DOI: 10.17469/O2106SLI000001

EMANUELA CRESTI, MASSIMO MONEGLIA

Introduzione agli Atti del LIV Congresso SLI

1. *Il Congresso*

Il volume raccoglie i testi delle relazioni e delle Demo dei corpora presentate nel LIV Congresso internazionale della Società di Linguistica Italiana "*Corpora e studi linguistici*" tenutosi online. Vorremmo però ripercorrere brevemente quella che è stata la vicenda del Congresso che nasceva sotto i migliori auspici con la ripresa di una tradizione di ospitalità della sede fiorentina, nella quale si era già tenuto nel 2000 il XXXIV Congresso dedicato a "*Italia linguistica anno Mille – Italia linguistica anno Duemila*" (Maraschio & Poggi-Salani 2003). Proprio facendo riferimento alla parte moderna degli studi che erano stati presentati allora, il Congresso intendeva portare una testimonianza dello sviluppo e della diffusione del settore dedicato alla raccolta e analisi di grandi corpora linguistici, che era stato prefigurato in tale occasione e che si è dimostrato poi acquisire un'importanza crescente proprio nel ventennio intercorso sulla base della diffusione e del consolidamento delle nuove tecnologie informatiche. Avrebbe dovuto tenersi a settembre del 2020, ma la pandemia di Covid ci ha subito costretto a posticiparlo di un anno, nella vana speranza che a tale distanza di tempo l'emergenza sarebbe stata superata. Ma nella primavera successiva è stato a tutti chiaro che non sarebbe stato possibile, non solo uno svolgimento in presenza, ma neppure in modalità mista, che per le forti restrizioni di accesso dei partecipanti ne avrebbe vanificato il senso. Dopo una discussione collettiva degli organizzatori fiorentini e dell'esecutivo della Società si è optato per una soluzione interamente da remoto.

È cominciata allora una ricerca un po' affannosa da parte del Comitato organizzatore di una piattaforma che assicurasse, oltre alle modalità di connessione informatica ormai condivise e a disposizione delle sedi universitarie, una tipologia di interazione multipla che con-

sentisse un adeguato svolgimento del Congresso, che in realtà era piuttosto complesso. Esso univa infatti alle lezioni plenarie e alla possibilità di domanda- risposta ad esse collegata, la presentazione di Demo di corpora, la presentazione di poster con i loro brevi *booster* riassuntivi, e infine lo svolgimento in parallelo di sei *work-shop* con la possibilità di passare da uno all'altro. Inoltre, dato il numero significativo di relatori esteri, la cui provenienza svariava su più fusi orari (includendo la gran parte dei paesi europei, ma anche il Giappone, la Cina, fino al Sudamerica), bisognava prevedere orari che permettessero la partecipazione di tutti. La scelta è stata quella di una società, Underline, che è specializzata nella gestione di grandi congressi e di cui nell'Ateneo di Firenze era stata già provata l'efficienza. Di conseguenza anche la lingua di gestione del congresso è stata l'inglese. Questa scelta ha anche comportato un valore aggiunto, ovvero la possibilità di avere un DOI per la registrazione audio-video delle presentazioni a cui gli autori hanno dato il consenso, che sono rimaste a disposizione per un anno sul sito del Congresso, fino al nuovo Congresso SLI, e che rimarranno indefinitamente disponibili nel *repository* di Underline. A conclusione della introduzione il lettore troverà la lista dei DOI da cui è possibile accedere alle presentazioni. Cogliamo quindi l'occasione per ringraziare tutto il *team* di Underline per l'efficienza e per la gentilezza che ci ha dimostrato. Un ringraziamento particolare a Sol Rosenberg, con cui è stato un piacere trattare, e a Damira Mrsic che ci ha seguito passo passo.

Naturalmente la scelta è stata costosa, ed è stata supportata in parte con le iscrizioni dei partecipanti e in parte con fondi di ricerca personali del Comitato organizzatore. Un caldo ringraziamento va all'amministrazione del dipartimento DILEF che ha curato la gestione finanziaria. A riconoscimento di tale impegno, il Direttivo e l'Assemblea dei Soci della SLI hanno stanziato un consistente contributo per la pubblicazione di questi Atti, per il quale, a nome del Comitato organizzatore, esprimiamo riconoscenza.

Dopo queste premesse, vorremmo fare tuttavia un passo indietro tornando al XXXIV Congresso, per ricordare che fu aperto presso il Salone dei 500 di Palazzo Vecchio da Giovanni Nencioni e che la prima relazione plenaria "*L'Italia linguistica in cammino nell'età della Repubblica*" fu tenuta da Tullio De Mauro, di cui era stato da poco pubblicato il GRADIT (De Mauro 2000). Ed è a questi maestri che

dobbiamo molto dei risultati a cui siamo giunti oggi. Fu organizzata in quell'occasione una tavola rotonda sui "*Grandi progetti in corso*", tra i quali possiamo trovare gran parte delle iniziative poi realizzate e presentate ora nel LIV Congresso, secondo una linea che possiamo chiamare di "tradizione e innovazione".

Una consolidata tradizione di raccolta di testi e di studi basati su di essi ha infatti avuto come conseguenza che l'italiano sia una delle lingue di cultura europee maggiormente rappresentate attraverso corpora. A questo proposito la sede fiorentina appariva particolarmente consona allo svolgimento della tematica del Congresso per la sua tradizionale attività di raccolta di corpora scritti, parlati e trasmessi, nonché di lessicografia su corpora storici e moderni. Non possiamo non ricordare l'ininterrotta opera secolare dell'Accademia della Crusca e l'impresa del Vocabolario (OVI), con la pubblicazione digitale del Tesoro della lingua italiana delle origini.

Ma se lo sfruttamento dei corpora si fonda su una importante tradizione lessicografica, esso si è successivamente allargato verso nuove prospettive, comprendendo studi su aspetti fonetici, prosodici, morfo-sintattici, grammaticali, testuali e pragmatici dell'uso linguistico. Esistono inoltre ambiti specifici nei quali l'indagine scientifica necessita di risorse appositamente concepite e realizzate, come la patologia linguistica, la prima acquisizione, l'apprendimento di lingue seconde. In particolare, gli studi sul parlato hanno prodotto grandi corpora, anche con riferimento alle varietà linguistiche dell'italiano, che hanno portato ad avanzamenti della ricerca negli ambiti della fonetica, della sintassi e della prosodia. In tempi recenti, inoltre, le prospettive di sviluppo si sono ulteriormente ampliate. La realizzazione di grandi risorse derivate dalla rete (*web corpora*) ha prodotto un cambiamento di scala dei dati a disposizione, ormai estesi a miliardi di *tokens*. Infine, la facilità di creare risorse multimodali, con sorgenti audio e video, amplia la ricerca sugli eventi comunicativi in contesti naturali.

In ogni caso, il Congresso non intendeva offrire una rassegna esaustiva dei corpora a disposizione in Italia, ma piuttosto fungere da palcoscenico per alcuni di essi, magari sorti in settori meno attesi, dei quali è stata portata in effetti testimonianza di un'ampia fioritura a evidenziarne l'impatto sulla ricerca linguistica. Per una panoramica più completa dei molti corpora italiani rimandiamo a pubblicazioni introduttive a questo settore di ricerca linguistica apparse nell'ultimo

decennio (Riccio 2016; Lenci, Montemagni & Pirelli 2016; Freddi, 2014; Barbera 2013; Cresti & Panunzi 2013).

Seppure evidentemente in modo del tutto parziale, si voleva anche mettere in rapporto l'esperienza italiana con quanto si muove entro il panorama internazionale ai fini di una proficua comparazione di metodologie di raccolta e studio. Sono stati quindi presentati nel Congresso una serie di Demo di corpora di lingue diverse dall'italiano, tra i quali ricordiamo: *The Research and Teaching Corpus of Spoken German (FOLK) and the Database for Spoken German (DGD); Russian oral discourse through the lens of a multichannel corpus; Rhapsodie, a prosodic and syntactic treebank for spoken French; Design and Analyses of Japanese Speech Corpora; Brazilian Portuguese: Spoken, Written and Diachronic Corpora; Corpus Val.Es.Co. 3.0; ESLORA: un corpus de español hablado en Galicia.* Solo il testo di alcuni di questi è presente nel volume, ma di tutti è a disposizione il Demo nel *repository* di Underline.

La risposta alle richieste del temario è stata significativa, ne sono prova i quattordici Demo e le lezioni plenarie che, se anche a seguito dell'esclusione delle sessioni parallele si sono limitate a sedici, hanno ugualmente permesso un'ampia panoramica dei diversi tipi di corpus e degli studi linguistici connessi.

Inoltre, la presentazione di diciassette poster, non compresi nel volume, ha arricchito ulteriormente la discussione nel corso del Congresso e ci fa piacere anticipare che una selezione di questi sarà pubblicata in un numero speciale della rivista *CHIMERA: Romance Corpora and Linguistic Studies*.

I sei workshop, celebrati in parallelo, sono stati organizzati rispettivamente da:

- Silvana Loiero per il GISCEL (*Apprendere e insegnare: il ruolo dei corpora*)
- Chiara Meluzzi e Sonia Cenceschi (*La linguistica forense dalla ricerca scientifica alla pratica legale*)
- Cecilia Andorno, Emilia Calaresu e Andrea Sansò (*La modalità parlata e il suo ruolo nei modelli grammaticali*)
- Francesca M. Dovetto, Tommaso Raso e Patrizia Sorianello (*Le patologie del linguaggio: studi e risorse tra cross-disciplinarietà e inter-disciplinarietà)*

– Silvia Micheli, Federica Da Milano e Gabriele Iannàccaro (*Ibridismo: per una sistematizzazione epistemologica*)
– Giovanna Alfonzetti, Franca Orletti e Emanuele Banfi (*Agire con le-parole e non solo. Indagini empiriche nelle diverse prospettive teoriche e metodologiche*)

Vorremmo infine ricordare le lezioni magistrali degli studiosi invitati che ci hanno onorato con la loro partecipazione: "*Segmenting and annotating multimodal corpus: inspirations and principles from the traditional Chinese medicine*" di Gu Yueguo (The Chinese Academy of Social Sciences), "*Allestimento, sviluppo e fruizione di DanteSearch, corpus delle opere volgari e latine di Dante con annotazione morfologica e sintattica*" di Mirko Tavoni (Università di Pisa) e "*Corpus-based research in English grammar*" di Bas Aarts (University College London). Mentre delle prime ci è stato fornito un testo per la pubblicazione, dell'ultima, che in ogni caso è consultabile tramite il DOI, vogliamo accennare brevemente i contenuti, invero assai rilevanti. **Aarts** fa un'introduzione storica al *Survey of English Usage*, che iniziato nel 1959 da Randolph Quirk, costituisce il primo corpus che raccoglie non solo l'uso scritto dell'inglese ma anche quello parlato, con ciò fondando la *corpus-linguistics* nella sua piena accezione. La compilazione interamente manuale dell'opera comprende la trascrizione ortografica, l'analisi sintattica, la scansione prosodica ed evidenzia una contrapposizione tra il concetto di *pausing unit* e la *sentence* chomskyana. Il modello di un corpus abbastanza contenuto (un milione di *items*), ma completamente processato e corretto, e quindi affidabile, ha portato alla iniziativa nota come ICE (*International Corpus of English*) con la proliferazione in tutti i paesi di lingua inglese di corpora concepiti nella stessa maniera e comparabili fra loro. L'autore propone poi l'affascinante discussione di due casi dibattuti nella linguistica inglese, quello della differenza di *as* e *for* in sintagmi preposizionali con reggenza nominale e quello del mantenimento del valore verbale del participio presente a scapito di un'interpretazione aggettivale. Aarts dimostra che solo l'uso dell'informazione presente nei corpora permette di avvalorare conclusioni su tali questioni, che rimarrebbero altrimenti speculative. Mostra infine il sito web ENGLICIOUS (*English language resource*), rivolto ai maestri e agli insegnanti delle

scuole secondarie come applicazione del *Survey* per sfruttare il corpus e la ricchezza della sua analisi sintattica.

2. *I contributi*

Il volume è diviso in due parti, nella prima sono illustrate le esperienze di costruzione dei corpora orali, anche multimodali, di quelli scritti sia contemporanei che diacronici, di quelli degli apprendenti, le loro finalità, i principi costitutivi e di annotazione, le modalità di accesso e di ricerca messe in atto. Nella seconda parte sono raccolti i contributi dedicati allo sfruttamento delle informazioni presenti nei corpora a fini di ricerca nei diversi settori della linguistica.

La prima parte è a sua volta idealmente divisa in una prima sezione che presenta i contributi relativi ai corpora orali e multimodali, una seconda dedicata alle principali esperienze condotte in Italia riguardo ai corpora di italiano scritto e infine una sezione specificamente dedicata ai corpora della lingua antica.

La prima parte è introdotta dall'intervento teorico e metodologico di **Gu Yueguo** della Chinese Academy of Social Sciences, che in essa presenta i riferimenti teorici sottostanti la lezione magistrale da lui tenuta in apertura del Congresso. Mentre la lezione, disponibile con le altre nel *repository* di Underline, illustra in dettaglio il grande progetto di raccolta di un corpus di parlato cinese "*The Spoken Chinese Corpus of Situated Discourse*", estremamente ampio e volto a testimoniare la produzione linguistica dalla nascita alla morte, *from womb to tomb*, il testo qui pubblicato è una riflessione profonda sulle principali teoresi sviluppate nel secolo scorso a fondamento della *corpus-linguistics*, con particolare focus sui corpora di parlato. Il contributo vale quindi come introduzione più generale a questo ambito di studi linguistici, oltre ad essere comprensivo delle motivazioni che hanno ispirato l'iniziativa cinese.

Gu esamina e discute le proposte dei fondatori della disciplina nella tradizione anglosassone, come quella di Leech (Leech 1992) – con un confronto con i capisaldi della proposta chomskyana razionalista e innatista – ma anche quelle di Chafe, Sinclair e Biber (Chafe 1992; Sinclair 1994; Biber et al. 2000). L'autore coglie nelle diverse esperienze un percorso di avvicinamento allo sviluppo di una *corpus-linguistics*, che, a suo giudizio, dovrebbe riflettere la *real-life experience*

del parlante, ovvero gli stati del parlante intesi nel loro insieme come una *total saturated experience*. È tale esperienza l'entità che costituisce il reale oggetto di ricerca scientifica, rispetto alla quale i linguaggi e le teorie linguistiche valgono solo come entità derivate.

In effetti, una *real-life experience* del parlante non può essere mai catturata da nessun modello e neppure in particolare da una delle possibili *corpus-linguistics*, che distruggono il loro oggetto nel momento stesso in cui cercano di fissarlo. Assunto che l'autore riprende in maniera esplicita dalla tradizione del pensiero taoista. Ma è inevitabile che ci muoviamo in tal senso.

A tal fine, sulla base della teoria ontologica del filosofo e matematico Mario Bunge (Bunge 1984; 2000), viene proposta una "sistematica" che organizza scientificamente "in un tutto" gli stati del parlante, includendo gli stati fisici, affettivi, – con ancora un interessante riferimento al sistema della medicina tradizionale cinese – e sociali, fino alle caratteristiche più proprie della ricerca linguistica, con considerazione, per esempio degli aspetti pragmatici, prosodici e gestuali. All'interno di tale quadro concettuale si fonda quindi una *ideal corpus-linguistics*, nella quale prende senso anche una *practical-linguistics*. Sono così finalizzati in modo scientificamente significativo i diversi strumenti tecnologici e logici attualmente a disposizione, che sono necessari per la costruzione di corpora orientati a diversi scopi, sia pratici che di ricerca linguistica, e possono documentare l'esperienza del parlante nei suoi molti campi.

Illustriamo di seguito con brevi sintesi i contributi riguardanti i corpora di parlato.

Takehiko Maruyama della Senshu University di Tokyo, nel suo articolo "*Designs and Analyses of Japanese Speech Corpor*a", presenta tre importanti risorse orali giapponesi realizzate dal National Institute for Japanese Language and Linguistics (NINJAL), ovvero il Corpus of Spontaneous Japanese (CSJ), il Corpus of Everyday Japanese Conversation (CEJC) e lo Showa Speech Corpus (SSC). Al di là dei criteri di trascrizione e annotazione che rendono comparabili le risorse in questione, appare di interesse generale la strategia di costruzione e il corpus design sviluppato da NINJAL ai fini della rappresentatività dei corpora orali (Koiso et al. 2018).

Anne Lacheret-Dujour, e **Paola Pietrandrea**, rispettivamente dell'Università Paris-Nanterre e dell'Università di Lille, propongono

il contributo *"Rhapsodie: Un treebank prosodico-sintattico per il francese parlato"* che costituisce un modello per l'annotazione multilivello dell'orale. Il corpus trascritto è allineato al suono per fonemi, sillabe, parole e turni. All'analisi delle dipendenze sintattiche è unita significativamente l'analisi macro-sintattica (Blanche-Benveniste et al. 1990) e l'analisi prosodica è distribuita ai livelli del periodo intonativo, delle sue unità interne e delle prominenze. Il modello prevede l'indipendenza di ciascun livello di analisi come premessa allo studio delle correlazioni e in special modo risulta significativa quella tra le dipendenze sintattiche e la realizzazione prosodica.

Emanuela Cresti, Lorenzo Gregori, Massimo Moneglia, Carlota Nicolás e **Alessandro Panunzi**, dell'Università di Firenze, nel contributo "*The LABLITA Speech Resources"* presentano tre risorse, tra le quali il Corpus di riferimento dell'italiano parlato LABLITA, raccolto in Toscana dal 1965 ad oggi (disponibile in rete attraverso la piattaforma ORFEO), il DB cross-linguistico dell'articolazione dell'Informazione IPIC, da questo derivato, e un corpus realizzato perché sia possibile integrare conoscenze sull'orale nell'acquisizione dello spagnolo L2. L'allineamento per enunciati (realizzato attraverso WINPITCH), così come le annotazioni di queste risorse, seguono il criterio della dipendenza dei livelli di annotazione dalla realizzazione prosodica, proprio della Teoria della lingua in Atto (Cresti 2000).

Caterina Mauri e **Silvia Ballarè** dell'Università di Bologna e **Eugenio Goria** e **Massimo Cerruti** dell'Università di Torino illustrano il corpus KIParla, la più recente risorsa di italiano parlato spontaneo rilasciata nel 2019, che comprende circa 70 ore di audio raccolte a Bologna (Modulo KIP) e a Torino (Modulo ParlaTO). Il corpus raccoglie conversazioni in diversi contesti di ambito universitario (lezioni, esami, ricevimenti, interviste semistrutturate e conversazioni libere) che, insieme ai metadati che le accompagnano, permettono di orientare le ricerche sulle variazioni diafasiche (tra soggetti colti) e diatopiche dell'italiano contemporaneo. KIParla è stato trascritto e allineato in ELAN, è aperto all'acquisizione di nuovi moduli ed è liberamente accessibile in rete attraverso la piattaforma NoSketch Engine.

Marco Biffi dell'Università di Firenze e **Francesca Cialdini** dell'Università di Modena e Reggio Emilia ricostruiscono la storia delle più importanti risorse dedicate alla rappresentazione dell'italiano orale nella varietà trasmessa. Queste sono state realizzate in vari pro-

getti a partire dalla metà degli anni 90' del secolo scorso per l'impulso di Nicoletta Maraschio e dell'Accademia della Crusca. Il contributo *"Banche dati per il trasmesso: il LIR e il LIT"* illustra rispettivamente il Lessico dell'Italiano Radiofonico e il Lessico dell'italiano televisivo, la strategia di campionamento per la compilazione delle due risorse e le modalità di interrogazione nelle nuove piattaforme on line.

In anni recenti lo studio dell'orale, tradizionalmente basato solo su dati audio, si è arricchito di componenti multimodali grazie a strumenti *software* robusti come ELAN, che consentono l'annotazione simultanea di tracce audio e video, e la loro integrazione con *speech software* (PRAAT), come mostrano anche i grandi corpora giapponesi e cinesi qui menzionati. **Giorgina Cantalini** della Scuola Paolo Grassi di Milano, nel contributo "*Corpus multimodale annotato per lo studio della gestualità co-verbale nel «parlato-parlato» e nel «parlato-recitato»",* presenta un corpus multimodale che, al di là degli scopi specifici di confronto della varietà recitata con il parlato spontaneo (Nencioni1976), può a suo modo costituire un modello per la costituzione e l'annotazione di un corpus italiano multimodale dotato di una annotazione simultanea di gesto e prosodia.

Federica Cominetti e **Edoardo Lombardi-Vallauri** dell'Università di Roma 3 con **Lorenzo Gregori** e **Alessandro Panunzi** dell'Università di Firenze, nel contributo "*IMPAQTS: un corpus di discorsi politici italiani annotato per gli impliciti linguistici*" presentano ancora una risorsa multimodale, in corso di realizzazione all'interno di un progetto PRIN coordinato da Lombardi-Vallauri e dedicato allo studio dell'implicito nel linguaggio politico. Il contributo dà una descrizione del grande corpus previsto, che va dalle origini della Repubblica ad oggi (più di 10 MW), dei criteri di bilanciamento necessari ai fini della sua rappresentatività e descrive in particolare il complesso schema di annotazione pragmatica necessario alla determinazione degli impliciti.

Come abbiamo anticipato esistono nuovi settori di ricerca sull'oralità e tra questi uno dei più importanti è quello che indaga sulla lingua di soggetti con patologia. **Francesca Dovetto** e il suo gruppo di ricerca all'Università di Napoli Federico II, insieme a **Raffaele Guarasci** dell'istituto ICAR-CNR, nel contributo "*Corpora di Italiano Parlato Patologico dell'età adulta e senile*" presentano tre corpora archiviati presso il Laboratorio scientifico LiSa all'Università di Napoli Federico II**:** *Corpus del parlato schizofrenico* (CIPPS) già

pubblicato in Dovetto & Gemelli 2012, Corpora della demenza in età senile, sia nella patologia più severa (*Alzheimer,* corpus CIPP-ma) che nella fase prodromica in cui si sviluppano i fattori di rischio (*Mild Cognitive Impairment*, Corpus CIPP-mci). Dei corpora sono illustrati i criteri di costituzione e di trascrizione e una ricca rassegna di studi da essi derivati.

Il congresso ha ricevuto un importante contributo per quanto riguarda la linguistica dei corpora in Brasile. In un lavoro collettivo è stata presentata la ricca varietà di corpora sia orali che scritti oggi disponibili per questa lingua, di cui vengono fornite le caratteristiche principali, i siti dei progetti e un elenco delle pubblicazioni ad essi relative. C-ORAL-BRASIL (**Mello & Raso**) costituisce la costola di portoghese brasiliano del corpus del parlato romanzo C-ORAL-ROM (Cresti & Moneglia 2005) di cui segue i principi di annotazione e di corpus design. NURC (dal 1969) è una iniziativa per la costituzione di corpora orali volta a documentare le varietà linguistiche delle principali capitali di stato brasiliane. *NURCdigital* (**Oliveira**) rende ora disponibile il sotto-corpus di Recife in un formato digitale di alta qualità, con annotazioni conformi agli standard internazionali. *Corpus Brasileiro* (**Sardinha**), è un mega corpus di più di un miliardo di parole (8% parlato, 92% scritto) rilasciato nella sua ultima versione nel 2015. Accanto a questo, un ulteriore mega-corpus, il *Corpus do Português* (**Davies**) comprende tre corpora: Storico (45 milioni di parole), Web (1 miliardo di parole) e ADESSO (1,1 miliardi di parole). *Linguateca* (**Freitas**), è un'infrastruttura dedicata ai corpora e alle risorse computazionali che permette l'accesso a risorse portoghesi brasiliane ma anche europee e di altre varietà. L'università USP-São Carlos ha anche sviluppato molti corpora e risorse computazionali (*NILC Corpora* **Aluísio**) dedicati rispettivamente alla valutazione delle competenze e a *treebank* annotati semanticamente. Infine, il corpus diacronico *Tycho Brahe* (**Galves-Chamberlland**) è un corpus di testi portoghesi di autori sia portoghesi che brasiliani nati tra il 1380 e il 1978.

Per quanto riguarda i corpora che documentano l'italiano scritto sia nella sua dimensione sincronica che nella comparazione diacronica sono raccolti i seguenti articoli.

Fabio Tamburini dell'Università di Bologna**,** nel suo contributo "*I corpora del FICLIT, Università di Bologna: CORIS/CODIS, BoLC*

e DiaCORIS" presenta tre grandi corpora, liberamente consultabili sul Web. Le risorse sono state sviluppate negli anni presso l'università di Bologna a partire dalla esperienza seminale del CORIS/CODIS. Possiamo dire che esso costituisce ancora il riferimento di base per la rappresentazione dell'italiano contemporaneo (online dal 2001 e costantemente aggiornato ogni tre anni). DiaCORIS, riproduce la struttura e le varietà testuali di CORIS in una prospettiva diacronica dall'Unità agli anni 2000. BoLC è un corpus bilingue (italiano/inglese) specifico per il confronto della terminologia giuridica nei due sistemi linguistici.

"I Corpora.unito.it", presentati nel lavoro di **Manuel Barbera, Carla Marello, Cristina Onesti** ed **Elisa Corino** dell'Università di Torino, rappresentano una raccolta storica per la linguistica dei corpora in Italia. Nel loro complesso questi comprendono i testi italiani nella loro più ampia varietà scritta: la varietà storica dell'italiano del Duecento (*Corpus Taurinense*), la lingua dei gruppi di discussione online (corpus multilingue *NUNC*), l'italiano accademico (*Athenaeum Corpus*), l'italiano di apprendenti non nativi (*VALICO*) confrontato con quello degli studenti italofoni (*VINCA*), l'universo del discorso legale in Italia (Jus Jurium), per finire con la lingua giornalistica (Corpus Segusinum). I corpora unito sono POS-taggati con *TreeTagger* e sono accessibili in rete.

L'articolo "*Il corpus MIDIA: concezione, realizzazione, impieghi*" di **Paolo D'Achille** di Uniroma3 e **Claudio Iacobini** dell'Università di Salerno presenta un corpus diacronico di lingua italiana che potremmo definire di seconda generazione rispetto ai corpora scritti appena citati. *MIDIA* è un corpus bilanciato della lingua italiana che comprende testi di diverso genere che vanno dall'inizio del XIII alla prima metà del XX secolo per un totale di circa otto milioni di occorrenze. Il corpus è liberamente consultabile in rete. La presentazione dà conto dei criteri di costituzione del corpus e fornisce esempi concreti del suo utilizzo per studi di tipo diacronico, volti in particolare a studi morfo-sintattici nelle varietà dei generi testuali.

Un ambito più specifico di applicazione della linguistica dei corpora è presentato da **Naomy Nagy** dell'Università di Toronto e **Chiara Celata** dell'Università di Urbino nel contributo "*Un corpus per lo studio della variazione sociolinguistica dell'italiano in contesto migratorio*". Il progetto *Heritage Language Variation and Change in Toronto*, ha

prodotto un corpus multilingue che raccoglie la produzione linguistica di dieci comunità alloglotte giunte in quell'area in seguito a ondate migratorie di diversa provenienza. Attraverso i corpora dell'emigrazione si intende studiare come le lingue vengono mantenute e trasmesse e i fattori che ne influenzano il cambiamento tra le generazioni. Il contributo presenta un corpus socialmente stratificato che presenta i dati relativi alla lingua della prima generazione di immigrati e di due generazioni successive e procede in un loro confronto con campioni della produzione di parlanti rimasti nei paesi di origine. In particolare, vengono illustrati alcuni risultati riguardanti il gruppo calabrese.

In seguito alla crescente disponibilità di testi in formato digitale anche per le lingue antiche sono stati sviluppati negli ultimi anni strumenti di Trattamento Automatico del Linguaggio ad essi dedicati. Tuttavia, dato che la loro affidabilità deve essere oggetto di valutazione, diventa necessario sviluppare corpora che funzionino rispettivamente da *trainig set*, su cui allenare gli algoritmi, e da *test set* per valutarne i risultati. **Rachele Spugnoli** dell'Università di Parma, **Matteo Pellegrini**, **Marco Passarotti** e **Flavio Cecchini** dell'Università Cattolica di Milano presentano "*EvaLatin 1.0: un Corpus per la Valutazione delle Tecnologie del Linguaggio Applicate al Latino*". Il corpus contiene testi latini in prosa e poesia sia di epoca classica che di epoca medievale e nella sua realizzazione è stata data particolare attenzione alla rappresentazione della variabilità diacronica e di genere. Con *EvaLatin 1.0* sono stati valutati in particolare i sistemi di lemmatizzazione e annotazione delle parti del discorso.

Siamo particolarmente lieti che anche il Congresso SLI abbia potuto celebrare l'anno dantesco attraverso la lezione magistrale "*Allestimento, fruizione e prospettive di DanteSearch*" tenuta da **Mirko Tavoni** dell'Università di Pisa e con il contributo di **Paola Manni** e **Rossella Mosti** "*Per Dante. Il VD e i corpora dell'italiano antico*". Il testo della lezione di Tavoni è collocato all'inizio della parte specificamente dedicata ai corpora della lingua italiana antica. In esso l'autore illustra le funzionalità, la storia e le attuali linee di sviluppo dell'infrastruttura *DanteSearch*, che è il corpus completo delle opere volgari e latine di Dante con annotazione linguistica in formato XML-TEI. La risorsa è stata concepita già a partire dalla fine del secolo scorso ed ha seguito una complessa evoluzione nella raccolta delle opere latine difficilmente reperibili online, di diverse collezioni di filologia

digitale che intanto si erano sviluppate ma anche di una massa di testi "incontrollati" reperibili in rete. *DanteSearch* è utilizzato nell'ambito di progetti di ricerca contigui, quali il *Vocabolario Dantesco,* il *Vocabolario Dantesco Latino* e il progetto ERC *LiLa-Linking Latin*, mettendo a disposizione funzionalità uniche di ricerca morfologica e sintattica illustrate in dettaglio nel lavoro. Insieme con *DanteSources,* è attualmente ripensato in ottica di *web* semantico, al fine di costituire una base di conoscenza fondata su logiche calcolabili (RDF, OWL).

L'opera del Vocabolario Dantesco (VD) si propone di raccogliere il lessico contenuto nelle opere volgari di Dante. **Paola Manni**, direttore del Vocabolario Dantesco per l'Accademia della Crusca, e **Rossella Mosti**, coordinatrice del Tesoro della lingua italiana delle Origini (*TLIO*), nel loro articolo **"***Per Dante. Il VD e i corpora dell'italiano antico"* oltre a descrivere le ragioni e i criteri metodologici che hanno ispirato l'iniziativa, mostrano come sono utilizzati i corpora e sottocorpora del *TLIO* per l'analisi lessicografica di Dante. Manni nota come lo studio sistematico delle parole coniate o introdotte da Dante, in cui "si coagula la coscienza dell'indice di creatività insito nel lessico dantesco e del suo lascito nell'italiano", sorga solo nel '900 rispetto a una tradizionale esegesi sui significati più sottili della Commedia iniziata già nel Trecento. In anni recenti, poi, le tecnologie informatiche hanno portato a cogliere aspetti che precedentemente si sottraevano a indagini sistematiche ed esaurienti, permettendo in particolare di inquadrare le parole di Dante nel loro contesto storico.

Il contributo di **Giulio Vaccaro** del CNR di Roma "*Rappresentatività e bilanciamento in un corpus di italiano antico: appunti sul Corpus TLIO"* presenta più in generale il *Corpus OVI* dell'italiano antico e in maniera specifica il *Corpus TLIO*, che ne costituisce il cuore. Su di esso si fonda il vocabolario e l'autore mostra come gli sviluppi realizzati negli ultimi anni permettano nuovi tipi di ricerca. Il lavoro propone considerazioni sulla composizione dei corpora ai fini della ricostruzione complessiva del lessico dell'italiano antico, per cui sarebbe necessario un bilanciamento, soprattutto per genere, oltre che per origine geografica e periodo.

La seconda parte del volume comprende, come si diceva, i contributi dedicati alle ricerche linguistiche basate su corpora. I contributi riguardano i livelli di strutturazione dello scritto e dell'orale, le cate-

gorie grammaticali di base, aspetti della linguistica testuale, il lessico in prospettiva diacronica e diatopica, gli italianismi, e infine gli studi sull'acquisizione dell'italiano L2. Anche solo per la diversità degli argomenti in essa presenti, emerge quindi in modo palpabile la varietà dei campi della ricerca linguistica nei quali l'utilizzo dei corpora si è dimostrato altamente significativo.

La sezione è aperta da due lavori che introducono ai problemi della segmentazione interpuntiva e prosodica rispettivamente della lingua scritta e di quella orale. **Angela Ferrari, Letizia Lala** e **Filippo Pecorari**, delle Università di Basilea e di Losanna, nel contributo "*La punteggiatura italiana attraverso i corpora. Teoria, sincronia e diacronia*" espongono i risultati di due grandi progetti attivati tra il 2015 e il 2020 all'Università di Basilea e dedicati allo studio della punteggiatura italiana in prospettiva sincronica e diacronica. Viene evidenziato il ruolo fondamentale svolto dai corpora nell'elaborazione di una teoria della segmentazione della lingua scritta attraverso la punteggiatura e nella descrizione dell'evoluzione storica del sistema interpuntivo. L'applicazione della *corpus-analysis* ha consentito agli autori di mostrare l'inadeguatezza della diffusa interpretazione e concezione della punteggiatura, secondo la quale essa dovrebbe segnalare gli snodi sintattici del testo o indicarne le curve intonative di realizzazione orale. Tali interpretazioni sono smentite dall'osservazione dei dati, che nei corpora confortano un'interpretazione comunicativo-testuale della punteggiatura secondo il *Modello Basilese* (Ferrari et al. 2008). A ogni segno di punteggiatura, quindi, è stata assegnata una definizione riconducibile all'una e/o all'altra delle due funzioni generali, testuale e comunicativa. L'uso dei corpora si è rivelato del resto necessario per tratteggiare le evoluzioni del sistema interpuntivo su larga scala, che caratterizzano il percorso storico dei diversi generi scritturali. Il contributo contiene però una avvertenza metodologicamente importante. La classica modalità di indagine corpus-based (*keyword in context*), nel momento che l'analisi riguardi fenomeni che coinvolgono ampie porzioni testuali, dimostra un proprio limite.

Il lavoro di **Philippe Martin** dell'Università Paris Cité "*Intonation of telephone conversations in a Customer Care service*" è dedicato alla segmentazione prosodica e al ruolo essenziale dell'informazione fornita dai corpora orali per corroborare le teorie in questo campo, spesso basate invece su parlato letto e su frasi di competenza. Lo studio è stato

effettuato su un corpus telefonico popolato dalle richieste dei clienti al servizio di trasporto della regione parigina, derivato dalla collezione di corpora di lingua francese ORFEO. Esso è annotato prosodicamente secondo il modello dell'Autore, "*Incremental storage concatenation model*" (Martin 2015), consistente in una strutturazione prosodica indipendente che assembla le parole prosodiche in una gerarchia con *dipendenza da destra*. Il modello si confronta con le assunzioni tradizionali di tipo fonologico, proponendo il ruolo centrale dei dati in favore di un nuovo modello di intonazione corpus-based. Coerentemente alle previsioni della teoria i contorni melodici sulle vocali accentate indicano rapporti di dipendenza a destra tra i gruppi accentuali, determinando la struttura prosodica degli enunciati analizzati.

Seguono due contributi dedicati a riflessioni teoriche indotte dallo studio dei corpora. Nel primo, **Anna-Maria De Cesare** dell'Università di Dresda, nel suo testo "*La concezione delle congiunzioni e degli avverbi negli schemi di annotazione dei corpora d'italiano scritto: breve ricognizione e alcune proposte",* affronta il tema sensibile dell'annotazione delle parti del discorso (PoS) nei corpora di italiano scritto, essenziale per il loro sfruttamento ai fini della ricerca linguistica. L'articolo è focalizzato sulle PoS relative alle congiunzioni e agli avverbi, e ricostruendo la loro concezione teorico-descrittiva nella linguistica dei corpora dell'italiano, rileva quanto essa sia vicina a quella della grammatica tradizionale. L'emergenza di elementi innovativi derivati dal lavoro di analisi su corpus permette invece all'autrice di formulare nuove proposte per la revisione delle due PoS, anche avvalendosi delle recenti messe a punto della ricerca teorica e delle infrastrutture computazionali.

Il contributo di **Jørn Korzen**, della Business School di Copenhagen, "*Cosa ci rivelano i corpora sulla complessità testuale dell'italiano*?" ci porta nell'ambito della linguistica del testo e in maniera specifica dello studio comparativo cross-linguistico della complessità testuale. Sulla base di corpora paralleli italiani e danesi (*Europarl*, *Mr. Bean, SugarTexts)* la maggiore complessità testuale dell'italiano è dimostrata da fenomeni computabili, quali il numero di proposizioni nel periodo e la testualizzazione finita vs. non-finita, ossia il grado della deverbalizzazione, che caratterizza in maniera tanto significativa l'italiano. Il peso dei due fenomeni è diverso nelle lingue indagate, tanto che la forma della testualizzazione danese, almeno come esemplificata, po-

trebbe apparire banale e semplicistica. Naturalmente, l'autore è ben consapevole che i due fenomeni trattati non esauriscono la definizione della complessità di una lingua, che dipende da una composizione e interrelazione di molti altri aspetti. Tuttavia, è proprio con l'ausilio di corpora comparabili che possiamo documentare in modo oggettivo quelle differenze di compattezza e di densità testuale, che possono giustificare la valutazione dell'italiano come "una lingua complessa" agli occhi/orecchie di un parlante danese.

Tre contributi sono poi dedicati al lessico e riguardano rispettivamente la variazione diatopica nell'italiano antico, i processi diacronici sottostanti alla formazione delle strutture polirematiche in italiano e lo studio sia in sincronia che in diacronica degli italianismi nelle lingue del mondo.

Maria Francesca-Giuliani, ricercatrice dell'OVI, nel suo testo "*Sulla diatopicità del repertorio lessicale degli antichi testi italiani"* offre un saggio dedicato alla rappresentazione del lessico dell'italiano delle origini nel *Tesoro della Lingua Italiana delle Origini* (*TLIO*) che è il corpus di riferimento "virtualmente" rappresentativo della situazione linguistica dei più antichi testi di area italiana, significativamente in assenza di una norma linguistica unitaria. L'articolo pone il problema del peso della variazione diatopica nell'assetto complessivo del repertorio lessicale del *TLIO*, attraverso valutazioni di ordine qualitativo e quantitativo, avvalendosi delle risorse di gestione allestite dall' OVI. Vengono discusse le strategie di accertamento della rappresentazione delle differenze lessicali legate a circuiti locali. L'articolo fornisce a conclusione un esempio relativo all' individuazione del lemmario marcato in diatopia del *Commento* alla Commedia del bolognese Iacopo della Lana. L'autrice indica una lista significativa di "lessemi ad attestazione monotestuale dei quali circa 1/7 del totale sono accertabili come localismi d'ambiente settentrionale e più precisamente emiliano-bolognese o veneto-veneziano". Questi presentano proprio quei caratteri psico-mentali del "vocabolario d'alta disponibilità", vincolato al quotidiano e al locale, che per esempio nella prospettiva del GraDIt è considerato parte integrante del lessico comune.

Il Contributo "*Diacronia e sincronia delle polirematiche con struttura preposizionale: un'analisi su corpora*" di **Vittorio Ganfi** dell'Università di Modena e Reggio Emilia e di **Valentina Piunno** dell'Università di Roma Tre, affronta dal punto di vista diacronico il sistema

delle polirematiche italiane, specificamente quelle con struttura di sintagma preposizionale. Mettendo a confronto latino, italiano antico e italiano contemporaneo – su dati rispettivamente estratti da *PHI Latin Texts, TLIO*, *OVI*, *corpus la Repubblica, corpus ITTenTen16*, *GraDIt* – è descritto il percorso storico che ha portato alla diversificazione funzionale del sistema di questo tipo di polirematiche. Ai fini dell'analisi, le polirematiche vengono distinte in relazione alla loro forma (struttura sintagmatica, tipo di preposizione e lessemi impiegati) e alla funzione che possono svolgere in contesto. Tra i risultati più significativi del lavoro può essere segnalato sul piano sincronico il consolidamento di alcune strutture con la loro lessicalizzazione o la "costruzionalizzazione" di uno schema, mentre sul piano diacronico emergono percorsi di generalizzazione di uno schema con possibilità di aumento o perdita di produttività nel tempo.

Matthias Heinz dell'Università di Salisburgo e **Lucilla Pizzoli** dell'Università degli Studi Internazionali di Roma nel contributo "*I vantaggi della ricerca su corpora per l'ampliamento e la verifica dei dati dell'OIM*" discutono le potenzialità rappresentate dalla ricerca linguistica su corpora, rappresentativi delle varietà linguistiche nelle quali sono stati censiti gli italianismi, secondo le prospettive di ricerca dell'Osservatorio degli italianismi nel mondo (OIM). I corpora consentono di rintracciare nell'uso l'effettiva circolazione dei prestiti e di misurarne il peso in modo più preciso rispetto alle fonti lessicografiche, nelle quali intervengono a volte fattori ideologici con sovrastima o sottostima del lemma in oggetto derivanti dall'impatto della lingua straniera sulla lingua ricevente. Per quanto riguarda la sincronia, poi, non sempre i neologismi sono censiti nei repertori lessicografici e anche in prospettiva diacronica emergono dati interessanti sugli scambi lessicali, viene così allargata la prospettiva complessiva della ricerca dei prestiti nella direzione del passato e del presente ma anche di comprensione delle tendenze in fieri.

Conclude il volume un lavoro che sfrutta le importanti basi di dati di apprendenti dell'italiano realizzate all'Università per Stranieri di Siena. **Andrea Listanti** e **Liana Tronci**, dell'Università per Stranieri, presentano lo studio, "*Ordini di apprendimento di strutture VS in Italiano L2: Uno studio sul corpus LIPS",* nel quale lo sfruttamento dei corpora degli apprendenti di italiano L2 è dedicato alla definizione dei processi di acquisizione della sintassi italiana in una delle sue strut-

ture marcate (il soggetto in posizione post-verbale). In letteratura, la posizione preverbale è associata generalmente al *topic* e quella post-verbale al *focus,* di conseguenza l'ordine VS nella maggior parte dei casi risulta pragmaticamente marcato perché il soggetto non ricopre la prototipica funzione topicale. Lo studio, coerentemente alle previsioni della *Teoria della Processabilità*, rivela un pattern di apprendimento produttivo a partire dai verbi che prevedono tale ordine in modo non pragmaticamente marcato, ma identifica nei corpora anche processi diversi, che si basano su acquisizioni formulaiche.

Vorremmo infine concludere questa introduzione esprimendo la nostra speranza che dall'insieme delle esperienze presentate il lettore possa apprezzare in maniera sempre più chiara la concezione di cosa sia un corpus, le problematiche legate alla gestione dei dati, gli indispensabili sistemi di metadati, e i requisiti di rappresentatività e interrogabilità dei corpora negli estesi domini che ormai si sono affermati.

3. *I DOI degli interventi al LIV Congresso SLI*

Relazioni su invito

- Gu Yueguo (The Chinese Academy of Social Sciences) *Segmenting and annotating multimodal corpus: inspirations and principles from the traditional Chinese medicine*: https://doi.org/10.48448/nn1m-bj12
- Bas Aarts (University College London) C*orpus-based research in English grammar*: https://doi.org/10.48448/bccc-8309
- Mirko Tavoni (Università di Pisa) *Allestimento, sviluppo e fruizione di DanteSearch, corpus delle opere volgari e latine di Dante con annotazione morfologica e sintattica*: https://doi.org/10.48448/1223-yt30

Plenarie

- Claudio Iacobini[1], Paolo D'Achille[2] ([1]Università di Salerno; [2]Università Roma Tre) *Il corpus MIDIA: concezione, realizzazione, impieghi*: https://doi.org/10.48448/aqxg-c670
- Federica Cominetti[1], Alessandro Panunzi[2], Edoardo Lombardi Vallauri[1], Lorenzo Gregori[2] ([1]Università Roma Tre; [2]Università di

Firenze) *IMPAQTS: un corpus di discorsi politici italiani annotato per gli impliciti linguistici*: https://doi.org/10.48448/zgr4-x188
- Rachele Sprugnoli[1], Matteo Pellegrini[2], Marco Passarotti[1], Flavio Massimo Cecchini[1] ([1]Università Cattolica del Sacro Cuore, Milano; [2]Università di Bergamo) *EvaLatin 1.0: un Corpus per la Valutazione delle Tecnologie del Linguaggio Applicate al Latino*: https://doi.org/10.48448/4esy-d156
- Chiara Alzetta[12], Felice Dell'Orletta[1], Simonetta Montemagni[1], Giulia Venturi[1], ([1]ILC-CNR; [2]Università di Genova), *Esplorazioni di treebank multilingue per studi tipologici*: https://doi.org/10.48448/e8x0-1x63
- Philippe Martin (Université Paris Diderot) *Intonation of telephone conversations in a Customer Care Service*: https://doi.org/10.48448/nrh4-8s92
- Anna-Maria De Cesare (Università di Dresda) *La concezione delle congiunzioni e degli avverbi nella linguistica dei corpora*: https://doi.org/10.48448/a0ne-p210
- Vittorio Ganfi[1], Valentina Piunno[2] ([1]Univ. degli Studi Internazionali di Roma; [2]Univ. Roma Tre), *Diacronia e sincronia delle poliremathiche con struttura preposizionale: un'analisi su corpora*: https://doi.org/10.48448/hvfa-qw07
- Angela Ferrari[1], Letizia Lala[2], Filippo Pecorari1 ([1]Universität Basel, [2]Université de Lausanne), *La punteggiatura italiana attraverso i corpora. Teoria, sincronia e diacronia*: https://doi.org/10.48448/rz5v-cj97
- Iørn Korzen (Copenaghen Business School), *Cosa ci rivelano i corpora sulla complessità testuale dell'italiano?*: https://doi.org/10.48448/raj3-kh31
- Paola Manni[12], Rossella Mosti[3] ([1]Accademia della Crusca; [2]Università di Firenze; [3]OVI-CNR), *Per Dante. Il VD e i corpora dell'italiano antico*: https://doi.org/10.48448/r3q9-3072
- Giulio Vaccaro (ISEM-CNR), *Rappresentatività e bilanciamento in un corpus di italiano antico: appunti sul Corpus OVI*: https://doi.org/10.48448/j1ct-k814
- Maria Francesca Giuliani (OVI-CNR), *Sulla diatopicità del repertorio lessicale degli antichi testi italiani*: https://doi.org/10.48448/dd3v-7975

– Matthias Heinz[1], Lucilla Pizzoli[2] ([1]Universität Salzburg; [2]Università degli Studi Internazionali di Roma), *L'uso dei corpora elettronici per l'OIM*: https://doi.org/10.48448/pz2q-ax51
– Naomi Nagy[1], Chiara Celata[2] ([1]University of Toronto; [2]Università di Urbino Carlo Bo), *A corpus for studying sociolinguistic variation in Italian in migratory settings: homeland and heritage comparisons*: https://doi.org/10.48448/1pyw-3k36
– Andrea Listanti[1], Jacopo Torregrossa[2], Liana Tronci[1] ([1]Università per Stranieri di Siena; [2]Goethe-Universität Frankfurt), *Ordini di acquisizione di strutture VS in italiano L2: uno studio basato sul corpus LIPS*: https://doi.org/10.48448/xb3x-fd73

Demo

– Julia Kaiser (Institut für Deutsche Sprache, Mannheim) *The Research and Teaching Corpus of Spoken German (FOLK) and the Database for Spoken German (DGD)*: https://doi.org/10.48448/64fe-a166
– Nikolay Korotaev, Vera Podlesskaya (Russian State University for the Humanities) *Russian oral discourse through the lens of a multichannel corpus*: https://doi.org/10.48448/9n0c-w283
– Anne Lacheret-Dujour[1], Sylvain Kahane[1], Paola Pietrandrea[2] ([1]Université Paris Nanterre, Laboratoire Modyco; [2]Université de Lille, STL), *Rhapsodie, a prosodic and syntactic treebank for spoken French*: https://doi.org/10.48448/6n57-hk16
– Takehiko Maruyama (Senshu University / National Institute for Japanese Languages and Linguistics), *Design and Analyses of Japanese Speech Corpora*: https://doi.org/10.48448/r0c1-5830
– Heliana Mello[1], Tommaso Raso[1], Sandra Aluisio[2], Tony Berber Sardinha[3], Mark Davies[4], Cláudia Freitas[3], Charlotte Galves[5], Miguel Oliveira[6] ([1]UFMG; [2]USP; [3]PUC, São Paulo; [4]Brigham Young University; [5]Unicamp; [6]UFAL), *Brazilian Portuguese: Spoken, Written and Diachronic Corpora*: https://doi.org/10.48448/6gdg-0090
– Salvador Pons[1], Margarita Borreguero[2] ([1]Universidad de Valencia, [2]Universidad Complutense de Madrid) *Corpus Val.Es.Co. 3.0*: https://doi.org/10.48448/7v8q-z950

- Victoria Vázquez Rozas (Universidade de Santiago de Compostela), *ESLORA: un corpus de español hablado en Galicia*: https://doi.org/10.48448/4zg4-gv06
- Marco Biffi, Francesca Cialdini (Università di Firenze), *Banche dati per il trasmesso: il LIR e il LIT*: https://doi.org/10.48448/7c2w-t760
- Giorgina Cantalini (Civica Scuola di Teatro Paolo Grassi), *Corpus Multimodale Annotato per lo studio della gestualità co-verbale nel parlato parlato e nel parlato recitato*: https://doi.org/10.48448/3k4n-hg49
- Emanuela Cresti, Lorenzo Gregori, Massimo Moneglia, Carlota Nicolas, Alessandro Panunzi (Università di Firenze), *Corpus dell'Italiano Parlato LABLITA; Corpora comparabili delle lingue romanze parlate (C-ORAL-ROM); Corpora didattici dello spagnolo (CORDIAL); Data Base interlinguistico dell'articolazione dell'informazione (DB-IPIC)*: https://doi.org/10.48448/1jrz-3810
- Francesca M. Dovetto[1], Alessia Guida[1], Anna Chiara Pagliaro[1], Raffaele Guarasci[2], Simona Trillocco[1], Sundra Sorrentino[1], Lucia Raggio[1] ([1]Università Federico II, Napoli, [2]ICAR CNR): *Corpora di italiano parlato patologico dell'età adulta e senile: CIPPS, CIPP-ma, CIPP-mci*: https://doi.org/10.48448/39js-b573
- Eugenio Goria[1], Caterina Mauri[2], Massimo Cerruti[1], Silvia Ballarè[1] ([1]Università di Torino; [2]Università di Bologna), *Il corpus KIParla*: https://doi.org/10.48448/jkp3-rk48
- Cristina Onesti, Carla Marello, Manuel Barbera, Elisa Corino (Università di Torino) *Corpora.unito; I corpora VALICO e VINCA*: https://doi.org/10.48448/hkms-vq47 e https://doi.org/10.48448/drhb-1918
- Fabio Tamburini (Università di Bologna), *I corpora del FICLI*: https://doi.org/10.48448/kh2s-3623

References

Barbera, Manuel. 2013. *Linguistica dei corpora e linguistica dei corpora italiana. Una introduzione*. Milano: Quasar.

Biber, Douglas & Johansson, Stig & Leech, Geoffrey & Conrad, Susan & Fingegan, Edward. 2000. *Longman Grammar of Spoken and Written English*. London: Longman.

Blanche-Benveniste, Claire & Bilger, Mireille & Rouget, Christine & Van den Eynde, Karel. 1990. *Le français parlé – études grammaticales.* Paris: Editions du Centre National de la Recherche Scientifique.

Bunge, Mario Augusto. 1984. Philosophical problems in linguistics. *Erkenntnis* 21. 107-173.

Bunge, Mario Augusto. 2000. Systemism: the alternative to individualism and holism. *Journal of Socio-Economics* 29. 147–157

Chafe, Wallace. 1992. The importance of corpus linguistics to understanding the nature of language. In Svartvik, Jan (ed.) *Directions in Corpus Linguistics: Proceedings of Nobel Symposium* 82, 79-97. Berlin and New York: Mouton de Gruyter.

Cresti, Emanuela. 2000. *Corpus di Italiano Parlato*. Firenze: Accademia della Crusca.

Cresti, Emanuela & Moneglia, Massimo (eds). 2005. *C-ORAL-ROM. Integrated Reference Corpora for Spoken Romance Languages*. Amsterdam: John Benjamins.

Cresti, Emanuela & Panunzi, Alessandro. 2013. *Introduzione ai corpora dell'italiano*. Bologna: Il Mulino.

De Mauro, Tullio.2000. *GRADIT: Grande dizionario italiano dell'uso*. Torino: UTET.

Dovetto, Francesca M. & Gemelli, Monica. 2013. *Il parlar matto. Schizofrenia tra fenomenologia e linguistica. Il corpus CIPPS*, Prefazione di Federico Albano Leoni, Seconda edizione rivista e integrata con DVD-ROM [audioregistrazioni e trascrizioni], Roma: Aracne. [2012 prima ed.]

Ferrari, Angela & Cignetti, Luca & De Cesare, Anna-Maria & Lala, Letizia & Mandelli, Magda & Ricci, Claudia & Roggia, Carlo Enrico. 2008. *L'interfaccia lingua-testo. Natura e funzioni dell'articolazione informativa dell'enunciato*. Alessandria: Edizioni dell'Orso.

Freddi, Maria 2014 *Linguistica dei corpora*. Roma: Carocci.

Koiso, Hanae & Den, Yasuharu & Iseki, Yuriko & Kashino, Wakako & Kawabata, Yoshiko & Nishikawa, Ken'ya & Tanaka, Yayoi & Usuda, Yasuyuki. 2018. Construction of the Corpus of Everyday Japanese Conversation: An Interim Report. In *Proceedings of LREC2018*, 4259-4264.

Leech, Geoffrey. 1992. Corpora and theories of linguistic performance. In Svartvlk, Jan (ed.), *Directions in corpus linguistics. Proceedings of Nobel Symposium 82, Stockholm, 4-8 August 1991,* 105-22. Mouton de Gruyter, Berlin/New York.

Lenci, Alessandro & Montemagni, Simonetta & Pirelli Vito. 2016 *Testo e computer. Elementi di linguistica computazionale*. Roma: Carocci;

Maraschio, Nicoletta & Poggi-Salani, Teresa. 2003. *Italia linguistica anno mille. Italia linguistica anno duemila*. Atti del XXXIV Congresso Internazionale di Studi, Firenze 19-21 ottobre 2000, Roma: Bulzoni

Martin, Philippe. 2015. *The structure of Spoken Language*. Cambridge: CUP.

Nencioni, Giovanni. 1976. Parlato-parlato, parlato-scritto, parlato-recitato. *Strumenti critici LX*. 1-56.

Riccio, Anna 2016. *Gli strumenti per la ricerca linguistica. Corpora, dizionari e database*, Roma Carocci;

Sinclair, John. 2004. Carter, Ronald (ed.), *Trust the Text: Language, Corpus and Discourse*. London: Routledge

ELAN: <https://archive.mpi.nl/tla/elan>

ENGLICIOUS: <http://www.englicious.org/>

OVI: <http://www.ovi.cnr.it/>

PRAAT: <https://www.fon.hum.uva.nl/praat/>

Sketch Engine & NoSketch Engine: <https://www.sketchengine.eu/nosketch-engine/>

Survey of Enghish Usage: <https://www.ucl.ac.uk/english-usage/>

Tree Tagger: <https://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/>

Underline: <https://underline.io/>

WINPITCH: <https://www.winpitch.com/>