



UNIVERSITÀ
CATTOLICA
del Sacro Cuore

www.unicatt.it

Centro interdisciplinare di ricerche per la Computerizzazione
dei segni dell'espressione - CIRCSE

LINKING LATIN. BUILDING A KNOWLEDGE BASE OF LINGUISTIC RESOURCES FOR LATIN (LILA)



This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme - Grant Agreement No 769994

Centro interdisciplinare di ricerche per la Computerizzazione
dei segni dell'espressione - CIRCSE
Email: marco.passarotti@unicatt.it
Telefono: +39 02 7234 2380
Sito web: centridiricerca.unicatt.it/circse



UNIVERSITÀ
CATTOLICA
del Sacro Cuore

Nonostante il settore di ricerca dedicato alla creazione, diffusione e utilizzo delle risorse linguistiche e degli strumenti di trattamento automatico per il latino (e, più in generale, per le lingue antiche) abbia riscontrato una forte crescita nel corso dell'ultimo decennio, le risorse linguistiche per il latino sono ancora prive d'interoperabilità. Ciò significa che l'informazione linguistica è sparsa in molti prodotti, che non parlano tra loro.

Tale situazione comporta uno sfruttamento limitato della ricchezza fornita da tutti quegli oggetti digitali che sono stati prodotti per il latino in anni di lavoro. Dal momento che il latino è una lingua morta (che, dunque, manca di parlanti nativi), tutto ciò che possiamo e dobbiamo fare è utilizzare al meglio l'informazione contenuta nei pochi e preziosi testi che sono sopravvissuti dal passato.

Questo significa:

- realizzare la migliore organizzazione possibile delle risorse linguistiche per il latino (accostate a servizi online di trattamento automatico del linguaggio - TAL -) al fine di ottenere un'integrazione fruttifera dell'informazione che esse forniscono, ovvero di estrarre e combinare informazioni provenienti da diverse fonti nel modo più efficiente possibile;
- mettere a disposizione risorse linguistiche di qualità controllata.

L'obiettivo del progetto LiLa consiste nel connettere e, in ultima istanza, sfruttare l'insieme di risorse linguistiche e strumenti di TAL costruiti fino ad oggi, al fine di superare la separazione tra dati linguistici, TAL e descrizioni della conoscenza, consentendo così ai ricercatori di usare al meglio le risorse e gli strumenti attualmente disponibili.

Per far fronte a tale sfida, LiLa mira a incorporare le risorse linguistiche per il latino nel framework Linked Data, rendendo possibile la loro pubblicazione, interconnessione e interazione online. A tal fine, il progetto sta costruendo una knowledge base per il latino fondata sui principi del paradigma Linked Data, per combinare tra loro dati provenienti da diverse risorse linguistiche, fornire servizi online di TAL e includere anche il latino nel multilingue Linguistic Linked Open Data cloud.

Impatto sociale

Il progetto si situa nel settore della Linguistica Computazionale e mira ad avere un impatto notevole sul modo di lavorare quotidiano dei classicisti (e non solo). Il classicista del terzo millennio dovrà essere in grado allo stesso tempo di consultare un volume cartaceo e di usare una base di conoscenza sul web.

Ciò apre anche una sfida formativa: le nostre università si trovano davanti alla necessità di rinnovarsi, offrendo agli studenti delle facoltà umanistiche competenze non solo essenziali per far avanzare la ricerca che verrà, ma anche capaci di aprire loro nuovi e promettenti sbocchi professionali post-laurea.

Esempi d'impatto sociale dei risultati del progetto

Il docente di liceo: potrà accedere a testi latini in formato digitale che sono raccolti in diverse collezioni in giro per il mondo e potrà usare strumenti di trattamento automatico del linguaggio per fare *e-learning* su di essi.

Ad esempio, potrà chiedere ai propri studenti d'identificare tutti i soggetti al nominativo di una porzione di testo e poi consentire loro di verificare nella base di conoscenza che il progetto sta costruendo quanto sia giusta la loro risposta (o se quella della macchina sia sbagliata).

Lo studente universitario: potrà accedere a testi, dizionari, lessici latini connessi tra loro e operare ricerche incrociate con un linguaggio unico.

Potrà lanciare diversi tipi di strumenti di analisi automatica del linguaggio, scaricare i dati e usarli per le proprie ricerche.

Il classicista: potrà sostenere le proprie ricerche con grandi massi di evidenza empirica, che gestirà facilmente grazie a interfacce di accesso di semplice uso e a connessioni semantiche soggiacenti in grado di mettere in relazione le varie informazioni sul medesimo oggetto (ad esempio, una parola).

Potrà rendere le proprie ricerche replicabili, documentando i procedimenti seguiti per arrivare al dato che supporta le conclusioni della ricerca.

Potrà avere sul proprio tavolo di lavoro quotidiano un accesso a una base di conoscenza ricca, dinamica e varia, che lo metterà nelle condizioni non solo di accedere a testi/dizionari in formato digitale (cosa che oggi già accade), ma anche alle connessioni tra essi, che abatteranno i confini tra i siti web che oggi li archiviano.

Nuove figure professionali

Il progetto sviluppa e utilizza metodi e tecnologie il più *language-independent* possibile. Ciò rende i suoi risultati ampiamente portabili su lingue diverse dal latino. I collaboratori del progetto (dagli studenti ai ricercatori) maturano competenze nel settore della linguistica computazionale oggi ampiamente richieste dal mercato del lavoro.

Ogni giorno vengono pubblicate centinaia di migliaia di parole sul web, in lingue diverse, su ogni argomento e nei più svariati registri. Chi è in grado di organizzare questa massa di dati in linguaggio naturale (per sua natura informale e potenzialmente ambiguo) ed estrarre da essa informazione strutturata ha in mano il mondo della comunicazione. Il linguista computazionale è la figura professionale più confacente a questa sfida, in quanto è in grado di comprendere (e realizzare) i processi che una macchina deve compiere per trattare automaticamente dati linguistici.