

LiITA, una Knowledge Base di risorse interconnesse per l'italiano

Eleonora Litta, Marco Passarotti, Paolo Brasolin¹

Valerio Basile, Cristina Bosco, Andrea Di Fabio²

¹ Università Cattolica del Sacro Cuore, Milano, Italia – eleonora.litta@unicatt.it

² Università di Torino, Italia – valerio.basile@unito.it

ABSTRACT (ITALIANO)

Il progetto LiITA (<https://www.liita.it/>) si propone di rendere interoperabile e quindi sfruttare la ricchezza di risorse linguistiche disponibili per la lingua italiana (corpora, edizioni digitali, lessici, dizionari e glossari) attraverso l'applicazione dei principi Linked Open Data. Questo articolo descrive i primi passi per lo sviluppo di una Knowledge Base aperta, in cui le risorse linguistiche sono rese interoperabili attraverso il collegamento delle voci delle risorse lessicali e delle occorrenze delle risorse testuali alle forme di citazione registrate in una raccolta di lemmi (Lemma Bank) della lingua italiana. Seguendo i principi del paradigma Linked Open Data, l'interoperabilità concettuale tra le risorse distribuite connesse in LiITA è ottenuta applicando un vocabolario per la descrizione della conoscenza comunemente utilizzato nel mondo dei Linked Open Data linguistici, adottando modelli ampiamente utilizzati, tra cui il principale è OntoLex-Lemon.

Parole chiave: italiano; Linked Open Data; interoperabilità; RDF

ABSTRACT (ENGLISH)

The LiITA project, a Knowledge Base of interconnected resources for Italian. The LiITA project (<https://www.liita.it/>) aims to make interoperable and thus exploit the wealth of linguistic resources available for the Italian language (corpora, digital editions, lexicons, dictionaries and glossaries) through the application of Linked Open Data principles. This paper describes the first steps for the development of an open Knowledge Base, in which linguistic resources are made interoperable by linking the entries of lexical resources and the occurrences of textual resources to the citation forms recorded in a collection of lemmas (Lemma Bank) of the Italian language. Following the principles of the Linked Open Data paradigm, conceptual interoperability between the distributed resources connected in LiITA is achieved by applying a vocabulary for the description of knowledge commonly used in the world of linguistic Linked Open Data, adopting widely used models, among which the main one is OntoLex-Lemon.

Keywords: Italian, Linked Open Data, interoperability, RDF

1. INTRODUZIONE

L'italiano è tra le lingue più ricche per numero di risorse linguistiche digitali, sia lessicali che testuali. Una ricerca effettuata sul CLARIN Virtual Language Observatory,¹ un'interfaccia che permette di trovare risorse linguistiche in vari domini, restituisce più di 8.000 risultati per la lingua italiana. Come altre lingue ricche di risorse, per l'italiano esistono risorse fondamentali, tra cui WordNet (Pianta et al. 2002 e Roventini et al. 2016), treebank mantenute nel repository di Universal Dependencies,² corpora storici³ e corpora di riferimento e monitoraggio della lingua scritta (*CORIS/CODIS*, Favretti et al. 2002) e parlata (*KIParla*, Mauri et al. 2019). Tuttavia, analogamente a quanto osservato per molte altre lingue, le risorse linguistiche disponibili per l'italiano presentano rilevanti eterogeneità in termini di formato dei dati, criteri di annotazione e tagset adottati. Questa variabilità ostacola l'integrazione e l'interoperabilità dei (meta)dati provenienti dalle diverse risorse, con effetti negativi sulla fruibilità delle stesse e sullo sviluppo di studi linguistici. Tali risorse, infatti, differiscono non solo per il tipo di informazioni fornite, ma anche per il livello di granularità applicato alla rappresentazione dello stesso oggetto linguistico, ovvero la parola, sia essa intesa come occorrenza in un corpus o come voce di una risorsa lessicale. Far interagire tale ricchezza di informazioni per sfruttare al meglio l'enorme patrimonio dei (meta)dati raccolti in decenni di lavoro, rappresenta una delle principali sfide odierne, non solo per quanto riguarda l'italiano, ma per tutte le lingue. Una linea di ricerca molto attiva si focalizza al momento sui cosiddetti Linguistic Linked Open Data (LLOD). L'obiettivo è quello di definire pratiche comuni per la rappresentazione e la pubblicazione di risorse

¹ <https://vlo.clarin.eu> (cons. 07/01/2025).

² <https://universaldependencies.org> (cons. 07/01/2025).

³ <https://www.corpusmidia.unito.it/>, <http://www.oivi.cnr.it/> (cons. 07/01/2025).

linguistiche secondo i principi del paradigma Linked Data, che sta alla base del Semantic Web. La Knowledge Base (KB) LiLa (Linking Latin) è probabilmente il principale caso d'uso LLOD attualmente disponibile. LiLa è una KB di risorse linguistiche per il latino rese interoperabili grazie alla loro rappresentazione e pubblicazione secondo i principi dei Linked Data. Sulla base dell'esperienza maturata con il progetto LiLa, e mediante il riutilizzo della relativa architettura, il progetto LiITA: Linking Italian⁴ si propone di realizzare una Knowledge Base (KB) nella quale le risorse linguistiche per l'italiano, pubblicate secondo i principi dei Linked Data, risultino pienamente interoperabili. Il presente contributo descrive lo sviluppo della componente centrale della KB LiITA: la cosiddetta "Lemma Bank", ovvero un repertorio di lemmi della lingua italiana concepito come nodo di interconnessione tra le diverse risorse linguistiche incluse nella KB.

2. LINKED DATA

Il concetto di Semantic Web⁵ si fonda sull'idea che i contenuti pubblicati sul World Wide Web debbano essere arricchiti con informazioni e metadati strutturati in modo da renderli accessibili, interrogabili e interpretabili non solo dagli esseri umani, ma anche da sistemi computazionali. Tale strutturazione è realizzata attraverso i Linked Data, che costituiscono l'infrastruttura portante del Semantic Web. A differenza di un Web fatto di ipertesti, in cui i link non sono interpretabili semanticamente, il Semantic Web è caratterizzato da collegamenti tra entità (o "oggetti") identificati univocamente mediante URI (Uniform Resource Identifier). I collegamenti tra oggetti nel contesto del Semantic Web acquisiscono significato semantico in quanto espressi mediante vocabolari controllati e formalizzati sotto forma di ontologie, ovvero modelli condivisi per la descrizione strutturata della conoscenza.

Il paradigma dei Linked Data si articola su quattro principi fondamentali:

1. Utilizzo degli URI come identificatori univoci e persistenti: ogni entità è identificata da un URI, che ne garantisce la disambiguazione e la tracciabilità nel tempo. Nel caso dei (meta)dati linguistici, le entità identificate includono oggetti (meta)linguistici quali occorrenze di parole in testi, lemmi nei dizionari, o parti del discorso;
2. Adozione di URI HTTP per l'accesso sul Web: ciò consente sia agli utenti umani sia ai sistemi automatici di interrogare direttamente le risorse identificate;
3. Utilizzo di standard semantici per la rappresentazione e il recupero delle informazioni: in particolare, RDF (Resource Description Framework) costituisce il modello dati alla base del Semantic Web. In RDF, l'informazione è strutturata in triple, composte da un soggetto, una proprietà e un oggetto, le cui relazioni semantiche sono definite da ontologie condivise. SPARQL⁶ (SPARQL Protocol and RDF Query Language) è il linguaggio di interrogazione standard per dati rappresentati in RDF, e consente operazioni complesse di recupero e inferenza sui (meta)dati;
4. Creazione di collegamenti tra URI: la presenza di riferimenti ad altri URI all'interno delle risorse consente la navigazione semantica tra dati distribuiti.

L'applicazione dei principi del paradigma Linked Data ai (meta)dati derivati da risorse linguistiche e la loro pubblicazione sul Web offre diversi vantaggi (Chiaros et al. 2013). In primo luogo, per quanto riguarda la rappresentazione e la modellazione dei (meta)dati, RDF è un modello particolarmente versatile, adatto a rappresentare metadati come quelli veicolati dai vari livelli di annotazione disponibili nelle risorse linguistiche (morfologia, sintassi, lemmatizzazione, ecc.). Inoltre, l'adozione di un modello di dati comune (RDF) consente due livelli di interoperabilità: l'interoperabilità strutturale (o sintattica), ovvero la capacità di sistemi diversi di elaborare dati scambiati utilizzando protocolli e formati condivisi (come HTTP e URI), e l'interoperabilità concettuale (o semantica), ovvero la capacità di un sistema di interpretare automaticamente e semanticamente le informazioni scambiate utilizzando un insieme comune di classi e categorie di dati definite in ontologie e vocabolari (Ide & Pustejovsky 2010).

⁴ <https://www.liita.it/>.

⁵ Introdotto da Tim Berners-Lee et alii 2001.

⁶ <https://www.w3.org/TR/rdf-sparql-query/> (cons. 08/04/2025).

Sebbene la lingua italiana non sia estranea ai Linked Data,⁷ LiITA rappresenta il primo tentativo sistematico di realizzare una risorsa strutturata sotto forma di Lemma Bank, concepita attorno alla parola come unità centrale e punto di intersezione tra le diverse tipologie di risorse linguistiche.

3. LA KNOWLEDGE BASE LiITA: ARCHITETTURA E MODELLO

La componente principale di LiITA consiste in una raccolta di forme canoniche di citazione (lemmi) per la lingua italiana, accessibile dal sito <https://www.liita.it/>, e scaricabile dalla repository del progetto.⁸

L'architettura della KB di LiITA si rifà a quella della LiLa KB per il latino,⁹ condividendone l'assunto centrale secondo cui la parola costituisce l'elemento cardine per l'interoperabilità tra risorse linguistiche: le risorse lessicali, come dizionari o lessici, che descrivono le proprietà delle parole, sono costituite da voci lessicali, e cioè parole, e le risorse testuali, come corpora e biblioteche digitali che consistono in raccolte di testi, sono costituite da token, e cioè occorrenze di parole.

Il collegamento ai rispettivi lemmi rende interoperabili in LiITA le voci lessicali e le occorrenze delle parole provenienti da risorse sparse e prive di interoperabilità, consentendo l'esecuzione di ricerche che spaziano tra le diverse risorse linguistiche. Ad esempio, è possibile cercare tutte le occorrenze del lemma *andare* in più corpora; oppure estrarre tutti quei token che hanno determinate proprietà fornite da una o più risorse lessicali.

Dato dunque il ruolo centrale delle parole nell'architettura di LiITA, la componente principale della KB è una raccolta di lemmi, forme di citazione lessicale adottate abitualmente nelle risorse linguistiche, chiamata Lemma Bank. Questi lemmi trovano corrispondenza, da un lato, nelle voci delle risorse lessicali e, dall'altro, nelle forme scelte per raccogliere tutte le occorrenze di una particolare parola nelle risorse testuali lemmatizzate.

Seguendo i principi del paradigma Linked Data, l'interoperabilità concettuale tra le risorse distribuite connesse in LiITA viene ottenuta utilizzando categorie per la descrizione della conoscenza comunemente utilizzate nel mondo dei Linguistic Linked Open Data. Nel caso specifico della Lemma Bank di LiITA, è stato adottato il modello specificato da OntoLex-Lemon (McCrae et al. 2017), uno dei modelli RDF più utilizzati per rappresentare e pubblicare le risorse lessicali Linked Data. Inoltre, i lemmi sono descritti in un'ontologia personalizzata basata sull'ontologia del progetto LiLa. Questa ontologia, da un lato, fornisce informazioni dettagliate su alcune caratteristiche morfologiche e linguistiche dei lemmi (ad esempio la parte del discorso e il genere grammaticale per i sostantivi) basandosi sul modello di annotazione OLiA [Cimiano et al., 151-155]. D'altro canto, l'ontologia LiLa definisce classi e proprietà per modellare il compito di lemmatizzazione, come la proprietà `lila:hasLemma`¹⁰ che collega i lemmi ai token di un corpus. Le entrate lessicali (`ontolex:LexicalEntry`)¹¹ sono intese come unità di analisi del lessico che possono raccogliere una o più forme (`ontolex:Form`)¹² e uno o più sensi lessicali (`ontolex:LexicalSense`)¹³, concetti lessicali (`ontolex:LexicalConcept`)¹⁴ o entità da ontologie.¹⁵ Una parola può inoltre essere scritta in forme diverse, chiamate varianti grafiche (written representations), rappresentate tramite la proprietà `ontolex:writtenRep`,¹⁶ e avere una o più varianti fonetiche (proprietà `ontolex:phoneticRep`)¹⁷. Una di queste

⁷ Esempi di risorse per l'italiano al momento disponibili come Linked Open Data sono il lessico Compl-it (<http://hdl.handle.net/20.500.11752/ILC-1007>, cons. 07/01/2025), ItalWordNet v.2 (<http://hdl.handle.net/20.500.11752/ILC-66>, cons. 07/01/2025) e una raccolta di nomi dal lessico PAROLE SIMPLE CLIPS (PSC) (<http://hdl.handle.net/20.500.11752/ILC-558>, cons. 07/01/2025).

⁸ https://github.com/LiITA-LOD/LiITA_LemmaBank.

⁹ <https://lila-erc.eu/data-page/> (cons. 08/04/2025).

¹⁰ <http://lila-erc.eu/ontologies/lila/hasLemma>.

¹¹ <http://www.w3.org/ns/lemon/ontolex#LexicalEntry>.

¹² <http://www.w3.org/ns/lemon/ontolex#Form>.

¹³ <http://www.w3.org/ns/lemon/ontolex#LexicalSense>.

¹⁴ <http://www.w3.org/ns/lemon/ontolex#LexicalConcept>.

¹⁵ Per sensi lessicali si intendono i sensi lessicalizzati: un senso corrisponde a una singola voce lessicale. Gli aspetti semantici che possono essere espressi da più parole sono invece rappresentati tramite concetti lessicali (Lexical Concept), che possono quindi avere più di una lessicalizzazione.

¹⁶ <http://www.w3.org/ns/lemon/ontolex#writtenRep>.

¹⁷ <http://www.w3.org/ns/lemon/ontolex#phoneticRep>.

forme, l'oggetto della proprietà `ontolex:canonicalForm`,¹⁸ è la forma che viene convenzionalmente scelta per rappresentare l'intero gruppo delle forme flesse di una voce lessicale. La Lemma Bank di LiITA consiste in una raccolta di queste forme, modellate nell'ontologia come "individui" della Classe `lila:Lemma`,¹⁹ a sua volta una sottoclasse di `ontolex:Form`, creata originariamente per il progetto LiLa²⁰ e adottata di conseguenza nella LiITA Lemma Bank.

I lemmi della Lemma Bank di LiITA non sono tuttavia vincolati da alcuna relazione con una voce lessicale, poiché la Lemma Bank non è una raccolta di voci lessicali, ma un set di forme canoniche di citazione. Questo riflette il ruolo peculiare della Lemma Bank come strumento per rendere interoperabili le risorse. Utilizzando la proprietà `lila:hasPos`,²¹ a ciascun lemma nella Lemma Bank viene assegnata una parte del discorso, utilizzando il tagset UPOS di Universal Dependencies. Nel caso in cui ci siano parole a cui sono assegnate più PoS nelle diverse risorse lessicali, vengono creati nella Lemma Bank più lemmi corrispondenti. Ad esempio, alla parola "rosso" vengono assegnate due PoS: aggettivo e sostantivo e nella Lemma Bank vengono creati due lemmi distinti con due diverse PoS rappresentate tramite la proprietà `lila:hasPos`.

Per armonizzare eventuali diversi criteri di lemmatizzazione applicati nelle molteplici risorse linguistiche disponibili per l'italiano, la Lemma Bank di LiITA implementa due proprietà specifiche aggiuntive. La prima è la proprietà simmetrica `lila:lemmaVariant`²² che collega diverse forme del paradigma flessivo di una parola che possono essere utilizzati come lemmi indipendenti. Il caso tipico è quello dei nomi con doppio plurale (nomi in *-o* con plurale maschile in *-i* e femminile in *-a*, come per esempio *braccio/bracci*, *braccia* in cui la differenza nel plurale indica una decisa differenza semantica).²³ Queste parole possono essere lemmatizzate sia nella forma plurale che nella forma singolare: grazie alla proprietà `lila:lemmaVariant` i due `lila:Lemma` *braccio* e *braccia*, sono collegati tra loro nella Lemma Bank. La seconda proprietà, `lila:hasHypolemma`²⁴ (e la sua proprietà inversa `lila:isHypolemma`)²⁵ collega i lemmi che fanno parte del paradigma flessivo di una stessa parola ma hanno parti del discorso diverse. È il caso degli aggettivi usati come avverbi, ad esempio *veloce* che può essere interpretato (e lemmatizzato) sia come una forma di aggettivo (quindi modellato come un `lila:Lemma`) sia come un avverbio (quindi modellato come un `lila:Hypolemma`,²⁶ una sottoclasse di `lila:Lemma`).

I participi passati e presenti sono un altro tipo di ipolemma (si vedano ad esempio il participio passato *narcotizzato* e il participio presente *narcotizzante*), a cui nella Lemma Bank viene assegnata la PoS aggettivo. I participi sono modellati come individui della classe `lila:Hypolemma` e sono collegati al loro lemma verbale (*narcotizzare*) tramite la proprietà `lila:isHypolemma`. Prescindendo quindi dal fatto che due differenti risorse lemmatizzino i participi secondo criteri diversi (vale a dire, che siano essi ricondotti ad una forma di citazione participiale o verbale), la Lemma Bank considera entrambe le casistiche e armonizza entrambe le possibilità in una.

I lemmi e l'informazione sulla parte del discorso che costituiscono la Lemma Bank si basano sulla base lessicale di una versione online del dizionario Nuovo De Mauro,²⁷ che ammonta complessivamente a circa 145.000 voci; di queste, 13.000 polirematiche sono state ritenute non rilevanti per gli scopi del progetto, in quanto i lemmatizzatori operano tipicamente su singoli token. Fino ad ora, circa 125.200 lemmi, e 16.330 ipolemmi sono stati estratti dalle restanti 131.000 voci. La logica di categorizzazione della parte del discorso della base lessicale originaria è stata adottata parzialmente e mappata sul tagset UPOS.²⁸ Poiché la categorizzazione grammaticale originaria del dizionario segue la tradizione linguistica italiana, è stato necessario armonizzare alcune parti del discorso con il sistema Universal Part-of-Speech (UPOS). In particolare, le congiunzioni, che nella fonte originale non distinguono tra forme coordinate e subordinate,

¹⁸ <http://www.w3.org/ns/lemon/ontolex#canonicalForm>.

¹⁹ <http://lila-erc.eu/ontologies/lila/Lemma>.

²⁰ <http://lila-erc.eu/ontologies/lila/>.

²¹ <http://lila-erc.eu/ontologies/lila/hasPOS>.

²² <http://lila-erc.eu/ontologies/lila/lemmaVariant>.

²³ Cf. Serianni, 1989, p.143.

²⁴ <http://lila-erc.eu/ontologies/lila/hasHypolemma>.

²⁵ <http://lila-erc.eu/ontologies/lila/isHypolemma>.

²⁶ <http://lila-erc.eu/ontologies/lila/Hypolemma>.

²⁷ <https://dizionario.internazionale.it/> (cons. 08/04/2025).

²⁸ Universal Part of Speech Tagset, <https://universaldependencies.org/u/pos/>.

sono state riallineate manualmente ai tag UPOS. Inoltre, l'estrazione iniziale ha temporaneamente escluso alcune categorie marginali, come prefissi, segni di punteggiatura e alcune abbreviazioni, il cui trattamento è previsto in fasi successive del progetto.

4. Conclusione e lavoro futuro

In questo contributo sono stati illustrati i primi sviluppi del processo di pubblicazione, secondo i principi dei Linguistic Linked Open Data (LLOD), di una raccolta di forme canoniche di citazione (lemmi) per la lingua italiana. Tale raccolta costituisce il nucleo della Lemma Bank, componente centrale del progetto LiITA, una knowledge base pensata per favorire l'interoperabilità tra risorse linguistiche italiane, sulla scia dell'esperienza maturata con la knowledge base LiLa per il latino. LiITA si propone di colmare l'attuale frammentazione delle risorse linguistiche italiane, fungendo da nodo centrale per l'integrazione di lessici e corpora esistenti e futuri. A tal fine, la Lemma Bank è progettata per accogliere e armonizzare differenti criteri di lemmatizzazione adottati nelle varie risorse, senza alterarne la struttura originaria, ma valorizzandone la specificità e garantendone al contempo la massima riusabilità all'interno dell'ecosistema Linked Data. Il passo immediatamente successivo nello sviluppo della KB LiITA, entro i termini del finanziamento corrente, consiste nel collegamento di una almeno una risorsa testuale e di almeno una risorsa lessicale.

Poiché LiITA intende adottare un processo resource-driven di acquisizione dei dati, la costruzione della Lemma Bank per rendere interoperabili le risorse distribuite per l'italiano adottando il paradigma Linked Data rimane inevitabilmente un processo aperto. Il collegamento di sempre più risorse alla KB richiederà l'inclusione di nuovi lemmi e l'implementazione incrementale della Lemma Bank di LiITA, sia attraverso l'estrazione di lemmi da varie fonti lessicali sia, in modo controllato, dalle risorse testuali.

Oltre all'estensione della Lemma Bank e al collegamento delle prime risorse, il progetto LiITA si propone anche lo sviluppo di servizi online per semplificare l'accesso ai dati. Il processo di collegamento di un testo o corpus nella KB verrà infatti supportato da uno strumento che permette lemmatizzazione automatica, PoS-tagging e collegamento. A questo scopo, è stato recentemente addestrato un nuovo modello Stanza (Qi et al. 2020) testato su tutte le treebank italiane disponibili in Universal Dependencies. Questo modello servirà come base per il processo di collegamento delle risorse testuali da includere nella KB di LiITA.²⁹ L'interrogazione avanzata dei dati offerta da tutte le risorse interconnesse in LiITA sarà facilitata da un'interfaccia grafica che aiuterà nel compito di scrivere complesse query SPARQL.

Infine, data l'architettura language-independent e l'adozione di un vocabolario di descrizione della conoscenza basato su modelli e ontologie ampiamente usati nella comunità LOD, il progetto LiITA può determinare un sostanziale avanzamento nell'accesso, pubblicazione ed utilizzo delle risorse linguistiche per l'italiano, come il progetto pioniere LiLa ha fatto per il latino, applicando una metodologia focalizzata sull'interoperabilità basata sull'approccio Linked Open Data.

RINGRAZIAMENTI

Il presente contributo è finanziato dall'Unione Europea – Next Generation EU, Missione 4 Componente 1, CUP J53D23017270001. Il progetto PRIN 2022 PNRR "LiITA: Interlinking Linguistic Resources for Italian via Linked Data" è realizzato congiuntamente dall'Università Cattolica del Sacro Cuore di Milano e dall'Università di Torino.

BIBLIOGRAFIA

- Berners-Lee, T., Hendler, J., & Lassila, O. (2001). The Semantic Web. *Scientific American*, 284(May 2001), 34–43.
- Chiarcos, C., Moran, S., Mendes, P. N., Nordhoff, S., & Littauer, R. (2013). Building a Linked Open Data Cloud of Linguistic Resources: Motivations and Developments (I. Gurevych & J. Kim, Eds.; pp. 315–348). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-35085-6_12
- Cimiano, P., Chiarcos, C., McCrae, J. P., & Gracia, J. (2020). *Linguistic Linked Data: Representation, Generation and Applications*. Springer International Publishing. <https://doi.org/10.1007/978-3-030-30225-2>

²⁹ Il modello è disponibile sulla repository Github del progetto: https://github.com/LiITA-LOD/LiITA_NLP_Models

- McCrae, J. P., Gil, J., Gràcia, J., Bitelaar, P., & Cimiano, P. (2017). The OntoLex-Lemon Model: Development and Applications. <https://www.semanticscholar.org/paper/The-OntoLex-Lemon-Model%3A-Development-and-McCrae-Gil/3ab2877e3cf9d8f7bad3a4fb9a03602010e00691>
- Favretti, R. R., Tamburini, F., & Santis, C. (2002). CORIS/CODIS: A corpus of written Italian based on a defined and a dynamic model.
- Ide, N., & Pustejovsky, J. (2010). What Does Interoperability Mean, Anyway? Toward an Operational Definition of Interoperability for Language Technology. in: Proceedings of the Second International Conference on Global Interoperability for Language Resources. Hong Kong, China, 2010.
- Mauri, C., Ballarè, S., Gorla, E., Cerruti, M., & Suriano, F. (2019). KIParla Corpus: A New Resource for Spoken Italian. Italian Conference on Computational Linguistics.
- Miller, E. (1998). An Introduction to the Resource Description Framework. *D-Lib Magazine, Journal of library administration*, 34 (2001) pp. 245–255.
- Passarotti, M., Mambrini, F., & Moretti, G. (2024). The Services of the LiLa Knowledge Base of Interoperable Linguistic Resources for Latin. In C. Chiarcos, K. Gkirtzou, M. Ionov, F. Khan, J. P. McCrae, E. M. Ponsoda, & P. M. Chozas (Eds.), Proceedings of the 9th Workshop on Linked Data in Linguistics @ LREC-COLING 2024 (pp. 75–83). ELRA and ICCL. <https://aclanthology.org/2024.lidl-1.10/>
- Petrov, S., Das, D., & McDonald, R. (2012). A Universal Part-of-Speech Tagset. In N. Calzolari, K. Choukri, T. Declerck, M. U. Doğan, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, & S. Piperidis (Eds.), Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC`12) (pp. 2089–2096). European Language Resources Association (ELRA). <https://aclanthology.org/L12-1115/>
- Pianta, E., Bentivogli, L., & Girardi, C. (2002). MultiWordNet: Developing an aligned multilingual database. In First International Conference on Global WordNet (pp. 293–302).
- Qi, P., Zhang, Y., Zhang, Y., Bolton, J., & Manning, C. D. (2020). Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. In A. Celikyilmaz & T.-H. Wen (Eds.), Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations (pp. 101–108). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2020.acl-demos.14>
- Roventini, A., Marinelli, R., & Bertagna, F. (2016). ItalWordNet v.2. <http://www.ilc.cnr.it/it/content/italwordnet>.
- Serianni, L. (1989). Grammatica Italiana, UTET.