Il latino sbarca nel Semantic Web. La Knowledge Base LiLa

Marco Passarotti[1]

Abstract

Questo articolo presenta LiLa, una Knowledge Base fondata sui principi del paradigma Linked Data (adattato nel Semantic Web) e dedicata al latino. Dopo una breve introduzione al Semantic Web e alla sua rilevanza per le risorse linguistiche, il lavoro si concentra dedicato alle diverse fase metodologiche attraversate lungo la stesa della cooperazione interdisciplinare, l'attività formale verso una tassonomia della risorse linguistiche disponibili per il latino, la creazione di risorse ad hoc per la lemmatizzazione e la creazione di dati in linguaggio RDF. Infine, viene illustrata l'architettura della Knowledge Base LiLa che si incentra attraverso alcuni esempi specifici di uso.

Abstract

This paper introduces the LiLa Knowledge Base, a collection of both lexical and textual resources for Latin made interoperable on the Web by applying the principles of the Linked Data paradigm (widely adapted in the Semantic Web). After discussing the methodological steps underlying the cooperation between linguists and computer scientists, attention focuses on a formal ontology of Latin linguistic resources for Latin, to the creation of ad hoc resources for lemmatization, and to the creation of data in RDF format. Finally, the architecture of the LiLa Knowledge Base is shown through a number of examples of its use.

Parole-chiave

Linguistica Computazionale, Risorse Linguistiche, Latino, Linguistica Lirbaol Open Data, Ontologie

Introduzione

Quando nel 1949 il gesuita rentrazionista Roberto Busa si sedette nello studio del gesuita capo dell'IBM Thomas Watson Sr. a New York aveva ben chiaro cosa sarebbe voluto realizzare. Si era convinto che l'automazione dei processi di analisi testuale, che sarebbero stati presto avviati, avrebbero portato una svolta metodologica nello studio dei testi: la computazione, transitando dalla quantità alla qualità, avrebbe permesso la decodificazione, intesa come la capacità di raggiungere la cosi fondato sul significato e sulla qualità.

Nel Novecento Roman Jakobson (in additio 1) ha scritto di questo: «una verificazione alla base di una rivoluzione metodologica in scienze». "Una rivoluzione che ancora oggi non è avvenuta, ma certamente a qualcosa all'ora, ma sulla ripetere parte busa.

Da l'inizio di una storia metodologica, che ha dato vita alla prima ma è 1974 è 1990 dell'Italia Thomasiana (Busa, 1974-1980), la Linguistica Computazionale (LC) ha cercato di seguire la strada Tomasiana d'Aquino, in modo di poter dialogare con la filosofia.

Nel frattempo, gli studi computazionali sul linguaggio si sono arricchiti, attraverso le tappe di Busa, di nuovi ambiti di ricerca, di nuove metodologie e nuovi strumenti informatici, mentre nasce la Banca del sapere ALPHC del Dante Alighieri Research Council (a Kerlini, 1996), la Linguistica Computazionale, che in storia vuole formare «un ponte tra le scienze dure e le scienze umane», che «consente l'uso di strumenti di analisi linguistica automatica fondamentale, quali quelle sulla lemmatizzazione (ovvero, la riduzione delle parole di un testo alle loro forme di base) e sull'etichettatura morfologica (ovvero, la attribuzione delle parole di un testo ad classi morfologiche, linguistiche) e dati morfologici (genere, numero, modo, tempo etc.), incontrano, il discorso affrontato comunque dovuto da cui l'utilizzo grammaticale di un testo.

Parallelamente, la maggioranza di strumenti di analisi linguistica automatica favorita lo sviluppo di corpora arricchiti con diffuse informazioni linguistiche, che hanno dato vita alle cosiddette banche dati testuali, che, in un secolo di studi critico, ha prodotto una vasta empirica nel trattamento automatico del linguaggio naturale. Tali risorse, sviluppate in ambiti di ricerca, sono state anche matrice, nonostante l'impegno e spesso le ore di studio, di metodologie, sviluppate e prodotte per un ambito di ricerca e che produce che la Linguistica Computazionale.

A partire dagli anni Novanta del secolo scorso, il mondo della ricerca ha cominciato a dare sempre più importanza alle risorse linguistiche (Cavalli, Fontaine e Mitkov 1997 e Linguistica Computazionale del CNR di Pisa), grazie, collaborazione di Busa ed Vittore Tommaso.

Il termine "linguistiche" nasce nascono dal linguaggio, macroscopicamente categorizzati in due tipi: risorse testuali o corpora, aspetto e metodologie (Lenci Tesoro, e Calzolari). Proprio con gli anni Novanta, la crescita progressiva di disponibilità di risorse testuali ha cominciato a mostrare i limiti di un approccio alla banca di dati, che, in un secolo di studi critico, ha prodotto una vasta empirica nel trattamento automatico del linguaggio naturale. Tali risorse, sviluppate in ambiti di ricerca, sono state anche matrice, nonostante l'impegno e spesso le ore di studio, di metodologie, sviluppate e prodotte per un ambito di ricerca e che produce che la Linguistica Computazionale.

La corsa empirica oggi è scoppiata anche alla cura dei testi che costituiscono il linguaggio. Tali risorse testuali, sviluppate dal lato. Quotidianamente vengono prodotti e «noi e diversi» testo accessibili in Rete «Internet» (Lenci, Lenci e Lenci), con il rischio di ritrovarsi, una «massa contenitore dalla piattaforma di Business Intelligence» (Lenci, Lenci e Lenci). Tali risorse testuali, sviluppate in ambiti di ricerca, sono state anche matrice, nonostante l'impegno e spesso le ore di studio, di metodologie, sviluppate e prodotte per un ambito di ricerca e che produce che la Linguistica Computazionale.

Nel 2000 Tim Berners-Lee, ideatore del World Wide Web, ha scritto: «Il Web ha cambiato il modo in cui le persone interagiscono con le informazioni e con il commercio.

tratta di un esempio, auspicabilmente virtuoso, di interdisciplinarità: un paradigma sviluppato originariamente per mettere in atto gli obiettivi del Semantic Web è applicato a (meta)dati linguistici e, nello specifico, a testi e lessici latini. Il naturale e intimo rapporto che gli studi classici mantengono con l'evidenza testuale è la ragione di fondo che spinge e motiva questa interdisciplinarità: come il latino è stata una delle prime lingue trattate a computer, grazie ai lavori di padre Busa, così oggi il latino è la lingua che può vantare il più granulare livello di interoperabilità tra le sue risorse grazie alla LiLa Knowledge Base, che consente l'interazione a livello della singola occorrenza di parola e della singola entrata lessicale.

L'obiettivo più prossimo è fare vivere davvero LiLa, portando le risorse digitali, gli strumenti computazionali di analisi linguistica e la loro interoperabilità sul tavolo di lavoro quotidiano del classicista: è un'ulteriore sfida, che riguarda non solo la ricerca, ma anche, e forse soprattutto, la formazione alla ricerca. Non sarà banale affrontarla, dato un certo grado di conservatorismo riottoso non infrequente nel mondo degli studi classici: ma è inevitabile.

Ringraziamenti

Il progetto "LiLa: Linking Latin" è finanziato dallo European Research Council (ERC) nell'ambito del programma Horizon 2020 – Grant Agreement N. 769994.

Bibliografia

Bamman, D. & Crane, G. (2007), The Latin Dependency Treebank in a cultural heritage digital library, in Sporleder, C., van den Bosch, A. & Grover, C. (eds.), Proceedings of the Workshop on Language Technology for Cultural Heritage Data (LaTeCH 2007) (33-40), Prague: Association for Computational Linguistics.

Berners-Lee, T. (2006), Linked Data, https://www.w3.org/DesignIssues/LinkedData.html.

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

Busa, R. (1974-1980), Index Thomisticus, Stuttgart-Bad Cannstatt: Frommann-Holzboog.

Camarda, D.V., Mazzini, S. & Antonuccio, A. (2012), LodLive, exploring the web of data, in Sack, H., Pellegrini, T., Presutti, V. & Pinto, H.S. (eds.), Proceedings of the 8th International Conference on Semantic Systems (197-200), New York: Association for Computing Machinery.

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

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

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

Eckhoff, H., Bech, K., Bouma, G., Eide, K., Haug, D., Haugen, O.E. & Jøhndal, M. (2018), The PROIEL treebank family: a standard for early attestations of Indo-European languages, "Language Resources and Evaluation", 52(1), 29-65.

Franzini, G., Peverelli, A., Ruffolo, P., Passarotti, M., Sanna, H., Signoroni, E., Venturi, V. & Zampedri, F. (2019), Nunc Est Aestimandum: Towards an Evaluation of the Latin WordNet, in Bernardi, R., Navigli, R. & Semeraro, G. (eds.), Proceedings of the Sixth Italian Conference on Computational Linguistics (1-8), Bari: CEUR Workshop Proceedings.

Korkiakangas, T. (2021), Late Latin Charter Treebank: contents and annotation, "Corpora", 16(2), 191-203.

Lassila, O & Swick, R.R. (1998), Resource description framework (RDF) model and syntax specification, http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.44.6030.

Litta, E. & Passarotti, M. (2019), (When) inflection needs derivation: a word formation lexicon for Latin, in Holmes, N., Ottink, M., Schrickx, J. & Selig, M. (eds.), Lemmata Linguistica Latina. Volume I: Words and Sounds (224-239), Berlin-Boston: Walter de Gruyter GmbH.

Mambrini, F., Litta, E., Passarotti, M. & Ruffolo, P. (2021), Linking the Lewis & Short Dictionary to the LiLa Knowledge Base of Interoperable Linguistic Resources for Latin, in Fersini, E., Passarotti, M. & Patti, V. (eds.), Proceedings of the Eighth Italian Conference on Computational Linguistics (1-7), Milano: CEUR Workshop Proceedings.

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

Minozzi, S. (2017), Latin WordNet, una rete di conoscenza semantica per il latino e alcune ipotesi di utilizzo nel campo dell'Information Retrieval, in Mastandrea, P. (a cura di), Strumenti digitali e collaborativi per le Scienze dell'Antichità (123-134), Venezia: Edizioni Ca' Foscari.

National Research Council (US). Automatic Language Processing Advisory Committee & ALPAC (1966), Language and Machines: Computers in Translation and Linguistics; a Report (Vol. 1416), National Academies.

Passarotti, M. (2019), The Project of the Index Thomisticus Treebank, in Berti, M. (ed.), Digital Classical Philology. Vol. 10 of Age of Access? Grundfragen der Informationsgesellschaft (299-320), Berlin-Boston: Walter de Gruyter GmbH.

Passarotti, M., Saavedra, B.G. & Onambele, C. (2016), Latin vallex. a treebank-based semantic valency lexicon for Latin, in Calzolari, N., Choukri, K., Declerck, T., Grobelnik, M., Maegaard, B., Mariani, J., Moreno, A., Odijk, J. & Piperidis, S. (eds.), Proceedings of the Tenth International Conference on Language Resources and Evaluation (2599-2606), Paris: European Language Resources Association.

Prud'Hommeaux, E. & Seaborne, A. (2008), SPARQL Query Language for RDF. W3C, https://www.w3.org/TR/rdf-sparql-query/

Schuurman, I., Windhouwer, M., Ohren, O. & Zeman, D. (2016), CLARIN concept registry: the new semantic registry, in De Smedt, K. (ed.), Selected Papers from the CLARIN Annual Conference 2015 (62-70), Linköping: Linköping University Electronic Press.

Sprugnoli, R., Passarotti, M., Corbetta, D. & Peverelli, A. (2020), Odi et Amo. Creating, Evaluating and Extending Sentiment Lexicons for Latin, in Calzolari, N., Béchet, F., Blache, P., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J. & Piperidis, S. (eds.), Proceedings of the Twelfth International Conference on Language Resources and Evaluation (3078-3086), Paris: European Language Resources Association.

Tenney, I., Dipanjan, D. & Pavlick, E. (2019), BERT Rediscovers the Classical NLP Pipeline, in Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (4593-4601), Florence: Association for Computational Linguistics.

* **Marco Passarotti** è Professore Ordinario di Glottologia e Linguistica presso l'Università Cattolica del Sacro Cuore di Milano. I suoi interessi di ricerca si concentrano sulla Linguistica Computazionale e, nello specifico, sullo sviluppo, l'uso e la distribuzione di risorse linguistiche e strumenti di trattamento automatico della lingua latina. Dal 2006 guida il progetto della Index Thomisticus Treebank. Nel 2009 ha fondato il centro di ricerca CIRCSE, di cui è attualmente Direttore. Dal 2018 è Principal Investigator di un ERC-Consolidator Grant (LiLa: Linking Latin).

PROUDLY POWERED BY WORDPRESS | THEME: EDITOR BY ARRAY