SOCIETÀ DANTESCA ITALIANA

STUDI DANTESCHI

Fondati da Michele Barbi

Pubblicati dalla Società Dantesca Italiana

LXXXVI

PER IL CENTENARIO DANTESCO
(1321-2021)

IN FIRENZE, LE LETTERE – 2021

INDICE

PER IL CENTENARIO DANTESCO (1321-2021)

SAGGI

NOTE
Il *Vocabolario Dantesco Latino* (*VDL*): primi risultati

PER IL CENTENARIO DANTESCO
(1321-2021)

NOTE

IL *VOCABOLARIO DANTESCO LATINO* (*VDL*): PRIMI RISULTATI

Marco Passarotti - Flavio Massimiliano Cecchini
Rachele Sprugnoli - Giovanni Moretti

UDante. L'annotazione sintattica dei testi latini di Dante

L'articolo descrive il lavoro di realizzazione di *UDante*, il corpus dei testi latini di Dante Alighieri annotato a livello sintattico in base ai criteri stabiliti dall'iniziativa *Universal Dependencies*. Dopo avere introdotto e motivato lo stile di annotazione adottato, l'articolo presenta nel dettaglio le fasi di costruzione di *UDante*, soffermandosi particolarmente sul processo di conversione del formato dei dati e sulla loro annotazione manuale. Viene, quindi, descritta l'integrazione dei testi di *UDante* nella *knowledge base* di *LiLa*, grazie a cui il corpus sintattico dei testi latini di Dante è reso interoperabile con altre risorse linguistiche per il latino. Infine, alcuni esempi d'interrogazione di *UDante* sono riportati con l'obiettivo di dimostrarne l'utilità in termini di supporto alla compilazione del *Vocabolario Dantesco Latino*.

"UDante". Syntactic Annotation of Dante Alighieri's Latin Texts

This paper describes the process of building *UDante*, a corpus that collects the Latin texts of Dante Alighieri enhanced with syntactic annotation according to the criteria established by the *Universal Dependencies* initiative. After introducing the annotation style adopted, the article details the phases of development of *UDante*, particularly focusing on the process of conversion of the data format, as well as on their manual annotation. The inclusion of *UDante* in the *LiLa knowledge base*, which makes the corpus interoperable with other linguistic resources for Latin, is then described. Finally, some examples of queries run on *UDante* are presented, to show how the corpus can support the creation of the *Vocabolario Dantesco Latino*.

Keywords: Dante Alighieri; Latin; Syntax; Treebanks; Linguistic Linked Data.

1. *Introduzione*

Corpora testuali, concordanze, lessici, thesauri e dizionari sono parte del lavoro quotidiano di chiunque faccia ricerca in ambito umanistico e, specificamente, linguistico, letterario e filologico.

A partire dagli anni Sessanta del Novecento, quando venne pubblicato il *Brown Corpus* della lingua inglese,[1] questi strumenti hanno

[1] http://korpus.uib.no/icame/manuals/BROWN/INDEX.HTM. W.N. Francis - H. Kučera, *Manual of information to accompany a Standard Sample of Present-day Edit-*

cominciato a essere disponibili su supporto elettronico. Tale disponibilità è esplosa negli anni Novanta grazie alla capillare diffusione della tecnologia digitale e al sempre più facile ed economico accesso a computer con prestazioni crescenti.

È in quegli anni che i detti strumenti iniziano a essere nominati con il termine, introdotto da Antonio Zampolli, «risorse linguistiche». Nell'ambito della linguistica computazionale, gli ultimi due decenni hanno visto una crescita sostanziale di progetti mirati alla costruzione delle risorse linguistiche essenziali per numerose lingue, riducendo sempre più il numero di quelle dotate di risorse insufficienti.

Se il primario interesse che ha sospinto il notevole sviluppo dell'area di ricerca dedicata alle risorse linguistiche è stato di tipo linguistico-computazionale per fini di trattamento automatico del linguaggio, la raccolta e la gestione dei dati testuali rappresenta una fase essenziale anche (e soprattutto) per chi si occupa di lingue antiche. L'assenza di parlanti nativi, unitamente alla limitata disponibilità di evidenza testuale, endemica quando si tratta di lingue morte, impone infatti che l'analisi e l'indagine delle lingue antiche siano necessariamente condotte sulla base dei dati che di quelle lingue sono arrivati a noi attraverso i secoli. Fare linguistica delle lingue antiche significa, dunque, fare linguistica dei *corpora*. Non stupisce, quindi, che tra i primi testi organizzati in *corpora* registrati su supporto elettronico ci siano stati quelli, in latino, di Tommaso d'Aquino, raccolti nel *corpus* dell'*Index Thomisticus* dal gesuita Roberto Busa,[2] o che una delle prime biblioteche digitali sia stata *Perseus*, avviata negli anni Ottanta proprio per mettere a disposizione in formato digitale testi in greco antico e in latino.[3]

Nello specifico delle lingue classiche, un tipo di risorsa linguistica sul cui sviluppo si è particolarmente concentrata la ricerca nel corso degli ultimi quindici anni sono stati i *corpora* arricchiti con annotazione sintattica, le cosiddette *treebank*. Il lavoro in quest'area iniziò nel 2006, quando i primi due progetti mirati alla realizzazione di *treebank* di te-

---

ed American English, for use with digital computers, Providence, Department of Linguistics of Brown University, 1979.

[2] R. 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.

[3] http://www.perseus.tufts.edu/.

sti latini furono avviati: la *Index Thomisticus Treebank* (IT-TB),[4] basata sui testi di Tommaso d'Aquino, e la *Latin Dependency Treebank* (LDT),[5] che raccoglie testi classici tratti dalla *Perseus Digital Library*. Fin dall'inizio, le due *treebank* condivisero i medesimi criteri di annotazione, che si sarebbero poi imposti come un riferimento nell'area dell'annotazione sintattica delle lingue antiche.[6]

La ricerca dedicata al *treebanking* è attualmente molto vitale. In particolare, nel corso dell'ultimo quinquennio, si è distinto per il proprio successo nel settore un progetto, nominato *Universal Dependencies* (UD),[7] dedicato alla raccolta di *treebank* annotate secondo i medesimi criteri.[8] Oggi UD fa fronte al problema non solo della dispersione,

---

[4] M. PASSAROTTI, *The Project of the Index Thomisticus Treebank*, in *Digital Classical Philology*, a c. di M. BERTI, Berlin-Boston, De Gruyter, 2019, pp. 299-320. Dato che nell'articolo sono presenti numerosi acronimi, al fine di facilitare il lettore se ne riporta qui di seguito la lista completa, con il relativo scioglimento: CoNLL-U (*Computational Natural Language Learning – Universal*); ERC (*European Research Council*); IAA (*Inter-Annotator Agreement*); IT-TB (*Index Thomisticus Treebank*); KB (*Knowledge Base*); LDT (*Latin Dependency Treebank*); LLCT (*Late Latin Charter Treebank*); LLOD (*Linguistic Linked Data*); PDT (*Prague Dependency Treebank*); PML-TQ (*Prague Markup Language - Tree Query*); TEI-XML (*Text Encoding Initiative – eXtensible Markup Language*); UD (*Universal Dependencies*); VDL (*Vocabolario Dantesco Latino*).

[5] D. BAMMAN - G. CRANE, *The design and use of a Latin dependency treebank*, in Proceedings of the Fifth Workshop on Treebanks and Linguistic Theories TLT~2006 (Praha, 1-2 dicembre 2006), a c. di J. HAJIČ, J. NIVRE, Praha, Istituto di Linguistica Applicata e Formale (UFÁL), 2006, pp. 67-78.

[6] D. BAMMAN - M. PASSAROTTI - R. BUSA - G. CRANE, *The annotation guidelines of the Latin Dependency Treebank and Index Thomisticus Treebank. The treatment of some specific syntactic constructions in Latin*, in Proceedings of the Sixth International Conference on Language Resources and Evaluation LREC'08 (Marrakech, 28-30 maggio 2008), a c. di N. CALZOLARI, K. CHOUKRI, B. MAEGAARD, J. MARIANI, J. ODIJK, S. PIPERIDIS, D. TAPIAS, Marrakech, European Language Resources Association (ELRA), 2008, pp. 71-76.

[7] https://universaldependencies.org.

[8] J. NIVRE - M.C. DE MARNEFFE - F. GINTER - Y. GOLDBERG - J. HAJIČ - C.D. MANNING - R. MCDONALD - S. PETROV - S. PYYSALO - N. SILVEIRA - R. TSARFATY - D. ZEMAN, *Universal Dependencies v1: A multilingual treebank collection*, in Proceedings of the Tenth International Conference on Language Resources and Evaluation LREC'16 (Portorož, 23-28 maggio 2016), a c. di N. CALZOLARI, K. CHOUKRI, T. DECLERCK, S. GOGGI, M. GROBELNIK, B. MAEGAARD, J. MARIANI, H. MAZO, A. MORENO, J. ODIJK, S. PIPERIDIS, Portorož, European Language Resources Association (ELRA), 2016, pp. 1659-1666.

ma soprattutto della divergenza annotazionale tra le *treebank* di diverse lingue (o, addirittura, di diversi progetti per la medesima lingua), fornendo un unico luogo e formato di pubblicazione dei (meta)dati delle risorse, oltre che un insieme di strumenti per l'interrogazione e la validazione dei contenuti delle *treebank*, e un comune manuale di annotazione e insieme di etichette per la registrazione delle parti del discorso, dei tratti morfologici e delle funzioni sintattiche. A cadenza regolare di sei mesi, viene pubblicata una nuova versione della raccolta di *treebank* di UD; la più recente, distribuita il 15 novembre 2021, include 227 *treebank*, che coprono 122 lingue, tra cui il greco antico, rappresentato da due *treebank*,[9] e il latino (cinque *treebank*).[10] Nonostante l'attuale disponibilità di ben sette *treebank* in UD per queste lingue, la loro copertura è lungi dall'essere non solo esaustiva, ma anche minimamente rappresentativa della testualità greca e latina. Nel caso specifico del latino, ad esempio, le altre quattro *treebank* di UD includono solo alcuni testi letterari classici e post-classici, documentari del-

[9] *Ancient Greek and Latin Dependency Treebank*: circa 200 000 parole (G.G.A. CELANO, *The Dependency Treebanks for Ancient Greek and Latin*, in *Digital Classical Philology*, a c. di M. BERTI, Berlin-Boston, De Gruyter, 2019, pp. 279-98). *Corpus* PROIEL: circa 210 000 parole (D.T.T. HAUG - M. JØHNDAL, *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 (Marrakech, 1° giugno 2008), a c. di C. SPORLEDER, K. RIBAROV, A. VAN DEN BOSCH, M.P. DOBREVA, M.J. DRISCOLL, C. GROVER, P. LENDVAI, A. LÜDELING, M. PASSAROTTI, Marrakech, European Language Resources Association (ELRA), 2008, pp. 27-34).

[10] Oltre a UDante, IT-TB: circa 450 000 parole tratte dai testi di Tommaso d'Aquino (PASSAROTTI, *The Project of the Index Thomisticus Treebank*, cit.). LDT-*Perseus*: circa 29 000 parole di testi di latino classico (BAMMAN-CRANE, *The design and use of a Latin dependency treebank*, cit.). *Corpus* PROIEL: circa 200 000 parole in testi della *Vulgata* e di latino classico e post-classico (HAUG-JØHNDAL, *Creating a parallel treebank*, cit.). *Late Latin Charter Treebank* (LLCT): circa 250 000 parole tratte da *chartulae* notarili toscane dell'VIII secolo (T. KORKIAKANGAS - M. PASSAROTTI, *Challenges in annotating medieval Latin charters*, in «Journal for Language Technology and Computational Linguistics», 26 (2011), 2, pp. 103-114; F.M. CECCHINI - T. KORKIAKANGAS - M. PASSAROTTI, *A New Latin Treebank for Universal Dependencies: Charters between Ancient Latin and Romance Language*, in Proceedings of the Twelfth International Conference on Language Resources and Evaluation LREC'20 (Marseille, 13-15 maggio 2020), a c. di N. CALZOLARI, F. BÉCHET, P. BLACHE, K. CHOUKRI, C. CIERI, T. DECLERCK, S. GOGGI, H. ISAHARA, B. MAEGAARD, J. MARIANI, H. MAZO, A. MORENO, J. ODIJK, S. PIPERIDIS, Marseille, European Language Resources Association (ELRA), 2020, pp. 933-942).

l'alto medioevo e filosofici tardo-medievali. In totale, si tratta di poco meno di un milione di parole, a fronte di un ammontare complessivo di circa 200 milioni presenti nel *corpus* dei testi latini e greci fino all'anno 600 d.C., un totale approssimativo calcolato dal progetto *Open Greek and Latin*,[11] che pure non considera il periodo successivo al VII sec. d.C. e, in particolare, i testi della letteratura neolatina (da Petrarca ai tempi moderni), il cui *corpus* di riferimento, CAMENA, raccoglie circa 50 milioni di parole.[12]

Le condizioni che precedono lo sviluppo di una nuova *treebank* sono diverse: si va dal caso estremo di testi non ancora digitalizzati fino a quello in cui i testi non solo sono già disponibili in formato digitale, ma sono anche arricchiti con metadati (ovvero, dati che descrivono dati) relativi a livelli di annotazione che precedono quello sintattico, nello specifico la lemmatizzazione e l'analisi morfologica. Questa è la condizione che sussiste per i testi latini di Dante forniti dal *corpus DanteSearch*.[13] A supporto della redazione delle entrate del nuovo *Vocabolario Dantesco Latino* (VDL)[14] e in occasione del 700° anniversario della morte del poeta, è stato avviato un progetto di collaborazione tra il VDL e il progetto *LiLa: Linking Latin*,[15] attivo presso il centro di ricerca CIRCSE dell'Università Cattolica del Sacro Cuore di Milano, con l'obiettivo di realizzare la *treebank* UD dei testi latini di Dante (nominata *UDante*), tramite l'arricchimento dei metadati di annotazione linguistica già forniti da *DanteSearch* con le relazioni sintattiche definite dalle regole di UD.

Dopo un capitolo (§2) dedicato alle grammatiche e agli stili di annotazione sintattica adottati nelle *treebank*, con una particolare attenzione ai principî di annotazione di UD, questo articolo descrive le fasi di realizzazione di *UDante* (§3) e l'integrazione della nuova *treebank* nella *knowledge base* di *LiLa*, che ne consente l'interoperabilità con altre risorse linguistiche per il latino (§4). Inoltre, l'articolo presenta al-

---

[11] https://opengreekandlatin.org.

[12] http://mateo.uni-mannheim.de/camenahtdocs/camena_e.html.

[13] https://dantesearch.dantenetwork.it. M. 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*, a c. di A. CERBO, M. SEMOLA, Napoli, Il Torcoliere-Officine Grafico-Editoriali d'Ateneo, 2011, pp. 583-608.

[14] http://www.vocabolariodantesco.it.

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

cuni esempi d'interrogazione di *UDante* per dimostrarne l'utilità in termini di supporto alla compilazione del VDL (§5). Infine, sono riportate alcuni brevi considerazioni conclusive (§6).[16]

2. *Annotazione linguistica di corpora testuali*

Il processo di annotazione linguistica di una risorsa lessicale o testuale consiste nell'arricchimento dei suoi dati attraverso l'inserimento di metadati linguistici, generalmente organizzati in termini di livelli di analisi linguistica. Nell'ambito specifico delle risorse testuali (*corpora*), i principali livelli di annotazione linguistica sono usualmente i seguenti:[17]

- annotazione morfologica: attribuzione a ciascuna occorrenza testuale della parte del discorso e degli eventuali tratti morfologici (es. genere, numero, modo, tempo, persona, grado, diatesi);
- lemmatizzazione: attribuzione a ciascuna occorrenza testuale della sua forma di citazione convenzionale (lemma);
- annotazione sintattica: registrazione delle relazioni sintattiche intra-frasali e attribuzione delle funzioni sintattiche alle parole e/o ai sintagmi;
- annotazione semantica: include un'ampia gamma di tipi di annotazione, tra cui la disambiguazione del senso delle parole in contesto, la registrazione della struttura retorica e/o informativa (ad esempio, le relazioni tematico-rematiche), la risoluzione delle coreferenze (ad esempio, quelle pronominali) e l'attribuzione di etichette di ruoli semantici (come 'Agente', 'Paziente' e 'Beneficiario').

La realizzazione di *UDante*, basata sui testi forniti da *DanteSearch*, già dotati di annotazione morfologica e lemmatizzazione, è consistita nell'arricchimento dei testi con i metadati sintattici. Ciò ha richiesto di

---

[16] La responsabilità principale dei singoli capitoli e paragrafi va attribuita come segue. Marco Passarotti: §1, §2, §2.1, §4 e §4.1. Marco Passarotti e Giovanni Moretti: §5. Flavio Massimiliano Cecchini: §2.2. Flavio Massimiliano Cecchini e Rachele Sprugnoli: §3. Rachele Sprugnoli: §4.2. Le conclusioni (§6) sono da ascrivere in pari misura a tutti gli autori.

[17] Maggiori dettagli sull'annotazione linguistica sono forniti dall'*Handbook of Linguistic Annotation*, a c. di N. IDE, J. PUSTEJOVSKY, Dordrecht, Springer, 2007.

assumere decisioni innanzitutto in merito al tipo di grammatica da utilizzare per rappresentare la sintassi delle frasi e, più in dettaglio, a quale stile di annotazione conformarsi.

2.1. *Grammatiche e stili di annotazione sintattica*

A partire dalla pubblicazione di *Syntactic Structures* di Noam Chomsky,[18] le cosiddette grammatiche a costituenti hanno dominato per decenni l'area d'indagine linguistica dedicata alla sintassi. Non a caso, le prime *treebank* disponibili furono la *IBM/Lancaster Treebank*[19] e la *Penn Treebank*,[20] entrambe contenenti testi in lingua inglese e annotate secondo stili a costituenti.

La situazione iniziò a cambiare intorno alla metà degli anni Novanta, in concomitanza con l'avvio della stagione dedicata alla produzione delle risorse linguistiche fondamentali per un sempre più ampio numero di lingue, quando il dominio delle grammatiche a costituenti venne intaccato dalla crescente adozione di stili di annotazione sintattica basati sulle grammatiche a dipendenze, il cui testo fondativo, gli *Éléments de syntaxe structurale* di Lucien Tesnière,[21] è sostanzialmente coevo a *Syntactic Structures*.[22] All'origine del successo delle grammatiche a dipendenze nello sviluppo di *treebank* sta sia la necessità di rappresentare la sintassi di lingue con caratteristiche diverse dall'inglese, come ad esempio la maggior libertà nell'ordine delle parole, che le grammatiche a costituenti sono meno adatte a trattare, sia il maggior livello di accuratezza garantito da strumenti stocastici di analisi automatica addestrati su strutture a dipendenze che su strutture a costituenti.

---

[18] N. CHOMSKY, *Syntactic Structures*, 's-Gravenhage, Mouton & Co., 1957.

[19] E.W. BLACK - R. GARSIDE - G.N. LEECH, *Statistically-driven computer grammars of English: The IBM/Lancaster approach*, Amsterdam-Atlanta, Rodopi, 1993.

[20] M.P. MARCUS - M.A. MARCINKIEWICZ - B. SANTORINI, *Building a large annotated corpus of English: The Penn Treebank*, in «Computational linguistics», 19 (1993), 2, pp. 313-330.

[21] L. TESNIÈRE, *Éléments de syntaxe structurale*, Paris, Editions Klincksieck, 1959.

[22] Sulla svolta dalle grammatiche a costituenti alle grammatiche a dipendenze nella linguistica computazionale si veda M. PASSAROTTI, *Well, It Depends. Reflections on the Dependency Turn in Computational Linguistics*, in *Formal Representation and the Digital Humanities*, a c. di P. COTTICELLI-KURRAS, F. GIUSFREDI, Newcastle upon Tyne, Cambridge Scholars Publishing, 2018, pp. 141-158.

Pur nelle differenze che sussistono tra grammatiche a costituenti e grammatiche a dipendenze, un tratto comune a entrambe è la rappresentazione della struttura sintattica in termini di particolari grafi aciclici connessi detti 'alberi', in cui i nodi sono organizzati in modo 'orientato' a partire da una 'radice' da cui si sviluppa l'intero albero, e i cui archi (detti anche 'rami') rappresentano relazioni gerarchiche binarie fra nodi. Per definizione, nell'albero un nodo può avere un solo padre, mentre il numero di eventuali figli non è fissato. La principale differenza tra le due grammatiche consiste nella natura stessa dei nodi e delle relazioni.

Per quanto riguarda i nodi, negli alberi a dipendenze essi sono occupati da elementi lessicali o segni d'interpunzione. Negli alberi a costituenti, invece, i nodi 'foglia', ovvero senza relazioni di cui siano testa, rappresentano gli elementi lessicali della frase (e, non a caso, sono detti 'simboli terminali'), mentre i nodi interni dell'albero sono occupati da nomi di parti del discorso e di sintagmi ('simboli non terminali').

In merito alle relazioni, negli alberi a dipendenze esse sono dette 'dipendenze' in quanto la presenza del nodo-figlio è soggetta a quella del nodo-padre; in particolare, la gerarchia tra i nodi di un albero a dipendenze è stabilita alla luce dell'approccio predicato-centrico di siffatte grammatiche, che rappresentano innanzitutto le relazioni predicativo-argomentali affidando al predicato il ruolo di nodo-padre dei propri complementi (siano essi argomenti, o aggiunti). I rami che connettono tra loro i nodi degli alberi a costituenti, invece, rappresentano relazioni gerarchiche di categorizzazione tra nodi-figli, più specifici, come ad esempio una parola di una frase, e nodi-padre, che corrispondono a categorie più generiche, come il nome di un tipo di sintagma (es. sintagma nominale), fino a risalire alla radice dell'albero, identificata da un generico '*start symbol*'.

Per i fini di *UDante*, è stato scelto di rappresentare la struttura sintattica delle frasi dei testi attraverso uno stile di annotazione a dipendenze. Nello specifico, è stato adottato lo stile di UD, in quanto includere i testi latini di Dante in una così ampia raccolta di *treebank* annotate secondo il medesimo stile ne pone il *corpus* nel pieno dello stato dell'arte corrente nel settore dei *corpora* sintattici. Ciò consente di godere dei vantaggi di essere parte della comunità che ruota intorno a UD, tra cui l'utilizzo dei numerosi strumenti sviluppati nell'ambito di UD per costruire, visualizzare e validare gli alberi sintattici, oltre che per produrli automaticamente e interrogarli intra- e inter-linguisticamente, lanciando ricerche comuni su più di 180 *corpora* contemporaneamente. Inoltre, l'adozione di un formato dei dati e di etichette di parti del

discorso che rappresentano lo standard *de facto* nel settore dei *corpora* annotati favorisce la distribuzione e l'uso di *UDante* anche presso il mondo della linguistica computazionale, ponendo così le basi per una più stretta, e auspicabilmente fertile, collaborazione con la filologia italiana e, più ampiamente, con la ricerca in ambito umanistico.

2.2. *"Universal Dependencies": storia, principî ed esempi*

Il progetto *Universal Dependencies* punta allo sviluppo di un'annotazione morfosintattica coerente interlinguisticamente, da utilizzare per la creazione di *treebank* compatibili fra loro dal punto di vista formale e strutturale, e applicabile al maggior numero di lingue possibile.

Il punto di partenza del progetto è stata una rielaborazione delle dipendenze universali di Stanford,[23] delle parti del discorso universali proposte da *Google*[24] e del sistema di annotazione morfosintattico *Interset*.[25]

---

[23] M.C. DE MARNEFFE - B. MACCARTNEY - C.D. MANNING, *Generating typed dependency parses from phrase structure parses*, in Proceedings of the Fifth International Conference on Language Resources and Evaluation LREC'06 (Genova, 22-28 maggio 2006), a c. di N. CALZOLARI, K. CHOUKRI, A. GANGEMI, B. MAEGAARD, J. MARIANI, J. ODIJK, D. TAPIAS, Genova, European Language Resources Association (ELRA), 2006, pp. 449-454; M.C. DE MARNEFFE - C.D. MANNING, *The Stanford typed dependencies representation*, in Proceedings of the workshop on Cross-Framework and Cross-Domain Parser Evaluation at Coling 2008 (Manchester, 23 agosto 2008), a c. di J. BOS, E. BRISCOE, A. CAHILL, J. CARROLL, S. CLARK, A. COPESTAKE, D. FLICKINGER, J. VAN GENABITH, J. HOCKENMAIER, A. JOSHI, R. KAPLAN, T. HOLLOWAY KING, S. KÜBLER, D. LIN, J.T. LØNNING, C. MANNING, Y. MIYAO, J. NIVRE, S. OEPEN, K. SAGAE, N. XUE, Y. ZHANG, Manchester, Coling 2008 Organizing Committee, 2008, pp. 1-8; M.C. DE MARNEFFE - T. DOZAT - N. SILVEIRA - K. HAVERINEN - F. GINTER - J. NIVRE - C.D. MANNING, *Universal Stanford Dependencies: A cross-linguistic typology*, in Proceedings of the Ninth International Conference on Language Resources and Evaluation LREC'14 (Reykjavík, 26-31 maggio 2014), a c. di N. CALZOLARI, K. CHOUKRI, T. DECLERCK, H. LOFTSSON, B. MAEGAARD, J. MARIANI, A. MORENO, J. ODIJK, S. PIPERIDIS, Reykjavík, European Language Resources Association (ELRA), 2014, pp. 4585-4592.

[24] S. PETROV - D. DAS - R. MCDONALD, *A universal part-of-speech tagset*, in Proceedings of the Eighth International Conference on Language Resources and Evaluation LREC'12 (Istanbul, 21-27 maggio 2012), a c. di N. CALZOLARI, K. CHOUKRI, T. DECLERCK, M.U. DOĞAN,, B. MAEGAARD, J. MARIANI, A. MORENO, J. ODIJK, S. PIPERIDIS, Istanbul, European Language Resources Association (ELRA), 2012, pp. 2089-2096.

[25] D. ZEMAN, *Reusable Tagset Conversion Using Tagset Drivers*, in Proceedings of the Sixth International Conference on Language Resources and Evaluation LREC'08

Accanto a un'impostazione tipologica il cui obiettivo è l'individuazione di categorie morfologiche e relazioni di dipendenza sintattica universali, ciascuna lingua può arricchire questo schema di base facendo uso di sottocategorie specifiche capaci di rendere conto delle proprie peculiarità. Il vantaggio di questo sistema non è solo teorico, bensì ha anche ricadute relative allo sviluppo e alla valutazione comparata di 'strumenti universali' per il trattamento automatico del linguaggio, come annotatori automatici di parti del discorso (ing. *POS-tagger*), analizzatori morfologici o analizzatori sintattici (ing. *parser*).

L'accesso al progetto è aperto e libero: la discussione fra più di 300 collaboratori diretti, o chiunque sia semplicemente interessato, si svolge principalmente sulla piattaforma di sviluppo e distribuzione di software *GitHub*[26] e tramite eventi come il recente *Universal Dependencies Workshop 2021*, organizzato nell'ambito del Syntaxfest 2021.[27]

Per quanto concerne il latino, le cinque *treebank* (includendo *UDante*) attuali assommano circa 980 000 singole parole (o '*token*'),[28] facendone una delle lingue maggiormente rappresentate in UD. Tuttavia, è interessante notare come nessuna di tali *treebank*, tranne UDante, sia stata annotata direttamente seguendo il formalismo di UD, ma ciascuna provenga dalla conversione (semi)automatica di una precedente annotazione: le *treebank Perseus*, IT-TB e LLCT originali (cfr. §1), in particolare, si basano sul livello analitico della PDT, la *Prague Dependency Treebank* della lingua ceca,[29] mentre PROIEL segue un proprio standard.[30] Ciò significa che la *treebank* di *UDante*, pur derivando parte della propria annotazione da un precedente lavoro (*DanteSearch*), è dal punto di vista sintattico la prima a essere creata *ab origine* secondo le linee guida di UD.[31]

(Marrakech, 28-30 maggio 2008), a c. di N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, D. Tapias, Marrakech, European Language Resources Association (ELRA), 2008, pp. 213-218.

[26] https://github.com/UniversalDependencies.

[27] https://syntaxfest.github.io/syntaxfest21/.

[28] Con '*token*' si indica la generica unità di segmentazione di un testo, indipendentemente dai criteri ortografici, filologici e/o linguistici in base a cui tale segmentazione è stata ottenuta.

[29] https://ufal.mff.cuni.cz/pdt2.0/doc/manuals/en/a-layer/html/.

[30] http://dev.syntacticus.org/proiel-annotators-handbook-v1.pdf.

[31] https://universaldependencies.org/guidelines.html.

Universal Dependencies si fonda sul concetto di dipendenza: ogni frase è analizzata come un albero sintattico formato da nodi che corrispondono ciascuno a una 'parola sintattica' (oltre che a simboli, o segni d'interpunzione) e sono in relazione diretta fra loro, senza l'intermediazione di nodi indicanti sintagmi (cfr. §2.1).

A questo aspetto fondante si accompagna in modo inestricabile la priorità data alle parole lessicali, dette anche 'piene' o categorematiche. Ciò significa, equivalentemente, che le parole funzionali ('vuote' o sincategorematiche), cioè quelle che fungono prevalentemente da connettori o operatori grammaticali e non esprimono un significato di per sé, vengono subordinate alle parole piene: in termini sintattici, di norma una parola funzionale dipenderà sempre da una parola piena e potrà comparire nell'albero solo come foglia (cioè, senza figli). A propria volta, gli elementi modificatori dipendono da ciò che modificano. In questo modo, la struttura sintattica viene costruita primariamente attorno agli elementi che svolgono il ruolo di predicati, argomenti e modificatori, e completata secondariamente dai componenti più marcatamente grammaticali. A ciò si accompagna la regola di limitarsi alle parole effettivamente presenti nella frase, senza creare nodi fittizi per gestire fenomeni come l'ellissi.[32]

In una frase come *ad te litteras misi*,[33] il pronome *te* dipende direttamente dal predicato *misi*, alla pari dell'oggetto diretto *litteras* [FIG. 1]. In questo stesso esempio notiamo invece che *ad* dipende da *te* in quanto elemento grammaticale che si limita a veicolare la relazione (semantica) fra le parole rappresentate rispettivamente dai lemmi *tu* e *mitto*. Osserviamo qui che la scelta di non porre la preposizione alla testa del suo sintagma di appartenenza permette un parallelismo sia fra diverse strutture analoghe all'interno della stessa lingua, es. *tibi misi litteras* [FIG. 2], che fra lingue diverse, es. l'equivalente tedesco *ich schickte dir einen Brief* [FIG. 3].

[32] È previsto un secondo livello di 'dipendenze estese' (ing. *enhanced dependencies*) dove fare uso di tali strumenti, che permettono di esplicitare relazioni più dettagliate ma non garantiscono più di aver a che fare con un grafo aciclico (cfr. §2.1).

[33] Le frasi in latino e tedesco di questo paragrafo non derivano da *corpora*, ma sono state create e annotate dagli autori a mero scopo esemplificativo; la frase in tahitiano è reperita da L. PELTZER - V.S. TUHEIAVA-RICHAUD, *Le Tahitien de poche*, Chennevières sur Marne, Assimil, 2011, p. 21, e annotata dagli autori.

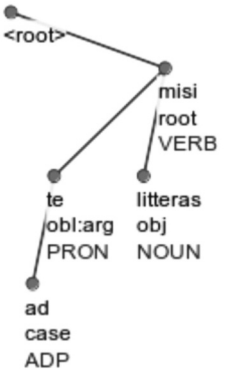

FIG. 1: *ad te litteras misi*

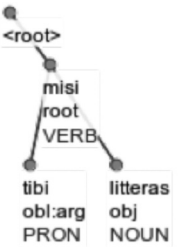

FIG. 2: *tibi misi litteras*

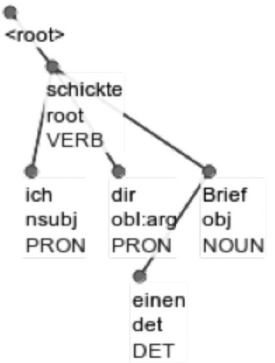

FIG. 3: *ich schickte dir einen Brief*

Nel secondo e terzo albero, una costruzione sintetica corrisponde a una più analitica nel primo albero, e la differenza è prettamente formale. Da un punto di vista universale, mentre la presenza di parole lessicali in qualsiasi lingua naturale è fuori discussione, gli elementi funzionali individuati come parole sintattiche autonome presentano un estremo grado di variabilità fra le varie lingue e sono in concorrenza con strategie più puramente morfologiche o sintattiche: così, mentre una lingua come il latino possiede paradigmi flessionali molto sviluppati e con vari gradi di fusività, in una lingua come il tahitiano, del tutto isolante, ogni parola è morfologicamente immutabile, le categorie e relazioni grammaticali sono espresse tramite particelle, e l'ordine dei costituenti all'interno di una frase è fisso. Si confronti, ad esempio, il tah. *ua 'ite 'oa ia mātou*[34] con l'equivalente lat. *nos vidisti*:

[34] Glosse: *ua '(marca perfettiva)'; 'ite* 'vedere'; *'oa* 'tu'; *ia* '(marca dell'oggetto)'; *mātou* 'noi (esclusivo)'.

FIG. 4: *nos vidisti*

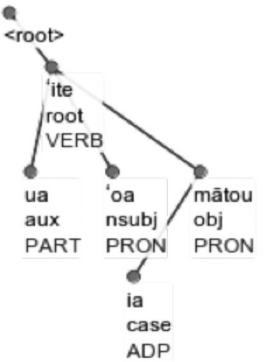

FIG. 5: *ua 'ite 'oa ia mātou*

La struttura portante `root-obj` dei due alberi [FIGG. 3 e 4 e 5] rimane identica, mentre a variare è solo la presenza o assenza di foglie. Notiamo che l'aspetto perfettivo del verbo *video* in latino si esprime tramite coniugazione e in modi diversi a seconda della classe verbale (es. *amo* darebbe qui *amavisti*), mentre in tahitiano ciò è rappresentato dalla particella preposta *ua*, valida per ogni radice verbale. Tuttavia, l'analisi delle due proposizioni procede parallelamente.

La separazione fra il livello semantico e la sua realizzazione sintattica è un altro concetto fondante di UD, che punta a rappresentare solo quest'ultimo e a pronunciarsi il meno possibile sul primo. La principale suddivisione posta per gli elementi all'interno di una proposizione si attua fra gli argomenti appartenenti al 'nucleo' (ing. *core*) e tutti gli altri, detti 'obliqui' (ing. *oblique*), in contrapposizione a una distinzione molto diffusa, ma di impronta semantica, fra 'complementi' necessari e 'aggiunti' facoltativi. Notiamo che, mentre un argomento nel nucleo è sempre un complemento, non vale l'implicazione inversa. Un argomento è individuato come nucleare se il suo comportamento sintattico si orienta verso quelli dei cosiddetti 'protoagente' o 'protopaziente' sintattici, i cui criteri di individuazione possono cambiare da lingua a lingua.[35] Per

[35] A.D. ANDREWS, *The Major Functions of the Noun Phrase*, in *Language Typology and Syntactic Description: Clause Structure*, a c. di T. SHOPEN, Cambridge, Cambridge University Press, 2007², pp. 132-223.

quanto riguarda il latino, li riconosciamo rispettivamente nei concetti tradizionali di soggetto espresso al nominativo e oggetto diretto espresso all'accusativo in una proposizione semplice non marcata; altri argomenti sono obliqui. Osserviamo che in certi contesti alcuni aspetti formali, come il caso, potrebbero subire mutazioni: in una costruzione 'assoluta', per esempio, sia il predicato in forma nominale che il soggetto assumono uno stesso caso strutturale, in latino di norma l'ablativo.[36] Le relazioni nucleari all'interno di una proposizione rimangono in ogni caso le stesse [FIG. 6].

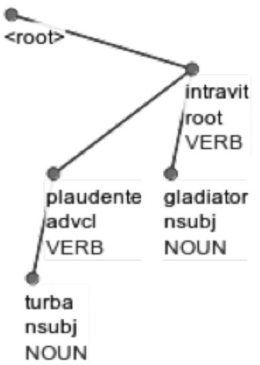

FIG. 6: *turba plaudente gladiator intravit*[37]

Dalle basi appena esposte si diparte tutto il formalismo di UD: un sistema ben definito e intuitivo ma potente al tempo stesso, capace di inglobare più livelli di rappresentazioni sintattiche, senza trascurare del tutto quelle semantiche. Applicarlo sistematicamente al latino si è rivelata una sfida stimolante che è ancora in corso e ha permesso e permetterà una sempre più approfondita comprensione dei suoi intimi meccanismi linguistici.

---

[36] T. NIKITINA - D.T.T. HAUG, *Syntactic Nominalization in Latin: A case of Non-Canonical Subject Agreement*, in «Transactions of the Philological Society», 114 (2016), 1, pp. 25-50.

[37] La relazione sintattica `advcl` indica una subordinata avverbiale, cioè l'equivalente proposizionale di un argomento obliquo nominale.

3. *"UDante": la* treebank *dei testi latini di Dante*

Questo capitolo presenta i dettagli relativi alla creazione della *treebank* dei testi latini di Dante. Più precisamente, viene spiegata la conversione dal formato TEI-XML di *DanteSearch*[38] nel formato CoNLL-U (vedi oltre) e descritta l'annotazione manuale. Particolare attenzione è rivolta alla descrizione di come le annotatrici siano state formate e affiancate nel processo di apprendimento di UD.

3.1. *Dati*

I testi latini di Dante (*Monarchia*, *De vulgari eloquentia*, *Egloghe*, *Epistole* e *Questio de aqua et terra*)[39] sono disponibili in formato elettronico all'interno del *corpus DanteSearch*. Tutti i testi sono stati già lemmatizzati e annotati morfologicamente da un gruppo di studenti e giovani ricercatori dell'Università di Pisa e sono codificati tramite marcatura TEI-XML. I file originali contenenti i testi annotati sono convertiti nel formato CoNLL-U,[40] lo standard usato da UD, manualmente controllati ed emendati nonché annotati sintatticamente usando una specifica interfaccia web.[41]

La conversione da TEI-XML a CoNLL-U è stata effettuata in maniera automatica con un programma creato appositamente per questo

---

[38] Le edizioni di riferimento sono le stesse di *DanteSearch*. Per maggiori informazioni si rimanda alla pagina web contenente i riferimenti bibliografici: https://dantenetwork.it

[39] Nel proseguo dell'articolo i titoli delle opere verranno abbreviate come segue: *Monarchia* = *Mon.*, *De vulgari eloquentia* = *DVE*, *Egloghe* = *Egl.*, *Epistole* = *Ep.*, *Questio de aqua et terra* = *Questio*.

[40] Si tratta di una rielaborazione del formato CoNLL-X, descritto in S. BUCHHOLZ - E. 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 c. di L. MÁRQUEZ, D. KLEIN, New York, Association for Computational Linguistics (ACL), 2006, pp. 149-164. Prende il nome dalla *Conference on Natural Language Learning* durante la decima (= X) edizione della quale è stato presentato. Nel contesto di UD, la U di *universal* sostituisce la X.

[41] J. HEINECKE, *ConlluEditor: a fully graphical editor for Universal Dependencies treebank files*, in Proceedings of the Third Workshop on Universal Dependencies UDW 2019 (Paris, 29-30 agosto 2019), a c. di A. RADEMAKER, F. TYERS, Paris, Association for Computational Linguistics (ACL), 2019, pp. 87-93.

scopo. Innanzitutto, il programma scansiona la struttura del file XML per identificare l'organizzazione interna del testo (ad esempio, la divisione delle opere in libri, capitoli etc.): questa informazione è riportata nel file CoNLL-U, in modo da poter recuperare facilmente la struttura originale del testo partendo dal quel formato. Quindi, la suddivisione in frasi viene prodotta e, per ogni *token*, i valori relativi al proprio *tag* <LM> nel file XML, cioè dove è contenuta l'informazione grammaticale, sono analizzati per estrarre il lemma, la parte del discorso e i tratti morfologici, nonché convertire i codici usati in *DanteSearch* in quelli adottati in UD. Un esempio di *tag* <LM> è il seguente: «<LM lemma="resono" catg="va1cis3">resonaret</LM>» (*DVE* I II 7).

Nello specifico, la parte del discorso e i tratti morfologici sono derivati dal valore dell'attributo `catg`, mentre i campi del formato CoNLL-U riservati all'informazione sintattica sono riempiti con un trattino basso (ovvero, _) e lasciati all'annotazione manuale.

La conversione dell'attributo `catg` è un processo non banale, perché i valori in esso contenuti sono stringhe di lettere e numeri senza una posizione fissa; ciò significa che ogni valore può avere una lunghezza diversa e lo stesso tratto morfologico può occupare posizioni diverse a seconda della parte del discorso.

In generale, UD richiede un'annotazione più granulare dei tratti morfologici rispetto a quella di *DanteSearch*. Riprendendo l'esempio di *resonaret*, il suo valore per `catg` è `va1cis3` ed è convertito come segue:

```
v  >  VERB
a  >  Voice=Act
1  >  InflClass=LatA
ci >  Aspect=Imp|Mood=Sub|Tense=Past|Verb-
      Form=Fin
s  >  Number=Sing
3  >  Person=3
```

Solo pochi (e rari) tratti annotati in *DanteSearch* non trovano una diretta corrispondenza in UD: tra questi segnaliamo le forme medievali dei verbi e i metaplasmi sia verbali che nominali.

Le annotatrici (cfr. §3.2), oltre ad annotare la sintassi, hanno il compito di verificare anche la correttezza della conversione automatica e modificare o aggiungere manualmente elementi non gestiti da essa. Ad esempio, un compito delle annotatrici consiste nel modificare la parte

del discorso dei nomi di popoli (es. «Veronenses», *DVE* I IX 4) che in *DanteSearch* sono marcati come nomi propri mentre, secondo UD, devono essere considerati aggettivi denominali il cui nome di base è un nome proprio.[42]

In [FIG. 7] è mostrato un estratto di file in formato CoNLL-U che riporta una frase del *De vulgari eloquentia*. Dopo tre righe di commento contrassegnate dal simbolo iniziale di cancelletto (#) e indicanti, rispettivamente, il numero progressivo della frase nell'opera (104), il suo testo originale e il suo riferimento nell'opera, parole e segni di interpunzione sono posti uno per riga. Ad esempio, nella prima riga dopo i commenti abbiamo l'indice della prima parola, la parola stessa, il lemma, la parte del discorso, la sequenza originale di codici nell'attributo `catg` di *DanteSearch*, i tratti morfolessicali, l'indice della parola da cui questa dipende nell'albero a dipendenze, il tipo di relazione di dipendenza, eventuali dipendenze estese (qui sempre vuoto), e uno spazio miscellaneo. Mostriamo anche l'albero corrispondente [FIG. 8].

```
# sent_id = DVE-104
# text = Quelibet enim partium largo testimonio se tuetur .
# citationhierarchy = Liber_Primus,x,Paragraphus_2
1   Quelibet    quilibet    DET dinsfn
Case=Nom|Gender=Fem|InflClass=LatPron|Number=Sing|PronType=Ind  7   nsubj   _   _
2   enim    enim    PART    c   _   7   discourse   _   _
3   partium pars    NOUN    sfp3g   Case=Gen|Gender=Fem|InflClass=IndEurI|Number=Plur   1   nmod
    _   _
4   largo   largus  ADJ ans1b   Case=Abl|Degree=Pos|Gender=Neut|InflClass=IndEur0|Number=Sing   5
   amod    _   _
5   testimonio  testimonium NOUN    sns2b   Case=Abl|Gender=Neut|InflClass=IndEur0|Number=Sing  7
   obl _   _
6   se  sui PRON    ppp3sfb Case=Acc|InflClass=LatAnom|Person=3|PronType=Prs|Reflex=Yes 7   obj _
    _
7   tuetur  tueor   VERB    vd2ips3
Aspect=Imp|InflClass=LatE|Mood=Ind|Number=Sing|Person=3|Tense=Pres|VerbForm=Fin|Voice=Pass  0
root    _
8   .   .   PUNCT   _   _   7   punct   _ _
```

FIG. 7: CoNLL-U della frase «Quelibet enim partium largo testimonio se tuetur.» (*DVE* I X 2)

[42] Si veda la definizione di aggettivo (ADJ) nelle linee guida di UD: https://universaldependencies.org/u/pos/ADJ.html

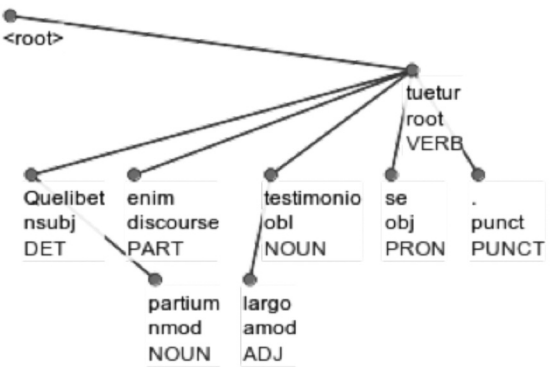

FIG. 8: Albero della frase «Quelibet enim partium largo testimonio se tuetur.» (*DVE* I X 2)

3.2. *Processo di annotazione*

Una parte importante del progetto *UDante* ha riguardato l'addestramento di un gruppo di quattro annotatrici (dottorande o addottorate all'Università di Pisa, con una solida conoscenza delle opere in questione e, in generale, della filologia latina e italiana) sul formalismo di UD, in modo da fornire loro le competenze per svolgere l'annotazione sintattica delle opere di Dante.

L'addestramento è stato organizzato in quattro fasi, intervallate da incontri periodici di discussione, durante i quali le annotatrici hanno annotato frasi di complessità sintattica crescente. In particolare, nelle prime tre fasi sono state selezionate frasi non necessariamente consecutive ma che presentassero vari tipi di strutture sintattiche, diverse lunghezze e differenti profondità dell'albero sintattico.

Lo scopo della prima fase è stato quello di fornire le basi di UD richiedendo l'annotazione di 15 frasi, per lo più tratte dal *DVE*. La seconda e la terza fase hanno invece riguardato 5 e 10 frasi rispettivamente (tratte sia dal *DVE* che dalle *Egl.)*, caratterizzate da strutture sintattiche più complesse rispetto a quelle della prima fase, come ad esempio predicazioni secondarie e proposizioni comparative.

Queste prime tre fasi sono state svolte dalle annotatrici lavorando in parallelo in modo da poter valutare il loro lavoro non solo dal punto di vista qualitativo ma anche da quello quantitativo, calcolando l'ac-

cordo tra di esse[43] (IAA, ing. *Inter-Annotator Agreement*). Dato che l'annotazione di dati linguistici comporta la formulazione di giudizi soggettivi, è fondamentale stabilire fino a che punto tali giudizi siano riproducibili e attendibili. Calcolare l'IAA significa, quindi, valutare il grado di coerenza tra gli annotatori nel trattare i medesimi elementi linguistici: un ampio accordo è considerato garanzia della validità dei dati annotati. L'IAA per ciascuna delle prime tre fasi è stato prodotto usando una misura statistica standard[44] che calcola il grado di accordo nell'annotazione rispetto a quello che ci si aspetterebbe se tale accordo fosse avvenuto per caso, ed è riportata in [TAV. 1]. I valori ottenuti, che possono variare tra 0 (corrispondente ad accordo nullo) e 1 (corrispondente ad accordo perfetto), sono sempre stati maggiori o uguali a 0,79, un valore buono, sia per quanto riguarda la struttura dell'albero sintattico, cioè la creazione delle relazioni sintattiche, che per la scelta delle etichette delle relazioni di dipendenza. Notiamo che la tendenza della *kappa* è a crescere, rappresentando un costante miglioramento; la flessione nella terza fase per quanto riguarda le relazioni sintattiche è fisiologica e dovuta al deciso salto di complessità, ed è stata superata già nella fase successiva.

	Fase 1	Fase 2	Fase 3
Relazione sintattica	0,80	0,83	0,79
Etichetta relazione	0,84	0,92	0,91

TAV. 1: Accordo tra annotatrici.

Nella quarta fase, a differenza di ciò che è avvenuto nelle precedenti, le annotatrici hanno lavorato su gruppi di frasi diverse: ognuna si è infatti concentrata su 10 frasi consecutive di una singola opera, la stessa di cui avrebbe preso in carico l'annotazione completa. Nello specifico, alle quattro annotatrici sono state affidate le seguenti opere: *Mon.*, *DVE*, i due componimenti bucolici di mano dantesca delle *Egl.* e le *Ep.* dalla I alla XII. I testi rimanenti di *DanteSearch*, ovvero l'epistola

[43] R. ARTSTEIN - M. POESIO, *Inter-coder agreement for computational linguistics*, in «Computational Linguistics», 34, (2008), 4, pp. 555-596.

[44] La metrica usata si chiama *kappa* di *Fleiss*.

XIII, i componimenti delle *Egloghe* di Giovanni del Virgilio e la *Questio*, sono invece stati annotati da tre annotatori aggiuntivi già esperti del formalismo di UD.

4. *"UDante" e "LiLa"*

Oltre alla pubblicazione della *treebank* dei testi latini di Dante nella distribuzione di UD, avvenuta nel maggio del 2021, *UDante* è stata allacciata alla *knowledge base* di *LiLa*, che connette tra loro risorse linguistiche per il latino distribuite sul web, al fine di consentirne l'interoperabilità.

L'inserimento di *UDante* in *LiLa* rappresenta un passo decisivo a supporto sia della distribuzione dei (meta)dati della risorsa, sia della valorizzazione del suo utilizzo, in quanto i (meta)dati di *UDante* possono così essere collegati a quelli delle altre risorse per il latino incluse in *LiLa*, siano esse lessicali o testuali, sfruttandoli in modalità incrociata a livello di loro interrogazione e svincolando così *UDante* dall'isolamento che subirebbe qualora essa fosse distribuita come una risorsa indipendente da tutte le altre.

Allargando lo sguardo al di là delle risorse per la lingua latina, l'allacciamento di *UDante* a *LiLa* pone il *corpus* nel vitale e crescente ambito dei cosiddetti *linguistic linked open data* (LLOD), che consistono in risorse linguistiche distribuite e rese interoperabili attraverso l'adozione di un paradigma sviluppato per i fini del *semantic web* e noto come *linked data*, in base al quale i (meta)dati delle risorse sono tra loro messi in relazione attraverso connessioni (*link*) semanticamente definite.[45]

[45] C. CHIARCOS - P. CIMIANO - T. DECLERCK - J.P. MCCRAE, *Linguistic linked open data (LLOD). Introduction and overview*, in *Representing and Linking Lexicons, Terminologies and Other Language Data*. Proceedings of the 2nd Workshop on Linked Data in Linguistics (LDL-2013): (Pisa, 23 settembre 2013), a c. di C. CHIARCOS, P. CIMIANO, T. DECLERCK, J.P. MCCRAE, Pisa, Association for Computational Linguistics (ACL), 2013, pp. 1-9.

4.1. *Cosa è "LiLa"*

Attualmente in via di sviluppo nell'ambito del progetto *LiLa: Linking Latin* (2018-2023), finanziato da una borsa *Consolidator* del Consiglio europeo della ricerca (ERC), la *knowledge base* (KB) di *LiLa* è un connettore di risorse linguistiche distribuite per il latino (ovvero, non necessariamente raccolte in un comune *repository* sul web) fondato sul paradigma dei *linked data*.

LiLa è una KB altamente lessicalizzata, quale conseguenza dell'assunto per cui tutte le componenti che forniscono o producono (meta)dati in essa hanno a che fare con le parole: le risorse testuali sono, infatti, composte da occorrenze di parole, le risorse lessicali forniscono (descrizioni di) proprietà di parole e gli strumenti di trattamento automatico del linguaggio operano su parole.

La connessione interoperativa tra risorse linguistiche avviene tramite la lemmatizzazione, un livello di annotazione ormai sufficientemente diffuso nelle risorse latine e la cui automazione è supportata da strumenti che attualmente garantiscono buoni livelli di accuratezza.[46] Le occorrenze delle singole parole (*token*) delle risorse testuali e le voci delle risorse lessicali vengono allacciate alla base lessicale di *LiLa*, che rappresenta il cuore della KB e consiste in un'ampia raccolta di circa 200 000 forme di citazione (lemmi) di parole latine, con relative varianti grafiche.[47]

LiLa è una KB *open-ended*, ovvero liberamente estendibile attraverso ulteriori connessioni a nuove risorse. Al momento, le risorse connesse in *LiLa* sono le seguenti:

– risorse testuali: *Index Thomisticus Treebank*, *Querolus sive Aulularia* (una palliata di fine IV secolo), *UDante*;

---

[46] R. SPRUGNOLI - M. PASSAROTTI - F.M. CECCHINI - M. PELLEGRINI, *Overview of the EvaLatin 2020 Evaluation Campaign*, in Proceedings of the 1st Workshop on Language Technologies for Historical and Ancient Languages LT4HALA 2020 (Marseille, 12 maggio 2020), a c. di R. SPRUGNOLI, M. PASSAROTTI, Marseille, European Language Resources Association (ELRA), 2020, pp. 105-110.

[47] M. PASSAROTTI - F. MAMBRINI - G. FRANZINI - F.M. CECCHINI - E. LITTA - G. MORETTI - P. RUFFOLO - R. SPRUGNOLI, *Interlinking through Lemmas. The Lexical Collection of the LiLa Knowledge Base of Linguistic Resources for Latin*, in *Current Approaches in Latin Lemmatization*, a c. di M. PASSAROTTI, in «Studi e Saggi Linguistici», 57 (2020), 1, pp. 177-212.

– risorse lessicali: *Latin WordNet*,[48] *LatinAffectus*,[49] *Word Formation Latin*,[50] *Etymological dictionary of Latin and the other Italic Languages*,[51] *Index Graecorum Vocabulorum in Linguam Latinam Translatorum*.[52]

4.2. *Allacciamento di "UDante" a "LiLa"*

Al fine di connettere i lemmi di *UDante* alla KB di *LiLa* è stata per prima cosa eseguita una corrispondenza di stringhe tra i lemmi dei testi e quelli nella KB tenendo anche conto della parte del discorso.

---

[48] G. Franzini - A. Peverelli - P. Ruffolo - M. Passarotti - H. Sanna - E. Signoroni - V. Ventura - F. Zampedri, *Nunc Est Aestimandum: Towards An Evaluation Of The Latin Wordnet*, in Proceedings Of The Sixth Italian Conference On Computational Linguistics CLiC-it 2019 (Bari, 13-15 novembre 2019), a c. di R. Bernardi, R. Navigli, G. Semeraro, in «Ceur Workshop Proceedings», s. AI*IA series, 2481 (2019), pp. 1-8.

[49] R. Sprugnoli - M. Passarotti - D. Corbetta - A. Peverelli, *Odi Et Amo. Creating, Evaluating And Extending Sentiment Lexicons For Latin*, in Proceedings of the Twelfth International Conference on Language Resources and Evaluation LREC'20 (Marseille, 13-15 maggio 2020), a c. di N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis, Marseille, European Language Resources Association (ELRA), 2020, pp. 3078-3086.

[50] E. Litta - M. Passarotti - F. Mambrini, *Derivations and Connections: Word Formation in the LiLa Knowledge Base of Linguistic Resources for Latin*, in «The Prague Bulletin of Mathematical Linguistics», 115 (2020), pp. 163-186.

[51] M. De Vaan, *Etymological Dictionary of Latin - and the other Italic Languages*, s. *Leiden Indo-European Etymological Dictionary Series*, vol. 7, Amsterdan, Brill, 2008; F. Mambrini - M. Passarotti, *Representing Etymology In The Lila Knowledge Base Of Linguistic Resources For Latin*, in Proceedings Of The 2020 Globalex Workshop On Linked Lexicography (Marseille, 12 maggio 2020), a c. di I. Kernerman, S. Krek, J. McCrae, J. Gracia, S. Ahmadi, B. Kabashi, Marseille, France, European Language Resources Association (ELRA), 2020, pp. 20-28.

[52] G.A.E.A. Saalfeld, *Tensaurus Italograecus: Ausführliches historisch-kritisches Wörterbuch der griechischen Lehn- und Fremdwörter im Lateinischen*, Wien, Carl Gerold's Sohn, 1884; G. Franzini - M. Passarotti - F. Mambrini - G. Moretti - F. Zampedri, *Græcissâre: Ancient Greek Loanwords in the LiLa Knowledge Base of Linguistic Resources for Latin*, in Proceedings of the Seventh Italian Conference on Computational Linguistics CLiC-it 2020 (Bologna, 1-3 marzo 2021), a c. di J. Monti, F. Dell'Orletta, F. Tamburini, Bologna, «Ceur Workshop Proceedings», s. AI*IA series, 2769 (2020), pp. 1-6.

Usando questa strategia, la percentuale di lemmi connessa in modo diretto ad un solo lemma nella KB varia a seconda dell'opera in esame: dal 68% del *DVE* all'84% delle *Ep.*. Il resto dei lemmi non connessi ricade in due possibili categorie: lemmi ambigui, cioè con possibile connessione a più di un lemma, e lemmi non presenti nella KB.

I lemmi ambigui sono in media l'8% per testo. Alcuni di questi sono stati disambiguati analizzando le caratteristiche morfologiche. Ad esempio, il lemma verbale *volo* può essere connesso a due rappresentazioni grafiche in *LiLa* ma viene disambiguato guardando alla coniugazione: in un caso si tratta di un verbo di coniugazione irregolare e nell'altro di un verbo di prima coniugazione. In alcuni contesti, invece, la disambiguazione può avvenire solo manualmente controllando ogni occorrenza: ad esempio, il lemma *campus* può essere connesso a due lemmi, con diverso significato (rispettivamente, 'campo' e 'ippocampo'), ma entrambi sostantivi maschili della seconda declinazione.

La restante parte dei lemmi non direttamente connessi a *LiLa* riguarda lemmi non presenti nella KB. La percentuale di lemmi di questo tipo varia a seconda del testo ma è alta (25%) nel *DVE*: di questi lemmi, il 10% corrisponde a parole non latine che non devono essere connesse alla KB. Un esempio è dato da parole in volgare italiano come «inanimatissimamente» (*DVE* II VII 6) o in provenzale come «bonté» (*DVE* II V 4).

I lemmi che, invece, sono stati aggiunti alla KB di *LiLa* possono essere classificati in quattro tipologie principali.

La prima tipologia è quella che raggruppa rappresentazioni grafiche diverse di lemmi già nella KB. Il caso più comune è quello di lemmi che presentano una monottongazione, ovvero la semplificazione dei dittonghi *ae* e *oe* in *e*. Tale riduzione dei dittonghi si registra in tutte le opere dantesche e per varie parti del discorso: ad esempio, *aequivoce* > «equivoce» (*Questio* 22), *praehonoratus* > «prehonoratus» (*DVE* I XIII 4). Altre variazioni grafiche riscontrate sono, tra le altre, l'alternanza tra *i* e *y* (es. *idealiter* > «ydealiter», *Questio* 46) e tra *s* e *z* (es. *introniso* > «intronizo», *Mon.* III IV 1), oltre che riduzioni consonantiche (es. *contemptrix* > «contemtrix», *Ep.* III 7).

La seconda tipologia di lemmi aggiunti alla KB è quella degli antroponimi. Sono numerosi soprattutto i nomi di persona tratti dal *DVE* e dalla *Mon.*. In particolare, nel primo caso, si è trattato per lo più di nomi di poeti come trovatori (es. *bertramus*, *namericus*) o autori toscani (es. *cavalcantis*, *guinizelli*, *mocatus*). Nel secondo, invece, i nomi sono quelli di personaggi dell'antichità greca, romana o etrusca: es. *eurialus*, *mutius*, *nicomacus*, *porsenna*.

Aggettivi e nomi etnici costituiscono la terza tipologia: si tratta, per la maggior parte, di lemmi del *DVE*. In questa opera, infatti, Dante menziona numerose popolazioni per discutere dei volgari municipali italiani: es. *aquileiensis*, *bononiensis*, *neapoletanus*.

L'ultima tipologia di nuovi lemmi connessi alla KB è quella delle neoformazioni, ovvero di termini che sembrano apparire per la prima volta in Dante perché non registrati in lessici e glossari noti.[53] Tali neologismi sono aggettivali, nominali o verbali formati per combinazione di prefissi o desinenze e spesso hapax: es. *adiuvalis* («adiuvalis», *Ep.* VIII 1), *nequitatrix* («nequitatrix», *DVE* I VII 2), *perpallesco* («perpalluit», *Egl.* II 30).

5. *Interrogazione di "UDante"*

Al fine di mostrare che tipo di supporto *UDante* possa fornire allo studio della sintassi dei testi latini di Dante e, più specificamente, alla compilazione delle entrate lessicali del VDL, questo capitolo riporta una serie di query d'esempio che è possibile lanciare sui dati di *UDante*. Il capitolo è diviso in due parti, rispettivamente dedicate a interrogazioni che possono essere operate su *UDante* e sulle altre *treebank* di UD, e a una ricerca incrociata che valorizza l'interoperabilità tra le risorse linguistiche attualmente allacciate alla KB di *LiLa*.

5.1. *"UDante" e le* treebank *di "Universal Dependencies"*

Le *treebank* di UD possono essere interrogate attraverso un servizio on-line basato sul linguaggio di query PML-TQ (*Prague Markup Language – Tree Query*).[54] Il servizio è accessibile presso l'indirizzo

[53] G. BRUGNOLI, *Il latino di Dante*, in *Dante e Roma.* Atti del Convegno di Roma (Roma, 8-10 aprile 1965), a c. di CASA DI DANTE, Firenze, Le Monnier, 1965, pp. 51-71.

[54] http://ufal.mff.cuni.cz/pmltq. J. ŠTĚPÁNEK - P. PAJAS, *Querying Diverse Treebanks in a Uniform Way*, in Proceedings of the Seventh conference on International Language Resources and Evaluation LREC'10 (Il-Belt Valletta, 17-23 maggio 2010), a c. di N. CALZOLARI, K. CHOUKRI, J. MARIANI, J. ODIJK, S. PIPERIDIS, M. ROSNER, D. TAPIAS, Il-Belt Valletta, European Language Resources Association (ELRA), 2010, pp. 1828-1835.

http://lindat.mff.cuni.cz/services/pmltq/#!/treebanks, ove è possibile selezionare la singola *treebank* su cui operare un'interrogazione.

A modo di esempio, si consideri il supporto che *UDante* può fornire alla compilazione di una voce del VDL come il nome *deus*. Attualmente, attraverso l'interfaccia web del *corpus DanteSearch* è possibile interrogare i dati in modo tale da visualizzare tutte le occorrenze del lemma *deus* nei testi di Dante, che assommano a un totale di 6 in *Purg.*, 3 in *Par.*, 1 in *VN* e 215 nelle opere latine. Rispetto a *DanteSearch*, i metadati sintattici di *UDante* consentono di organizzare le 215 occorrenze di *deus* nei testi latini di Dante secondo il loro comportamento sintagmatico, ad esempio rispondendo a domande come le seguenti.

La prima domanda chiede di quali verbi sia soggetto *deus* (e di quali sia oggetto). In termini di query, questa domanda consiste nel ricercare negli alberi a dipendenze di *UDante* tutte le occorrenze di nodi occupati da una forma del lemma *deus* che dipendono da un verbo (di cui si può specificare la diatesi) attraverso la relazione di dipendenza usata per i soggetti nominali (`nsubj`). Qualora, invece, si volessero isolare le occorrenze di *deus* come oggetto diretto di un verbo, la relazione di dipendenza da utilizzare sarebbe `obj`.

Una seconda domanda interroga il *corpus* chiedendo quali siano i nomi modificati da *deus*. La query che fornisce risposta a questa domanda consiste nel cercare nodi nominali che governano *deus* come loro modificatore nominale (`nmod`).

L'albero riportato in [FIG. 9] presenta 2 occorrenze di *deus*, corrispondenti rispettivamente ad output prodotti dalla prima e dalla seconda query descritte. La prima occorrenza (*Deus…fecit*) è un caso di *deus* come soggetto nominale di un verbo. Nell'albero di [FIG. 9], ciò è rappresentato dalla dipendenza diretta del nodo della forma *Deus* (che ha lemma *deus*) rispetto al nodo della forma *fecit*, attraverso la relazione sintattica `nsubj`, assegnata a *Deus*. La seconda occorrenza del lemma *deus* nella frase in questione mostra, invece, un nome (*vicarius*) modificato da una forma di *deus* (*Dei*), il che è rappresentato nell'albero di [FIG. 9] come una dipendenza diretta del nodo della forma *Dei* rispetto a quello di *vicarius*, attraverso la relazione `nmod` assegnata a *Dei*.

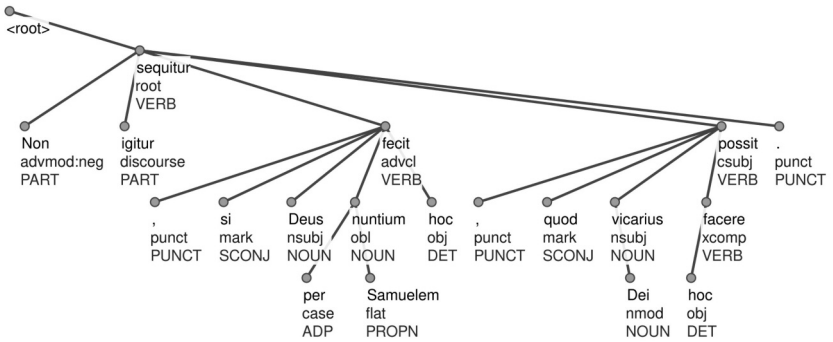

FIG. 9: Albero della frase «Non igitur sequitur, si Deus per nuntium Samuelem fecit hoc, quod vicarius Dei hoc facere possit.» (*Mon.* III VI 6)

Un'altra domanda cui *UDante* può rispondere consiste nel chiedere di quali strutture formate da copula e nome del predicato sia soggetto *deus*. A livello di query, questa domanda richiede di ricercare i nodi di una forma di *deus* che siano soggetti nominali di un nodo che, a propria volta, governi un altro nodo attraverso la relazione `cop` (copula), avendo anche la possibilità di indicare che questo nodo sia occupato da una forma del verbo *sum*.

L'albero in [FIG. 10] mostra uno degli output di questa query. L'occorrenza del lemma *deus* (*Deus*) è il soggetto nominale (`nsubj`) di un nodo (*dictator*) che governa una forma di *sum* (*est*) cui è assegnata la relazione attribuita alle copule (`cop`).

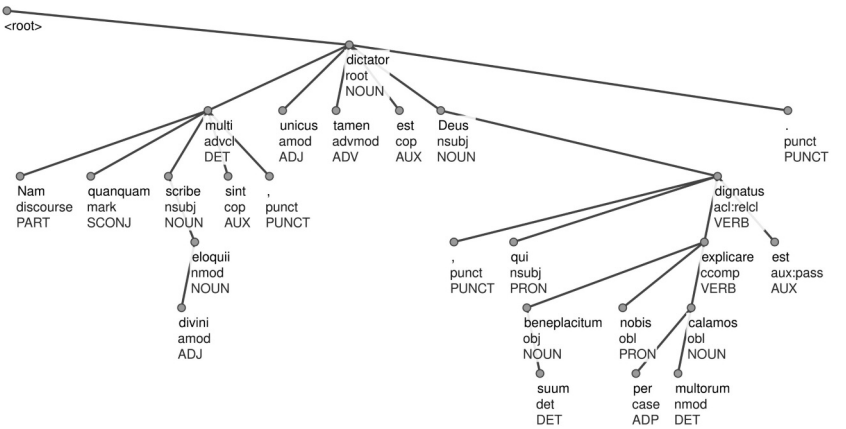

FIG. 10: Albero della frase «Nam quamquam scribe divini eloquii multi sint, unicus tamen dictator est Deus, qui beneplacitum suum nobis per multorum calamos explicare dignatus est.» (*Mon.* III IV 11)

Oltre a fornire in output le singole occorrenze testuali che corrispondono alle condizioni espresse nelle query, il linguaggio PML-TQ consente anche di produrre liste, i contenuti delle cui colonne possono essere definiti dagli utenti. Ad esempio, richiamando le domande sopra riportate, è possibile produrre l'elenco (ordinato per numero decrescente di occorrenze) dei verbi di cui *deus* è soggetto, dei nomi modificati da *deus* e dei lemmi che fanno da predicato nominale in una struttura copulativa di cui un'occorrenza di *deus* funga da soggetto.

Si noti, infine, come l'innesto di *UDante* in UD permetta di fare interagire il *corpus* delle opere latine di Dante con tutte le altre *treebank* lì registrate. Ciò consente, ad esempio, di lanciare una stessa query non solo sui dati di *UDante*, ma su tutte le *treebank* latine di UD, rendendo dunque possibile un'analisi comparativa tra i comportamenti sintattici di una specifica parola in *corpora* che raccolgono testi latini di epoche, autori e generi diversi.

5.2. *"UDante" e le risorse linguistiche di "LiLa"*

Al fine di mostrare il valore aggiunto dell'inclusione di *UDante* nella KB di *LiLa*, questo breve paragrafo descrive informalmente una ricerca d'esempio che può essere lanciata su *LiLa* presso https://lila-erc.eu/sparql/. La ricerca sfrutta l'interoperabilità tra le risorse linguistiche che *LiLa* consente, particolarmente concentrandosi su aspetti di tipo lessicale, al fine di evidenziare il supporto che l'allacciamento di *UDante* a *LiLa* può fornire alla compilazione delle entrate del VDL.

Sfruttando (meta)dati linguistici tratti da risorse sia di tipo testuale che lessicale rese interoperabili attraverso *LiLa*, è possibile ricercare nelle opere di *UDante* e/o in tutte quelle degli altri *corpora* allacciati a *LiLa* le occorrenze testuali delle parole che condividono l'appartenenza a una specifica famiglia morfologico-derivazionale. Una query che miri a produrre un siffatto risultato fa uso dei (meta)dati forniti da *UDante* e dagli altri *corpora* di *LiLa* (risorse testuali), oltre che dell'informazione morfologico-derivazionale registrata nella raccolta di forme di citazione di *LiLa* (tratta dalla risorsa lessicale *Word Formation Latin*), dove ciascun lemma del latino classico è connesso a una base lessicale, o a più di una base nel caso di parole composte. Tutti i lemmi connessi a una medesima base sono, quindi, membri di una medesima famiglia lessicale morfologico-derivazionale.

La query in questione può concentrarsi, ad esempio, sulle occorrenze testuali dei lemmi connessi alla base 231 di *LiLa*, che riunisce i membri della famiglia morfologico-derivazionale del verbo *loquor*. In *LiLa*, a questa base sono associati 135 lemmi, tra cui, ad esempio, *collocutor*, *stultiloquentia* e *suaviloquus*. Nello specifico del *DVE*, la query restituisce che 7 dei lemmi connessi alla base 231 presentano almeno un'occorrenza nel testo: *loquor* (106), *locutio* (60), *loquela* (28), *eloquentia* (10), *alloquor* (2), *turpiloquium* (2). Inoltre, la query fornisce informazione in merito alla presenza dei membri della famiglia di *loquor* nei testi degli altri *corpora* allacciati a *LiLa*, ad esempio permettendo di consultare le 15 occorrenze di *locutio* nella *Summa contra Gentiles* di Tommaso d'Aquino, tratta dalla *Index Thomisticus Treebank*. Infine, sfruttando i metadati sintattici di *UDante* e delle *treebank* presenti in *LiLa*, è possibile concentrarsi solo sulle occorrenze che manifestano una certa funzione sintattica, come ad esempio sugli usi di *locutio* come soggetto nominale (relazione di dipendenza `nsubj`).

6. *Conclusioni*

Attualmente, le risorse linguistiche, e in particolare i *corpora* annotati a livello sintattico, non sono (più) raccolte di (meta)dati linguistici utili ai soli ricercatori che lavorano nel settore del trattamento automatico del linguaggio e, più generalmente, nella linguistica computazionale. Unitamente agli strumenti per produrle, distribuirle e interrogarle, oltre che alle piattaforme e ai metodi per farle interagire tra loro, le *treebank* forniscono un accesso, fino a pochi anni fa indisponibile, a sostanziali quantità di evidenza empirica a supporto di ricerche anche negli ambiti umanistici più tradizionali, come sono gli studi sui testi scritti in lingue antiche e/o di autori letterari.

La collaborazione tra il VDL e *LiLa*, mirata alla realizzazione di *UDante*, è stata avviata proprio per innestare la raccolta dei testi latini di Dante nello stato dell'arte del settore delle risorse linguistiche, arricchendoli con annotazione che faccia ricorso a formati, modelli di rappresentazione, etichette e criteri di loro applicazione che sono ormai standard *de facto*, in virtù della loro applicazione su numerosi *corpora* in molte lingue. Tale innesto è avvenuto in due fasi. Innanzitutto, adottando per *UDante* lo stile di annotazione di UD, che ha consentito di armonizzare l'annotazione dei dati a quella delle più di 220 *treebank* oggi ivi disponibili; quindi, allacciando *UDante* alla KB di risorse latine interoperabili di *LiLa*, il che ha posto *UDante* nelle condizioni di poter interagire con le molte risorse per il latino che sono state sviluppate nel corso dei decenni.

Tutto il lavoro è stato guidato dalla coscienza che l'unico modo che oggi abbiamo ancora a disposizione per poter sentire la voce di Dante è leggere i suoi testi e che, dunque, è nostro preciso dovere valorizzarli e studiarli al meglio, utilizzando i metodi più avanzati per gestire ed estrarre l'informazione che essi veicolano, con l'obiettivo di produrre nuova conoscenza. Perseguendo un approccio altamente interdisciplinare, *UDante* è stato creato facendo incontrare i metodi e gli strumenti dei linguisti computazionali con la competenza linguistica specifica degli annotatori sulla lingua latina e sui testi di Dante: un'ibridizzazione che ha arricchito entrambe le parti, portando agli umanisti la dimensione interoperativa delle risorse linguistiche e l'innovazione analitico-gestionale del dato nella sua interezza, e ai computazionali l'acribia filologica e la cura del dettaglio del dato nella sua specificità.

Ringraziamenti

Gli autori ringraziano Daniela Corbetta, Federica Favero, Federica Gamba, Martina de Laurentiis, Giulia Pedonese, 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*.

Marco Passarotti, Flavio Massimiliano Cecchini, Rachele Sprugnoli, Giovanni Moretti (Università Cattolica del Sacro Cuore di Milano): marco.passarotti@unicatt.it flavio.cecchini@unicatt.it rachele.sprugnoli@unicatt.it giovanni.moretti@unicatt.it

FINITO DI STAMPARE
NEL MESE DI SETTEMBRE 2022
PER CONTO DI
EDITORIALE LE LETTERE
DALLA TIPOGRAFIA BANDECCHI & VIVALDI
PONTEDERA – PISA