scribe*.

Le coppie estratte sono in numero esiguo nell'ITTB (8), mentre in Dante troviamo una varietà più ampia con 45 coppie testa-modificatore che, oltre ad *eloquium*, *locutio* e *loquela*, in comune con l'ITTB, coinvolgono anche *eloquentia*.

In Tommaso d'Aquino si nota il riferimento alla sfera della parola divina, cui fanno capo *locutio transumptiva* come espressione figurativa tipica delle Scritture,[2] *eloquium sacrum*, *modus locutionis* (che nella *Summa contra Gentiles* indica il modo in cui le verità divine sono rivelate) e infine *gratia locutionis* e *usus loquelae* in riferimento alla concessione del linguaggio da Dio all'uomo proprio allo scopo di accedere a tali verità.

In Dante *locutio* è principalmente usata all'interno della definizione contrastiva dei due tipi di espressione linguistica che fanno capo alla facoltà di linguaggio esclusivamente umana: del volgare come *locutio vulgaris* e del latino come *locutio secundaria* o *gramatica*. Il significato di *nomen rei actae* che indica il risultato dell'azione, ovvero l'espressione, la parola, il discorso, è attestato

[1] TAVONI, *Qualche idea su Dante*, cit., pp. 295-334 e cfr. *poetor* in VDL.

[2] ROY JOSEPH DEFERRARI, IGNATIUS MCGUINESS, BARRY, M. INVIOLATA, *A lexicon of St. Thomas Aquinas based on the Summa theologica and selected passages of his other works*, Washington, Catholic University of America Press, 1948, p. 646 b.

in locuzioni come *antiqua locutio*, *Ytalie locutio*, *pulcra locutio* (riferito al volgare bolognese a DVE I XV 2) e si sovrappone parzialmente alla semantica di *loquela*, che in Dante indica la specifica lingua (*Ytalie loquela*, *tuscana loquela*, *nova loquela*, *optima loquela*).[1] Il piano retoricamente codificato della *locutio* è l'*eloquentia*, che si attesta soltanto nel DVE in relazione al volgare (*vulgaris eloquentia* e *vir eloquentiae* per indicare Sordello). Infine, la parola di Dio è espressa in Dante con il ricorso a *eloquium* modificato dall'aggettivo *divinus,* significativamente al di fuori del DVE, che sviluppa il pensiero linguistico sul piano razionale.

Dal punto di vista degli elementi che modificano i sostantivi della famiglia morfologica di *loquor*, in Dante si osserva una tipologia di amod di tipo valutativo (*decens*, *optimus*, *pulcher*, *purus*, *verus*) e classificatorio (*antiquus*, *maternus*, *novus*, *primus*, *secundarius, tuscanus*). Analogamente, fra i sostantivi che i deverbali in questione modificano come nmod si individuano elementi formali (*forma*, *signum*), strutturali (*exordium, pars*) ed epistemologici *(doctrina*, *ratio, ydemptitas*) esclusivi del DVE che evidenziano la fine ricerca linguistica che Dante approfondisce nel trattato.

5. Conclusioni

In questo articolo abbiamo descritto il processo di sviluppo della *treebank* UDante, che raccoglie i testi latini di Dante Alighieri tratti dal corpus *DanteSearch*, arricchendoli con un'annotazione sintattica a dipendenze in accordo con i criteri adottati dallo stile di UD. Inoltre, l'articolo presenta il lavoro di allacciamento di UDante alla *Knowledge Base* di risorse latine interoperabili LiLa, che ne consente l'interazione via web con altri corpora, lessici e dizionari per la lingua latina attraverso l'adozione dei principî del paradigma *Linked Data* e, nello specifico, di ontologie condivise sviluppate per la rappresentazione di (meta)dati linguistici.

Questi due lavori hanno innestato i testi latini di Dante nello stato dell'arte dei corpora sintattici e nella rete interoperabile delle risorse linguistiche per il latino, ponendo così le condizioni

[1] Cfr. *locutio* e *loquela* in VDL.

per la piena valorizzazione dei (meta)dati registrati nel corpus. Infatti, dopo una lunga fase dedicata alla costruzione di risorse linguistiche fondamentali per diverse lingue e diverse loro varietà testuali (stilistiche, cronologiche, tipologiche), nel corso degli ultimi due decenni è maturata nella comunità scientifica la convinzione di (almeno) tre necessità: primo, raccogliere le risorse in *repository* condivisi, superandone così la dispersione; secondo, arricchirle con metadati descrittivi e livelli di annotazione linguistica in accordo con criteri e *tagset* condivisi; terzo, renderle interoperabili sul web tramite il ricorso a modelli comuni di descrizione della conoscenza.

Oggi l'infrastruttura europea CLARIN[1] fa fronte alla prima di queste necessità, rappresentando il punto di raccolta e accesso di migliaia di risorse linguistiche di diverso tipo per centinaia di lingue.[2] A propria volta, UDante nasce con il proposito di far fronte alla seconda esigenza, trasformando un corpus originariamente annotato con uno stile proprietario e autoreferenziale in una risorsa che si accorda ai criteri di altre centinaia del medesimo tipo. Infine, l'inclusione di UDante in LiLa risponde al terzo bisogno, ponendo la *treebank* dei testi latini di Dante in una relazione aperta e interoperabile con altre risorse linguistiche del latino.

Si tratta, dunque, di un lavoro di progressivo svincolamento dei (meta)dati di *DanteSearch* dai confini della singola risorsa nativa, per dare loro cittadinanza in un ecosistema aperto, costituito da metodi, modelli e formati condivisi, in cui essi possano interagire con molte altre risorse. Oggi la tecnologia connessa al paradigma *Linked Data*, mutuata dal web semantico, e i modelli ontologici che la comunità scientifica sta elaborando per la rappresentazione dei dati linguistici consentono di automatizzare, rendere replicabile ed estendere sia quantitativamente che qualitativamente il trattamento dei testi antichi (e del lessico che essi veicolano) che rappresentano quell'evidenza empirica che è l'essenziale

[1] https://www.clarin.eu.

[2] Nella collezione CIRCSE di CLARIN (https://dspace-clarin-it.ilc.cnr.it/repository/xmlui/handle/000-co-111/525) sono in fase di pubblicazione i dati sia di UDante che della sua serializzazione RDF (Turtle).

supporto a qualsiasi analisi minuziosa del dato testuale specifico. Proprio nel trattamento e nella conseguente analisi e interpretazione dei dati testuali risiede il fertile punto d'incontro tra la metodologia tradizionale di ricerca linguistica, letteraria e filologica, attenta al dettaglio e altamente competente su di esso, e l'approccio computazionale, che pone nelle mani del ricercatore umanista non solo una quantità di evidenza empirica mai prima d'ora disponibile, ma anche un alto grado di qualità di estrazione di essa da risorse di diverso tipo e la replicabilità del processo del suo trattamento.

Ringraziamenti

Gli autori ringraziano Flavio Massimiliano Cecchini, Daniela Corbetta, Federica Favero, Federica Gamba, Martina de Laurentiis, Andrea Peverelli ed Elena Vagnoni.

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