Umanistica Digitale – ISSN: 2532-8816 – n. 15, 2023
F. Bianco – Recensione: *Linguistica computazionale. Introduzione all'analisi automatica dei testi*
DOI: http://doi.org/10.6092/issn.2532-8816/17260

# Recensione: Ježek, E. & Sprugnoli, R. (cur.). 2023. *Linguistica computazionale. Introduzione all'analisi automatica dei testi*. Bologna: il Mulino

Francesco Bianco

Univerzita Palackého v Olomouci
francesco.bianco@upol.cz

## Abstract

La presente recensione è dedicata al volume *Linguistica computazionale* di Elisabetta Ježek e Rachele Sprugnoli, un'agile e densa introduzione all'uso di strumenti computazionali nella ricerca linguistica e, più in generale, nel trattamento automatico del linguaggio naturale. Gli otto capitoli del volume, particolarmente adatto ai neofiti, presentano sinteticamente un'ampia varietà di discipline, attività, approcci, strumenti e risorse. Particolare spazio è dedicato, nella trattazione, al ruolo dell'intelligenza artificiale.

**Parole chiave**: IA, intelligenza artificiale, linguistica computazionale, TALN, trattamento automatico del linguaggio naturale.

*This review is dedicated to the book* Linguistica computazionale *by Elisabetta Ježek and Rachele Sprugnoli, which serves as a concise and comprehensive introduction to the use of computational tools in linguistic research and, more broadly, in natural language processing. The eight chapters of this volume provide a succinct overview of a wide range of disciplines, activities, approaches, tools, and resources, making it particularly suitable for beginners in the field. Special attention is given to the role of artificial intelligence throughout the book.*

**Keywords**: AI, Artificial Intelligence, Computational Linguistics, Natural Language Processing, NLP.

## Recensione

Uscito nei primi mesi del 2023, il volume di Elisabetta Ježek e Rachele Sprugnoli arriva sul mercato in un momento di rinnovata attenzione per il settore del trattamento automatico del linguaggio naturale (TALN; ingl. *NLP* 'Natural Language Processing'[1]) e, di riflesso, per la linguistica computazionale. Il recente avvento di ChatGPT – la cui prima versione risale

1 Curiosamente, nel volume non si riporta mai, neppure come mera indicazione terminologica, l'acronimo italiano *TALN* – o *TAL* 'trattamento automatico del linguaggio' –. *TALN* è citato solo come acronimo del progetto italiano *Text Analytics and Natural Language processing* (pp. 139, 231).

al novembre del 2022 –, nel quadro più generale degli stupefacenti risultati dell'intelligenza artificiale (IA) generativa, ha stimolato la curiosità di schiere di utenti e sollecitato un dibattito cui partecipano, in varie sedi, studiosi e intellettuali di estrazione molto varia. Se tale contesto si presentava come una preziosa occasione, si può dire che le autrici abbiano saputo coglierla.

Gli otto capitoli del libro sono segmenti di un percorso coerente e progressivo, attraverso il quale il lettore non (ancora) professionista arriva a farsi un'idea compiuta del campo di studi. Di tale percorso possono giovarsi studenti o studiosi in fase di avvicinamento alla linguistica computazionale, ma anche – azzardiamo – semplici profani animati da una viva curiosità intellettuale per i temi trattati – la fase storica, si diceva, è favorevole a questo genere di inclinazioni –. Ciò è reso possibile dalla felice architettura dell'opera, che può essere suddivisa in due parti principali: la prima, corrispondente ai primi quattro capitoli, fornisce le basi teoriche propedeutiche alla lettura di ciò che segue; la seconda, corrispondente ai secondi quattro capitoli, presenta le attività ascrivibili alla disciplina che costituisce – meglio: alle discipline che costituiscono – l'oggetto del libro.

Nel chiarire di cosa parli il volume, il primo capitolo (*Definizione, scopi e cenni storici*; pp. 17-35) esibisce alcune scelte delle autrici, fondamentali e non scontate: su tutte, la distinzione tra linguistica computazionale – intesa come «studio delle lingue effettuato con metodi computazionali» – e il TALN – «analisi computazionale di testi scritti o orali mirata a risolvere dei compiti» concreti: traduzione automatica, generazione di riassunti, *sentiment analysis*, etc. –, visti come campi separati, benché proficuamente interconnessi (p. 18), e il legame, presentato come indissolubile, tra linguistica computazionale e IA, al punto da definire la prima come «un'area» della seconda (p. 26) – affermazione che, presa alla lettera, escluderebbe qualsiasi applicazione informatica alla ricerca linguistica che non si serva dell'apprendimento automatico dal perimetro della linguistica computazionale. La storia della disciplina, nel quadro più ampio di quella dell'informatica umanistica[2], e la lettura stessa dell'intero volume, attenuano tale identificazione, pur senza nascondere il ruolo centrale svolto dall'IA nella linguistica computazionale e nel TALN contemporanei.

Questa premessa spiega come, dopo aver fornito le *Basi di linguistica* (§ 2, pp. 37-61) e le *Basi di statistica* (§ 3, pp. 63-82) – due capitoli degni di nota per ricchezza, chiarezza, accessibilità e sintesi, che forniscono al lettore inesperto, senza sovraccaricarlo, tutti gli strumenti per affrontare con fiducia i contenuti seguenti –, le due autrici dedichino il quarto capitolo (pp. 83-112) proprio all'*Apprendimento automatico* (ingl. *machine learning*), cuore dell'IA – non solo quella applicata al linguaggio –, presentando in maniera chiara e sintetica problemi, strumenti, metriche, modelli, con particolare attenzione alle architetture basate sulle reti neurali artificiali, che rappresentano lo stato dell'arte per quanto riguarda numerosi aspetti del TALN. Un aspetto, forse marginale ma – a parere di chi scrive – meritorio, in un testo non destinato a esperti di IA, sta nel presentare l'azione di macchine che eseguono compiti di TALN (generazione e classificazione di testi, traduzione automatica, estrazione di conoscenza, etc.) come una *simulazione* del comportamento (p. 23), piuttosto che come un'approssimazione delle facoltà cognitive dell'essere umano: «[s]imulare il comportamento linguistico umano», scrivono le autrici già nella *Premessa* (pp. 13-15), «non significa acquisire intelligenza linguistica nei termini in cui questa è sviluppata e posseduta dagli esseri umani, specialmente se tale simulazione è raggiunta con metodi statistici» (p. 14). Questa

2 Ai rapporti tra l'informatica umanistica e il TALN è dedicato il quadro di approfondimento 7.1 (pp. 191-192).

prospettiva, che ci sembra di poter estendere alla nozione stessa di IA – etichetta, in tal senso, fuorviante –, rende particolarmente preziosa la distinzione, di cui si è detto, tra TALN, inteso come disciplina applicativa e strumentale, e linguistica computazionale: la quale, pur applicando gli stessi strumenti agli stessi dati, ha come fine proprio la comprensione della facoltà umana del linguaggio e delle sue manifestazioni storicamente determinate, ossia le lingue storico-naturali.

Il quinto capitolo (*Semantica distribuzionale e tipi di vettori*; pp. 113-132) costituisce il proseguimento ideale di quello precedente, presentando le rappresentazioni vettoriali delle parole e delle frasi – soprattutto i cosiddetti *embedding* –, che costituiscono le *feature* principali su cui addestrare modelli di apprendimento automatico, soprattutto neurali – per es. *BERT*; cfr. Devlin *et al.* [3] –.

Il sesto capitolo, dedicato all'*Annotazione dei testi* (pp. 133-163), incrocia il dominio della linguistica dei corpora, ponendo l'attenzione su un aspetto fondamentale di ogni ricerca di tipo quantitativo: la raccolta e il trattamento dei dati. Oltre a presentare i diversi livelli di annotazione – corrispondenti ai livelli di analisi linguistica – e i formati di codifica del testo – argomento rilevante non solo per la linguistica computazionale, ma anche per altri domini, come quello della filologia digitale; cfr. Pierazzo [8], Pierazzo & Mancinelli [9] –, le autrici si soffermano sull'annosa questione dell'organizzazione del lavoro di annotazione, con un interessante e attualissimo approfondimento sugli aspetti legali (p. 163).

Nel settimo capitolo sono presentati i principali *Task di «Natural Language Processing»* (pp. 165-192), distinti nelle tre classi del *preprocessing* (per es. la *tokenizzazione*), della classificazione (per es. il *parsing*) e della generazione (per es. la traduzione automatica). Il capitolo, pur nell'estrema sintesi, offre una visione sufficientemente articolata, se non esaustiva, delle possibilità offerte dalle attuali tecnologie. L'approccio ai problemi concreti mostra, sul campo, i complessi rapporti che intercorrono tra linguistica computazionale e TALN e, più in generale, tra ricerca di base – quella sul linguaggio – e ricerca applicata –. Lo studioso di scienze umane troverà in questo capitolo una ricca fonte di ispirazione per applicare gli strumenti dell'analisi automatica del linguaggio alle proprie ricerche.

Chiude il libro una rassegna critica di *Strumenti per l'analisi dei testi* (§ 8, pp. 193-208), che fa da complemento ai capitoli precedenti, articolata in varie classi: archivi e *repository*, *pipeline*, *demo online*, librerie, servizi *cloud*, strumenti per la valutazione, risorse per l'utilizzo dei corpora e strumenti per l'analisi statistica. La scelta di aver riunito la rassegna in coda al volume sortisce due effetti positivi: snellisce la trattazione dei vari argomenti, che guadagna sintesi e chiarezza, e offre al lettore un luogo facilmente individuabile dove cercare ciò che gli occorre per proseguire la propria formazione o mettere in pratica le nozioni apprese. Sotto questo profilo, non avrebbero guastato indicazioni mirate su risorse, fra la moltitudine di quelle disponibili, per approfondire gli argomenti trattati nei vari capitoli del libro.

Seguono l'ultimo capitolo un ricco e aggiornato elento di risorse bibliografiche (pp. 211-223) e di siti web (pp. 225-232; nella versione online, i link sono navigabili)[3].

3 Riguardo alla bibliografia, segnalo alcune imprecisioni che potrebbero fuorviare il lettore: a p. 99 è citato un contributo di «Lecun, Kavukcuoglu & Farabet [5]» – verosimilmente, corrispondente a LeCun *et al.* [5] –, assente nella bibliografia finale; nella stessa pagina, Manning & Schütze [6] è erroneamente citato come «Manning e Schütze [6]»; l'entrata di de Vries *et al.* [4], citato a p. 128 come «Vries *et al.* [4]», è inserita, nella bibliografia, a p. 215, sotto la lettera *D*, cosa che rende difficile

Leggere il libro è un'esperienza gradevole: dettaglio non secondario quando si ha a che fare con un testo destinato principalmente ai neofiti di un settore specialistico. Merito della autrici, capaci di esporre con chiarezza e senza monotonia, evitando i tecnicismi superflui e glossando quelli – non pochi – necessari. Stona, sotto questo punto di vista, la scelta di stampare a caratteri molto piccoli, tali da affaticare il lettore; una scelta dettata, probabilmente, dal desiderio di far confluire, nel numero limitato di pagine concesso ai volumi di una collana nota per la propria agilità, una ragguardevole quantità di nozioni. A tale esigenza si devono, verosimilmente, anche alcune scelte di contenuto: limitare gli esempi pratici, evitare di scendere nei dettagli dell'implementazione degli strumenti presentati – mai le autrici riportano, per es., brani di codice – e non corredare il volume di un glossario finale – le glosse sono intratestuali –. Si tratta, a parere di chi scrive, di scelte abbastanza comprensibili, considerate anche le molte opzioni offerte dal web per approfondire gli argomenti di proprio interesse e la rapida obsolescenza che hanno, in informatica – specie in settori fortemente dinamici come quello dell'IA –, i dettagli tecnici – un algoritmo è decisamente più longevo delle sue implementazioni concrete. Tuttavia, ci si può chiedere se la versione digitale dell'opera, accessibile attraverso la piattaforma Pandoracampus, non avrebbe potuto essere arricchita da qualche materiale e supplementare, piuttosto che riproporre, semplicemente, il testo di quella cartacea.

## References

[1]. Busa, Roberto. *L'analisi linguistica nell'evoluzione mondiale dei mezzi di comunicazione*. In *Almanacco Bompiani: Le applicazioni dei calcolatori elettronici alle scienze morali e alla letteratura*, Milano: Bompiani, 1962, pp. 103-108, 117.

[2]. Busa, Roberto. *The Annals of Humanities Computing:* The Index Thomisticus. In «Computers and the Humanities», 14, 1980, pp. 83-90.

[3]. Devlin, Jacob *et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. In Burstein, Jill *et al.* (a cura di), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Minneapolis, MN)*, Stroudsburg (PA): Association for Computational Linguistics, 2019, pp. 4171-4186.

[4]. de Vries, Wietse *et al. Make the best of cross-lingual transfer: Evidence from PoS tagging with over 100 languages*. In Smaranda Muresan *et al.* (a cura di), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, Stroudsburg, (PA): Association for Computational Linguistics, 2022, pp. 7676-7685.

[5]. LeCun, Yann *et al. Convolutional Networks and Applications in Vision*. In *Proceedings of 2010 IEEE International Symposium on Circuits and Systems*, Paris: IEEE, 2010, pp. 253-256.

[6]. Manning, Christopher D. & Schütze, Hinrich. *Foundations of Statistical Natural Language Processing*, Cambridge (MA): MIT Press, 1999.

trovarla; a p. 191, nel quadro di approfondimento dedicato ai rapporti tra NLP e *digital humanities*, ci si riferisce a un articolo di Padre Roberto Busa del 1962 – verosimilmente, Busa ([1]) –, di cui tuttavia non v'è traccia nella bibliografia – dove è citato, invece, Busa ([2]) –; a p. 173 è citato «Nasar *et al.* [7]», anch'esso assente in bibliografia – forse corrispondente a Nasar *et al.* (2021) –.

[7]. Nasar, Zara *et al. Named entity recognition and relation extraction: State-of-the-art*. In «ACM Computing Surveys», 54, 1, 2021, pp. 1-39.

[8]. Pierazzo, Elena. *La codifica dei testi*, Roma: Carocci, 2005.

[9]. Pierazzo, Elena & Mancinelli, Tiziana. *Che cos'è un'edizione scientifica digitale*, Roma: Carocci, 2020.