DOI: 10.17469/O2106SLI000013

MANUEL BARBERA, ELISA CORINO, CARLA MARELLO, CRISTINA ONESTI

Corpora.unito.it

Il contributo presenta sinteticamente gli esiti del pluriennale lavoro di ricerca presso l'Università di Torino per la creazione di corpora di lingua scritta liberamente interrogabili in rete, nati dal desiderio di analizzare l'italiano, e successivamente altre lingue, nella varietà dei testi – dall'italiano del Duecento alla lingua "digitata" dei gruppi di discussione online, dall'italiano accademico all'italiano di apprendenti non nativi, fino alla lingua che caratterizza l'universo del discorso legale in Italia – varietà di lingua che hanno imposto un significativo sforzo di riflessione (meta)linguistica e computazionale per la messa a punto e la standardizzazione di adeguate strategie di annotazione dei dati.
Se ne forniscono qui i dati descrittivi principali, rimandando alla demo per alcune schermate e *queries* esemplificative.

Parole chiave: corpus linguistics, varietà dell'italiano, CQP, lingua scritta.

1. *Introduzione*

Il portale www.corpora.unito.it, distributore dei corpora approntati in bmanuel.org, affonda le sue radici in un progetto FIRB[1] di una ventina di anni fa e riveste storicamente un ruolo significativo per la linguistica dei corpora italiana, in quanto palestra di allenamento per progetti successivi, in termini di riflessione (linguistica e computazionale) sui principali processi di predisposizione di corpora di lingua scritta e strumenti informatici per il trattamento delle lingue naturali.

Nato dal desiderio di analizzare l'italiano, e successivamente altre lingue, nella varietà dei testi, il progetto ha elaborato negli anni più di un miliardo e mezzo di token, mettendo a disposizione della comunità scientifica cinque corpora ad accesso libero (con due ancora in

[1] "L'italiano nella varietà dei testi. L'incidenza della variazione diacronica, testuale e diafasica nell'annotazione e interrogazione di corpora generali e settoriali": progetto FIRB RBAU014XCF 2001, coordinatore Carla Marello.

fase di implementazione, cfr. par. 6), di cui è possibile visionare una demo al link https://doi.org/10.48448/hkms-vq47 con alcuni esempi d'uso.

La collaborazione iniziale con l'IMS di Stuttgart (*Institute für maschinelle Sprachverarbeitung*) ha apportato elementi fondamentali: un POS-tagger (*Tree Tagger*), il sistema *Corpus WorkBench* (CWB) e soprattutto il *Corpus Query Processor*, CQP, base informatica per tutte le risorse del gruppo.

La preparazione dei corpora con CQP per essere interrogati ha raccolto un'importante sfida in termini di standardizzazione, vagliando un insieme di annotazioni morfosintattiche per parte del discorso e di articolazione interna del testo in paragrafi che potesse valere per tutte le lingue e per tutti i testi.

Tali corpora sono adatti anche per ricerche di tipo testuale, poiché interrogabili senza alcuna restrizione di contesto.

L'attuale interfaccia di interrogazione rende inoltre facoltativa la conoscenza del linguaggio di CQP, proponendo alcuni tasti guidati per l'inserimento di *queries*, volti ad ampliare il bacino d'utenza delle risorse anche ai non addetti ai lavori.

Non trascurabile la riflessione portata avanti con esperti legali interessati ai problemi del diritto d'autore relativamente a banche dati ed altre opere collettive, che è stata necessaria per impostare l'assetto legale dei corpora, proponendo una possibile soluzione con Licenze *Creative Commons Share Alike* (v. Barbera *et al.* 2007).

La dimensione delle risorse analizzate più specificamente in questa sede è di seguito illustrata:

- Corpus Taurinense: 259.299 token
- Athenaeum Corpus: 306.927 token
- NUNC (considerando tutti i sottocorpora in cinque lingue europee): oltre 600 milioni di token per ogni lingua.

2. *Corpus Taurinense*

Il *Corpus Taurinense*, o CT, è costituito da ventidue testi fiorentini della seconda metà del XIII secolo, annotati e completamente disambiguati per parti del discorso, categorie morfosintattiche, genere letterario, caratteristiche filologiche ed articolazione paragrafematica del testo, portando le esperienze e le tecniche più avanzate della linguistica dei cor-

pora dalle lingue moderne a quelle antiche. Costruito, infatti, secondo specifiche EAGLES>ISLE compatibili nel formato CWB e rilasciato sotto licenza *Creative Commons Share Alike*, è liberamente consultabile alla sua homepage http://www.bmanuel.org/projects/ct-HOME.html.

Un'accurata documentazione è disponibile in Barbera (2009), che costituisce anche una sorta di vademecum dell'aspirante costruttore di corpora ed un punto di riferimento in particolare per la linguistica dei corpora dell'italiano antico (cfr. anche Barbera & Marello 2000/2003).

3. *Athenaeum*

Corpus di italiano scritto accademico, *Athenaeum* è stato costruito con testi prodotti dall'Università degli Studi di Torino, POS-taggati e classificati per argomento e tipo testuale.

È costituito da tre componenti: la rivista "L'Ateneo"; la newsletter "Dall'Università"; materiale amministrativo prodotto internamente o per il sito di ateneo UniTo, raccogliendo in totale 306.927 token, 32.221 type, 11.748 lemmi.

Alcune ricerche nel volume Barbera *et al.* 2007 hanno tratto specificamente vantaggio dall'interrogazione gratuita di questa risorsa online.

4. *NUNC*

L'interesse per la varietà dei testi sopra accennata ha trovato un fertile campo di ricerca nell'incontro con i newsgroups, gruppi di discussione a libero accesso rappresentativi di una lingua mediata dal web: un tipo di comunicazione scritta ed offline, ma con un grado di interattività simile a quello della comunicazione faccia a faccia e fonte di molteplici registri linguistici presenti nella lingua "digitata" (cfr. in particolare le riflessioni di Barbera & Marello 2011). Da qui la creazione dei NUNC, *Newsgroups UseNet Corpora*, anche con sottocorpora specialistici, in italiano, inglese, francese, spagnolo, tedesco.

Per la lingua italiana, che rappresenta la sezione più corposa (con più di 280 milioni di token), l'insieme degli scambi dei due corpora Generic-1 e Generic-2 deriva dalle gerarchie complete di *newsgroups.it* e *free.it* (con una successiva suddivisione nelle due parti per mera praticità computazionale).

Il periodo di raccolta dati è compreso negli anni 2002-2006.

Si tratta della suite di corpora su cui maggiormente si sono concentrati studi applicativi di vario genere, sfruttando anche il cotesto ampio restituito dal corpus, che consente ricerche testuali di ampio respiro (cfr. Marello 2007; Corino & Onesti 2012, solo per citare un paio di esemplificazioni tra le tante).

La presenza nei testi di abbreviazioni ed emoticon, così come di frequenti "sporcature" del testo, di spam e post OT ("out of topic") o crossposting, oltre all'abbondanza di testo ripetuto (spesso effetto del quoting) hanno rappresentato non facili scogli dal punto di vista dell'elaborazione informatica, ma sono stati in buona parte ovviati da una complessa preparazione dei testi, attuata attraverso moduli di filtraggio, tokenizzazione e markuppatura.

5. *Valico e Vinca*

Il learner corpus VALICO (italiano di apprendenti stranieri) e VINCA (il suo corpus appaiato, con scritti di studenti italofoni), sono stati progettati in seno a *bmanuel.org* e distribuiti inizialmente da *corpora.unito.it*, ma hanno poi trovato una nuova sede nel sito www.valico.org (cfr. Corino & Marello 2017). VALICO in particolare rappresenta ad oggi uno dei maggiori corpora di italiano di stranieri liberamente accessibili in rete. Per un approfondimento, si veda la demo dedicata: https://doi.org/10.48448/drhb-1918.

Brevemente, i seguenti aspetti li caratterizzano: raccolta di testi a partire da stimoli iconici, corpus e base di dati sociolinguistici degli autori dei testi interrogabili congiuntamente, ricerche attraverso un approccio georeferenziato, integrato con la risorsa orale "le voci di VALICO" ed esercizi tratti dal corpus stesso, possibilità di confrontare le ricerche sul corpus con il corpus VINCA, composto da testi di italofoni a partire dagli stessi stimoli.

Il nuovo arricchimento della risorsa prevede l'annotazione automatica della sintassi (seguendo lo schema di annotazione standard de facto delle *Universal Dependencies*) e la sua correzione manuale in una sezione gold per la quale è stata realizzata anche l'annotazione manuale degli errori degli apprendenti.

6. *Progetti in corso d'opera*

6.1 Jus Jurium

Risorsa attualmente in beta, il corpus giuridico *Jus Jurium* vuole documentare il discorso giuridico oggi esistente in Italia in tutti i suoi generi.

Il corpus è etichettato per parti del discorso ed ha un robusto markup testuale e diplomatico (Barbera & Onesti 2014, Onesti 2011; 2012): tra le sue finalità, in particolare, è infatti quella di poter interrogare in modo "ricco" i testi, intersecando la loro definizione diplomatica con il loro assetto linguistico e testuale.

Jus Jurium è un insieme di più subcorpora, che seguono la "vita" delle leggi dal loro concepimento nelle discussioni parlamentari, alla loro codificazione in regole normative, alla loro applicazione nei procedimenti giudiziari, raccogliendo dunque un ventaglio differenziato di tipi testuali: a tutti si è tentato di attribuire etichette comuni per marcare l'articolazione interna del discorso e poter in futuro muoversi su interrogazioni separate delle parti di testo. Si veda anche Barbera *et al.* 2017.

6.2 Corpus Segusinum

Il *Corpus Segusinum*, ancora in fase di implementazione, è il primo sottocorpus di una auspicabilmente più ampia raccolta di dati scritti da varietà di italiano giornalistico, con una particolare attenzione alla realtà regionale della stampa piemontese, volta anche a colmare una lacuna della *corpus linguistics* italiana nel considerare la stampa a tiratura locale.

Sorto intorno a due intere annate del giornale *La Valsusa*, una delle testate italiane più antiche, punta altresì a ripensare le possibilità di annotazione di varietà di linguaggio giornalistico, proponendosi dunque un obiettivo anche metodologico nell'enucleare tratti peculiari del tipo testuale "articolo di giornale" così come degli altri tipi di testo ricorrenti nella stampa periodica, locale e non.

Con questo strumento si potrà quindi accedere a tradizionali ricerche per lemma e calcoli di frequenze delle occorrenze; a singole parti del discorso grazie al POS-tagging; a ricerche specifiche su titoli, sottotitoli e occhielli; a ricerche specifiche nelle civette di prima pagina (e diversamente negli incipit delle girate); a ricerche mirate per luoghi, rubriche o testatine del giornale; a parole chiave degli articoli e di al-

tri generi testuali talvolta negletti, quali recensioni, inserzioni, echi di cronaca, comunicati stampa, ecc. (v. anche Barbera & Onesti 2010).

Ringraziamenti

Nel corso degli anni numerosi sono stati i contributi al progetto e altrettante le persone cui essere grati: in questa sede un doveroso e sincero ringraziamento va almeno ad Adriano Allora, Simona Colombo, Marco Tomatis e Luca Valle.

Riferimenti bibliografici

Barbera, Manuel. 2007. I NUNC-ES: strumenti nuovi per la linguistica dei corpora in spagnolo. *Cuadernos de filología italiana* XIV. 11–32.

Barbera, Manuel. 2009. *Schema e storia del Corpus Taurinense*. Alessandria: Edizioni dell'Orso.

Barbera, Manuel & Corino, Elisa & Onesti, Cristina (a cura di). 2007. *Corpora e linguistica in rete*. Perugia: Guerra Edizioni.

Barbera, Manuel & Corino, Elisa & Onesti, Cristina. 2017. Linguistica giuridica italiana on line. Dalle banche dati alla linguistica dei corpora. In Tafani, Laura & Ziller, Jacques (a cura di), *Atti della Giornata di studio "Il linguaggio giuridico nell'Europa delle pluralità"*, 7 novembre 2016, Roma, 123–150. Roma: Senato della Repubblica.

Barbera, Manuel & Marello, Carla. 2000/2003. 'Corpus Taurinense: italiano antico annotato in modo nuovo'. In Maraschio, Nicoletta & Poggi Salani, Teresa (a cura di). *Italia linguistica anno Mille – Italia linguistica anno Duemila. Atti del XXIV Congresso internazionale di studi della Società di linguistica italiana (SLI). Firenze 19-21 ottobre 2000*, 685–693. Roma: Bulzoni.

Barbera, Manuel & Marello, Carla. 2011. Tra scritto-parlato, Umgangssprache e comunicazione in rete: i corpora NUNC. In Antonini, Anna & Stefanelli, Stefania (a cura di), *Studi di Grammatica Italiana* XXVII (2008, recte 2011) = *Per Giovanni Nencioni. Convegno Internazionale di Studi. Pisa – Firenze, 4-5 Maggio 2009*, 157–185. Firenze: Le Lettere.

Barbera, Manuel & Onesti, Cristina. 2010. Dalla Valsusa in avanti: i corpora di stampa periodica locale. *Rivista Internazionale di Tecnica della Traduzione* | *International Journal of Translation* 12. 103–116.

Barbera, Manuel & Onesti, Cristina. 2014. Markup testuale ed articolazione diplomatica: linguistica dei corpora per testi giuridici. In Barbera, Manuel & Carmello, Marco & Onesti, Cristina (a cura di), *Traiettorie sulla linguistica giuridica*, 23–35. Torino – Tricase, bmanuel.org: Youcanprint.

Corino, Elisa & Marello, Carla. 2017. *Italiano di stranieri. I corpora VALICO e VINCA*. Perugia: Guerra Edizioni.

Corino, Elisa & Onesti, Cristina. 2012. Agreement and disagreement in newsgroup interaction. In Campagna, Sandra & Garzone, Giuliana & Ilie, Cornelia & Rowley-Jolivet, Elizabeth (eds.), *Evolving Genres in Web-mediated Communication*, Linguistic Insights, vol. 140, 197–213. Bern: Peter Lang.

Marello, Carla. 2007. Does Newsgroups "Quoting" Kill or Enhance Other Types of Anaphors?. In Korzen, Iørn & Lundquist, Lita (eds.). *Comparing Anaphors between Sentences, Texts and Languages. Proceedings of the international symposium held at the Copenhagen Business School*, September 1st-3rd 2005. "Copenhagen Studies in Language" 34, 145–157. Frederiksberg: Samfundslitteratur Press.

Onesti, Cristina. 2011. Methodology for Building a Text-Structure Oriented Legal Corpus. *Comparative Legilinguistics*, 8/2011. 37–50.

Onesti, Cristina. 2012. 意大利语法律语料库的建设 [Construction of an Italian legal corpus]. *Journal of Guangdong University of Foreign Studies*, vol. 23, n. 3 (tradotto da Ge Yunfeng 葛云峰). 42–48.