6 | nuova serie

S L I

Società di Linguistica Italiana

Corpora e Studi Linguistici

Atti del LIV Congresso Internazionale di Studi
della Società di Linguistica Italiana
(Online, 8-10 settembre 2021)

Corpora and Linguistic Studies

Proceedings of the LIV International Congress
of the Società di Linguistica Italiana
(Online, 8-10 September, 2021)

a cura di
Emanuela Cresti - Massimo Moneglia

2022

S L I | Società di Linguistica Italiana

Corpora e Studi Linguistici

Atti del LIV Congresso Internazionale di Studi
della Società di Linguistica Italiana
(Online, 8-10 settembre 2021)

Corpora and Linguistic Studies

Proceedings of the LIV International Congress
of the Società di Linguistica Italiana
(Online, 8-10 September, 2021)

a cura di
Emanuela Cresti - Massimo Moneglia

Milano 2022

S L I | Società di Linguistica Italiana

La Società di Linguistica Italiana (SLI), costituitasi a Roma nel 1967, ha lo scopo di promuovere studi e ricerche nel campo della linguistica, attraverso la creazione di una comunità di studiosi nel cui ambito trovi pieno riconoscimento e appoggio ogni prospettiva di ricerca linguistica teorica e applicata. La SLI tiene ogni anno un congresso internazionale di studi, e pubblica in volume alcuni dei contributi presentati al congresso. I manoscritti vengono valutati tramite un processo di revisione tra pari. Dal 2018 i volumi sono pubblicati con accesso libero a tutti gli interessati.

Edizione realizzata da
Officinaventuno
Via F.lli Bazzaro, 18
20128 Milano - Italy
email: info@officinaventuno.com
sito: www.officinaventuno.com

ISBN edizione cartacea: 978-88-97657-55-2
ISBN edizione digitale: 978-88-97657-56-9

Indice

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/>

PARTE PRIMA

I CORPORA

DOI: 10.17469/O2106SLI000002

GU YUEGUO

Reflections on the Foundation of Corpus Construction: An Argument for Experience-based Conceptualization

This paper presents reflections on the foundation of corpus linguistics, specifically around two basic issues: (1) Does corpus linguistics, bearing the label *linguistics*, contribute to our understanding of language in general or only specific languages? (2) Does corpus linguistics contribute to our understanding of man, or the mind of man, or individuals as speakers? Four positions are critically reviewed, covering three orientations, viz. product-oriented, process-oriented and experience-oriented. It is argued that experience-oriented conceptualization is a viable direction for future development. The tripartite division of labour in corpus construction – conceptual corpus linguistics, ideal corpus linguistics and practical corpus linguistics – is proposed and demonstrated. It points out that Bunge's systemism and ontology should be adopted as the foundation of corpus construction.

Keywords: experience-oriented conceptualization, corpus ontology, systems thinking.

1. *Preliminary Remarks*

Reflections on my 35-year compilation of the Spoken Chinese Corpus of Situated Discourse (the SCCSD, www.multimodalgu.com) have made me become suspicious of some basic issues about corpus linguistics in general:

1. Does corpus linguistics, bearing the label *linguistics*, contribute to our understanding of language in general or only specific languages?
2. Does corpus linguistics contribute to our understanding of man, or the mind of man, or individuals as speakers?

All practitioners of corpus compilation know through years of sleepless nights that these issues are central to corpus building of any kind, big or small. Practitioners also are aware that they have also been pon-

dered upon by some corpus pioneers. This paper first presents a brief critical review of the pertinent literature. The bulk of the paper will be dedicated to an argument for a holistic and experiential approach to the theorization of corpus linguistics, which will throw, hopefully, some light on the two fundamental issues.

2. *Corpus construction and its bearing on language/languages*

Historically, corpus construction witnesses an event that has exerted ever-lasting consequences, namely the use of computer. So significant is it that Leech (1992), based on it, draws a distinction between computer corpus linguistics and non-computer corpus linguistics. His distinction goes beyond sheer recognition of technology. He writes (1992:106): "I wish to argue that computer corpus linguistics (henceforth CCL) defines not just a newly emerging methodology for studying language, but *a new research enterprise*, and in fact *a new philosophical approach to the subject*." (italics added). In what sense does CCL constitute a new research enterprise? Leech points out:

> "On the face of it, a computer corpus is an unexciting phenomenon: a helluva lot of text, stored on a computer. But the computer's ability to search, retrieve, sort, and calculate the contents of vast corpora of text, and to do all these things at an immense speed, gives us the ability to *comprehend*, and to *account for*, the contents of such corpora in a way which was not dreamed of in the pre-computational era of corpus linguistics." (Leech, 1992: 106; italics original)

Speedy comprehension and massive data-based accounting-for still make a vivid picture of the present-day corpus linguistics research. These technical advantages aside, what is Leech's "a new philosophical approach to the subject"? His view is best seen through a comparison he makes with "other approaches in linguistics", primarily "the Chomskyan paradigm". It is summarized in four focuses as follows:

1. Focus on linguistic performance, rather than competence
2. Focus on linguistic description, rather than linguistic universals
3. Focus on quantitative, as well as qualitative models of language
4. Focus on a more empiricist, rather than rationalist view of scientific inquiry.

Leech's four-focus characterization of "a new philosophical approach" does indeed indicate something new only if we view it comparatively.

Each focus by itself, however, can hardly be seen as something new, for it has been argued for and pursued elsewhere in linguistics long before CCL. There are two features standing prominently in Leech's argumentation. One is that it is bi-polemic in the sense that it is dichotomous between two opposing opposites: performance vs. competence; specific vs. universalistic; empiricism vs. rationalism. The other is reductionistic in the sense that the speaker, the ultimate owner, the agent, of any human language, is reduced to non-existence, since it plays no role in the theorization of what language is. In other words, the three pairs of dichotomies about language just mentioned is language without speaker. Language is thus treated as if it were an object independent from the owner/agent.

It must not be construed here that Leech has failed to see the fact that behind every human language, every piece of text there is an ultimate producer/speaker/owner. This fact is taken for granted and plays little role as a key concept in theory-building.

The weakness of Leech's conceptualization of CCL is somehow amended by Chafe's view about corpus linguistics. Chafe reports the outcome of his pondering "more deeply the place of corpora within linguistics. Since I believe they are an absolutely crucial part of the linguistic enterprise..." he still sounds quite apologetic: "I would like to take advantage of this symposium to try to articulate some ideas about how corpora further the *ultimate goal of understanding the nature of language*. I hope I will be pardoned for starting in a philosophical vein." (Chafe, 1992: 79; italics added). As indicated above, Leech's "philosophic" venture is through bi-polemic dichotomy and reductionism, Chafe's in comparison is certainly deeper, for he wants to explore, via corpus linguistics, the *nature of language*. He argues thus:

> "we will never get much farther than we have gotten up to now unless we accept and integrate into our work the realization that *language cannot be separated from the mind as a whole*, that understanding language and understanding the mind are at bottom the same endeavor". (Chafe, 1992: 80; italics added)

It is crucial to note that Chafe does not adhere to the modularity theory of mind, as most notably advocated by Chomsky (1986) and Fodor (1983). He holds an opposite view "that language is an inseparable part of total mental activity." (Chafe, 1992: 81). He feels "joy whenever they discover a way in which some linguistic phenomenon can

be characterized as motivated and functional – explainable within a larger, coherent picture of the mind." (Chafe, 1992: 81). So language, as seen by Chafe, is free from the tyranny of competence-vs-performance dichotomy. This view of language runs consistent and has been continuously pursued in Chafe's long academic career (e.g., Chafe, 1994, 2018).

Chafe's investigation commences from the mind's eye, and he wants to know how the mind processes "speaking", which "is natural to the human organism... humans are 'wired up' to speak and listen". In this endeavor, "corpora have led to the discovery of two important constraints on the way the mind processes information during the production ... of language." (Chafe, 1992: 88). The two constraints are: "the light subject constraint", and "the one new idea constraint". Simplistically paraphrased, the first constraint refers to the phenomenon that the sentence subject tends to be light in information weightiness, while the second to the hypothesis that each intonation unit is limited to no more than one idea that is new.

Chafe hence asks: "What, then, is a 'corpus linguist'?"

> "I would like to think that it is a linguist who tries to understand language, and behind language *the mind*, by carefully observing extensive natural samples of it and then, with insight and imagination, constructing plausible understandings that encompass and explain those observations". (Chafe, 1992: 96; italics added)

This paper, meanwhile applauding Chafe's depiction of the tasks of corpus linguists, feels somewhat disappointed at his step short of embracing the whole-person speaker as the basic naturally given unit on which the conceptualization of corpus linguistics should be founded. We shall turn to this theme in section 4 below.

3. *Sinclair's view of corpus linguistics*

Our brief review, no matter how brief, cannot be complete without a look at Sinclair's position. To the best of my knowledge, Sinclair (2004) is the most outspoken corpus linguist who explicitly calls for giving priority to dialogical model of language. Integrating the written and spoken language has remained a recurrent theme in many of Sinclair's works. His effort is best appreciated when it is seen in a larger context. Linell (2005) makes a list of 101 points proving the

written language bias in linguistics, the bias actually felt by quite a few corpus linguists as well. Biber et al. (2000: 1038) observe that "there is a compelling interest in using the resources of the LSWE[1] Corpus of spoken language transcriptions to study what is characteristic of the grammar of conversation." But immediately after they admit that the "fact that this Corpus material consists of transcriptions — speech rendered in written form — means that even here, the reliance on the written form of the language cannot be escape..." In spite of this written bias, they still feel that

> "the existence of such a large body of transcribed speech makes it feasible to seek an answer to the following question, which has recently excited considerable interest: *is there a distinctive grammar of spoken language, operating by laws different from those of the written language? If so, what are these laws, and what are the functional or other principles underlying them*?" (italics added)

Their answer is: "the same 'grammar of English' can be applied to both the spoken and the written language." What is extremely interesting to note, however, is that the authors of the chapter have to admit constantly the difficulty in putting the jacket of written language grammar onto rebellious conversation. Take the notion of *sentence* for example. They observe:

> "Whereas the **sentence** has been treated, traditionally and in modern theory, as the fundamental structural unit of grammar, *such a unit does not realistically exist in conversational language*". (Biber et al., 2000: 1039; bold original; italics mine)

Carter and McCarthy (2006: 9) write:

> "Most books on the grammar of English have had a bias towards the written language. For many centuries dictionaries and grammars of the English language have taken the written language as a benchmark for what is proper and standard in the language, incorporating written, often literary, examples to illustrate the best usage. Accordingly, the spoken language has been downgraded and has come to be regarded as relatively inferior to written manifestations".

Carter and McCarthy seem to spare no efforts in counterbalancing the written language bias by inventing, quite readily without feeling

[1] Longman Spoken and Written English.

awkward, fresh metalanguage to capture unique features found in spoken discourse.

So much for the larger context. Let us return to Sinclair. If Biber et al try to integrate the spoken into the written, Sinclair attempts to do the opposite, i.e., integrating the written into the spoken. In his paper "The internalization of dialogue" (2004 [1999]), he puts forward the hypothesis that "much of the complexity of sentence grammar can be explained as the internalization of features of spoken interaction. This hypothesis is quite plausible, considering the fact that the written form in any literate culture is a much later development than the spoken one. Sinclair observes that "it must be reiterated that by the time it became possible to develop a written form of a language, the structures were already capable of great complexity." (ibid., p. 104). This echoes Chafe's view of speaking as human species' wired-up property touched upon in section 2 above.

What is significant about Sinclair's position is that he goes beyond embracing spoken samples in corpus building, and proposes to conceptualize corpus linguistics on the basis of speaking instead of writing. Two terms, viz. *dialogic language* vs. *monologic language*,[2] are proposed to help conceptualize the transition from the spoken to the written. Dialogic language is language in an interactive mode, whereas monologic language does not require elaborate contributions from other participants. The internalization involves reduction and compression of information, with gains or losses of independence or interdependence. Take "a move" in dialogic language for example. It is independent in its dialogic environment, but when it is internalized in monologic language, it becomes a sentence, and "loses its property of constituting an independent utterance". The sentence, however, "gains the facility of posture as a property of main clauses". (Sinclair, 2004: 114).

The notion of internalization in fact originates from his earlier paper "Planes of discourse" (2004 [1982]), where Sinclair writes:

[2] Sinclair's use of the term *language* in the distinction needs to be construed with care in this paper. In oral-aural cultures, there exists no monologic language. In such cases the term *language* overlaps with the term *natural language*, i.e. the sense being used in this paper. In literate cultures, on the other hand, there exist both dialogic and monologic languages. Here the term *language* does not overlap with *natural language*.

> "Language in use has two aspects: at one and the same time it is both a continuous negotiation between participants, and a developing record of experience. The negotiation aspect highlights interaction and will be called the *interactive plane* of discourse. ...
> ...
> The other aspect of language in use is the developing record of experience. On a small scale, in a conversation, say, or reading a letter, it can be seen as gradual sharing of relevant experience by recalling previous words and phrases and reworking them in the new contexts provided by a movement on the interactive plane.
> ...
> The stage-by-stage tally of the record of experience will be called the *autonomous plane* of discourse, because it is concerned with language only and not with the means by which language is related to the world outside." (Sinclair, 2004: 52-53; italics original)

The autonomous plane of discourse, looked at on a larger scale, as pointed out by Sinclair, is "a continuous internalization of experience, from the world outside to the inner space of language. The process is both individual and collective, and, where written down, forms the most explicit record we have of human evolution." (Sinclair, 2004: 53).

In a word, Sinclair adopts the dialogical model of language in use as a reference framework against which the written sentence grammar is critically examined. He exceeds Chafe in two aspects. One is that, unlike Chafe, he does not take the current status of corpus linguistics as given, and attempts to make it anew through reconceptualization. The other is that Sinclair's notion of speaker is much fuller than Chafe's. As pointed above, Chafe incorporates human language and mind, whereas Sinclair incorporates human language use into (1) "continuous negotiation between participants", and (2) "a developing record of experience". Surely the concept of experiencing participant presupposes a thinking mind, but the reverse inference is invalid. The experiencing participant will be referred to as a whole person below. To which we turn.

4. *The participant's total saturated experience: A step further from Sinclair*

The experiencing participant possesses not only a thinking mind, but also multiple sensory modalities interacting with the outside world.

Language use in the real-life everyday world is naturally multimodal and experientially real. By "naturally multimodal" is meant that real-life situated discourse involves what Goffman calls "bodily activity" on the speaker/performer's side, and "naked senses" on the addressee/receiver's side. The speaker's *current bodily activity* makes ***embodied messages*** (Goffman's terminology).

> "When one speaks of experiencing someone else with one's *naked senses*, one usually implies the reception of embodied messages. This linkage of naked senses on one side and embodied transmission on the other provides one of the crucial communication conditions of face-to-face interaction." (Goffman, 1963: 15; italics mine).

As emphasized by Goffman, ordinarily in using the naked senses to receive embodied messages from others, one also makes oneself available as a source of embodied messages for others. This is why we emphasized above that Chafe's notion of mind is not the same with Sinclair's notion of participant.

The embodied messages, produced and received/interpreted by both sides, consciously and sub-consciously, make moment-by-moment experienced reality. Gu (2009) proposes the adoption of the term *total saturated experience* (TSE for short) to refer to Goffman's face-to-face interaction with naked senses and embodied messages, and of the term *total saturated signification* (TSS for short) to talk about the total meanings constructed out of the total saturated experience by the acting co-present individuals.

Gu in series studies (2006a, 2009, 2013, 2016) attempts to make Sinclair's suggestion about the experiencing participant a *central concept* in corpus linguistics. Sinclair's "a developing record of experience" is reconceptualized primarily in terms of three inter-related concepts, *experiencer-experiencing-experience* (i.e., EEE model for short), and secondarily the "developing recording" is dynamically modeled through the formula "the (dimensional) self {...{...}...}". Briefly, the participant fills this formula with multiple layers of data corresponding to the multiple selves through multiple sensory modalities of interacting with the outside world (see further discussion in 5.3 below).

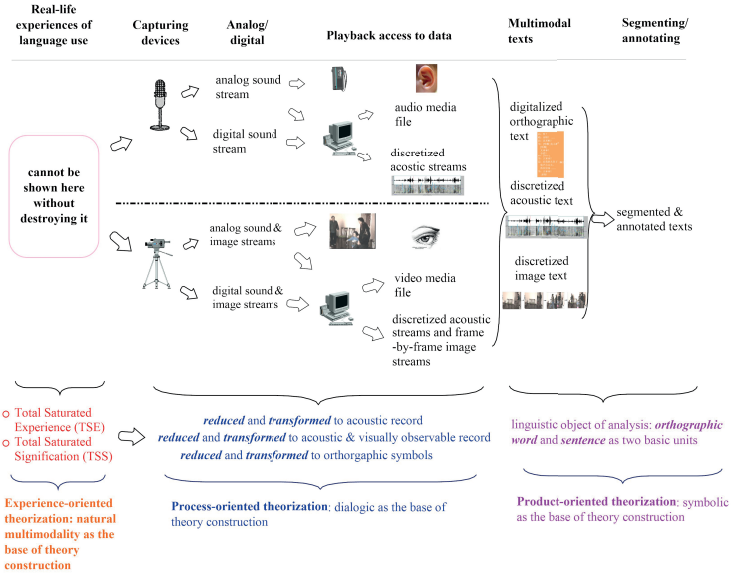
Gu (2009: 435) contrasts three perspectives in theorizing the study of language use: (1) product-oriented, (2) process-oriented, and (3) real-life experience-oriented. The first has been the most predominant one, whereas the second has been voiced as a potential alternative as

suggested by Sinclair. The third, in contrast, is little heard of. The three contrastive perspectives can be illustrated by the ensuing scenario.

> "Think about what happens to language use when the audio/video-recording device (or any other data-capturing devices) is switched on? It transforms real-life language-using experiences onto cassette tapes or hard disks. What is recorded on the tape or hard disk becomes data — real-life language use data. A corpus linguist then has an option of accessing the data: (1) by playing the medium back to listen to via the naked ear or watch it via the naked eye; and/or (2) transcribing the data into orthographic words or symbols before analyzing them; and/or (3) segmenting and annotating the data using tools such as PRAAT, ELAN, etc."

The whole process is graphically represented in Figure 1.

Figure 1 - *From real-life experience to text annotation*

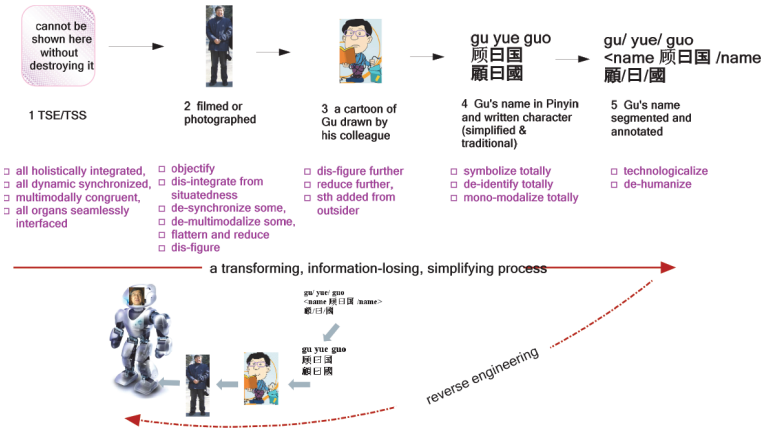


The product-oriented theorization shown on the bottom of the right side draws data from orthographic texts; the process-oriented theorization in the middle relies on dialogic data; the experience-oriented theorization attempts to develop a theory by modeling the natural multimodal interaction in the real-life world.

As shown in the figure, on the leftist side, it is the occurrence of real-life language use. It is a moment-by-moment creation of life experience, of which language use is a crucial part. It constitutes a total saturated experience and hence total saturated signification. This real-life experience cannot be captured by any device without destroying it. This echoes the fundamental tenet of Taoism: "Conceived of as having no name, it is the Originator of heaven and earth; conceived of as having a name, it is the mother of all things."[3]

The move from the product-oriented to the process-oriented, then to the experience-oriented and finally to the TSE-TSS can be seen as a reverse engineering process. This view assumes that language use experience lived by participants constitutes the ultimate goal of understanding (in the sense as defined by Chafe) to be achieved by corpus linguistics. Those corpus linguists adopting this view will endorse Gu's argument that the product-oriented theorization is based on *very much impoverished, if not distorted data*. This may be illustrated by the analysis of the author himself from the real-life person, alive and kicking, to his photograph, to his cartoon, and finally to his orthographic name (see Figure 2).

Figure 2 - *From real life to orthographic text and vice versa*



The bullet items under each stage indicate its data properties, and the transformations and distortions are clearly seen by comparing the

[3] Legge's translation of Laozi's Daodejing (Legge, 2008 [1891]).

data properties between the stages. At Stage 1 represented by TSE-TSS Gu is an all-round living system; at Stage 2 he is objectified, and flattened into a lifeless 2-dimensional figure; at Stage 3 he is further distorted and disfigured; at Stage 4 he is totally symbolized and de-identified – which can be testified if someone, having never seen him, is asked to pick Gu up from a crowd only by a name slip; at Stage 5 Gu's orthographic personal name is segmented and annotated for computer to do personal name recognition.

Now a corpus linguist is given a task of making a hominoid robot of the author capable of speaking Chinese, if given a corpus of orthographic texts. This is reverse engineering from Stage 5 back to Stage 1. Note that such reverse engineering tasks are not imaginative or mind-experiment: Paleoanthropologists, paleographers, and historiographers face challenges like this in their daily routines.

The illustration shows a self-defeating element intrinsically built in as a result of its linguistic stance and research methodology corpus linguists assume. By the time they finish crunching words, sentences, patterns, etc., and declare their findings about language use, the picture drawn is likely to be skewed in the way Indian blind men report their findings about the elephant!

Those corpus linguists who do not endorse Chafe's ultimate understanding view may counterargue that, since it is impossible to attain full data sets mapping the TSE-TSS, it becomes pointless to pursue experience-oriented theorization. This counterargument, if adhered to, will hinder advancement of corpus linguistics enterprise. We shall show some benefits of such theorization and discuss pertinent ontological and methodological issues.

5. *Tripartite division of labour in theorizing corpus linguistics*

5.1 Thinking in general systems theory

This paper proposes a solution to fending off the self-defeating element mentioned above. It proposes a tripartite division of labour of theorizing corpus linguistics, namely *conceptual corpus linguistics*, *ideal corpus linguistics*, and *practical corpus linguistics*. Conceptual corpus linguistics adopts the stance of looking at language use as it being experienced and lived by participants, and extracts know-how knowledge from it by way of constructing experiential constructs. Ideal corpus

linguistics, on the other hand, adopts a stance of looking at language use as an object to be modeled, and makes data models on the basis of the experiential constructs. Finally practical corpus linguistics looks at language use as texts, and its central task is to handle practicalities of text processing, meanwhile adopting experiential constructs and data models as its blueprints.

The hallmark of such conceptualization of corpus linguistics is *its general systems theory approach from real-life experience to text processing, and to data reverse engineering*. It does not stop at looking at language use as phenomenological reality. It goes further than that. For technically speaking, language use as it being experienced is an open-ended, continuous, and emergent analogue data type. It has to be modeled, sampled, quantized, and discretized to produce machine-friendly data type for efficient and effective processing. Experiential reconstruction and data modeling as research methodologies are no less important than orthographic transcription and word-crunching.

5.2 Bunge's ontological theory: A synopsis

Before we proceed, a synopsis of Bunge's theories of ontology (or metaphysics) and systemism will be helpful in case that they have not caught interest of some corpus linguists. Bunge differentiates "exact philosophy" (Bunge, 2016: Chapter 9) from inexact philosophies. His exact philosophy refers to philosophizing in such a way that all the basic concepts of philosophy, e.g., substance, thing, process, property, state, possibility, are defined in the metalanguage of basic set theory and pertinent mathematics. In Bunge's theorization, the terms ontology and metaphysics are interchangeable. He adopts "a naturalistic (or materialistic) world view", which is taken to "be the ontology of factual science and technology" – keep it in mind that corpus linguistics is technology-driven.

> "On this view the world is composed exclusively of concrete changing things: everything else is the invention of particular concrete things such as ourselves. Clearly, on this view language cannot exist in the same way as stars and people exist, i.e. in and by themselves. On this view *what are real are not languages but people, or other rational beings, engaged in producing, conveying or understanding linguistic expressions.* Asking whether language exists is like asking

> whether life or mind exist. The answer is an unqualified 'No'. There are no autonomous languages any more than there is life or mind by itself. There are instead minding animals and, in particular, animals capable of speaking and understanding speech. This, *the production and understanding of speech, we take to be the primary linguistic fact*. Everything else about language is construct – starting with language itself. Shorter: speech is real, language is not." (Bunge, 1984: 112-3; italics added)

It is essential to bear in mind that it is the minding animals capable of speaking and understanding that are real and concrete. Language is a construct out of speech. Bunge draws a distinction between "the philosophy of language", a part of ontology and epistemology, and "the philosophy of linguistics", a part of the philosophy of science (Bunge, 1984: 111). The former is concerned with such basic questions as "What is language?"

> "The basic question, 'What is language?' is an ontological one in the same category as 'What is life?'
>
> ...
>
> The nature of the question is likely to be better understood in attempting to answer the related question 'How does language exist?'. According to idealism language exists by itself, either as a sort of Platonic idea pre-existing people and hovering above them, or as a human creation though one that is immaterial. Needless to say, there is no empirical evidence for either variety of idealism." (Bunge, 1984: 112)

Bunge advocates materialism that holds that language does not exist by itself, contrary to idealism's position. The postulate that language does not exist by itself becomes self-evident only with reflection on the fact that there is no language without speakers. If language exists, its existence is secondary and derivative. Ironically, linguists accept this basic fact on the one hand, they ignore it while theorizing language on the other. It is a wide spread practice that language is regarded as an autonomous, independent system existing by itself. In other words, while linguists adopt a materialistic philosophy of language, they apply an idealistic philosophy of linguistics in practice. There are some serious consequences as a result of the discrepancy between the materialistic ontology and the idealistic practice. As Dillinger points out:

> "... linguists' theory is not developed as explicitly as their practice; the study of language and languages is fragmented, each subspecialty proceeds quite autonomously from the others; theoretical writing and textbooks present the field as a potpourri of activities without any explicit relations between them; and mutually exclusive "approaches" proliferate, each championing the study of one or a few fragments of the whole." (Dillinger, 1990: 10)

Dillinger's critique can readily be testified by the reading experiences students of linguistics have had while reading literature covering branches of linguistics, e.g. phonetics, phonology, formal syntax, formal semantics, pragmatics, discourse analysis, you name it. The reading of each, put in a Chinese proverb, constitutes an experience as difficult as crossing a mountain between them.

Bunge offers a revealing comment about "philosophical problems in linguistics" as follows:

> "... whoever regards language as an ideal object cuts the ties of pure linguistics with the other five branches of linguistics. Worse: *he isolates linguistics from the system of factual sciences*, all of which study concrete (material) things. And insulation is the mark of pseudoscience." (Bunge, 1984: 112; italics added)

Two pieces of background information are helpful in order to appreciate fully Bunge's critique: (1) Bunge's application of general systems theory to linguistics, and (2) his application of factual sciences to linguistics.

Bunge's favourite term for general systems theory is systemism, which is in contrast with individualism and holism. It refers to "a whole systemic worldview". It is centered in the following postulates:

1. Everything, whether concrete or abstract, is a system or an actual or potential component of a system;
2. systems have systemic (emergent) features that their components lack, whence
3. all problems should be approached in a systemic rather than in a sectoral fashion;
4. all ideas should be put together into systems (theories); and
5. the testing of anything, whether idea or artifact, assumes the validity of other items, which are taken as benchmarks, at least for the time being.

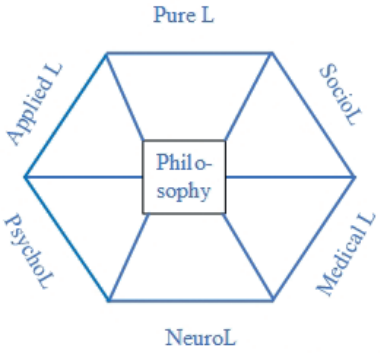
(Bunge, 2000: 149)

A *system,* concisely put, is "an organized whole in which parts are related together, which generates emergent properties and has some purpose." (Skyttner, 2005: 58). Systems are usually classified into four types: (1) concrete, (2) conceptual, (3) abstract, and (4) unperceivable. Alternative classifications are often made (1) between natural and man-made systems; (2) between living and non-living systems. The concrete system is the most common and also called physical system.

Everything can be modeled as a system, so language is a system, and linguistics is also a system. System-modeling like this, surly advantageous, has a defect of making people overlook the fact that *language system and linguistics system have very different referents, the former of which is the set of real people, whereas the latter is the set of man-made theories*. The factual sciences, on the other hand, study concrete/material things, and linguistics, if it wants to protect itself from falling into pseudoscience, must study concrete/material objects, e.g., speeches as recorded and collected in corpus. Logically, linguistics aiming at idealized and abstract phenomenon violates the postulates of factual sciences, hence it is pseudoscientific at best.

Bunge has drawn a "linguistic hexagon" visualizing the philosophy of linguistics – not to be confused with the philosophy of language – as shown in Figure 3.

Figure 3 - *Bunge's linguistic hexagon (Bunge, 1984: 111)*



Note: The linguistic hexagon or system of disciplines that study language. Here 'L' stands for 'linguistics'. Pure L is construed as the study grammars, which – since Chomsky (1965) – include syntax, semantics, and phonology.

Bunge's linguistic hexagon must not be construed strictly literally, that is, it can easily be expanded into a polygon like heptagon, octagon, and so on. The essential message is that "the different conceptions of language are related not only to the diversity of linguistic schools – each of them attached to its own philosophy – but also to the current fragmentation of the study of language into half a dozen different disciplines. These disciplines, that are only tenuously connected to one another..." (Bunge, 1984: 109).

5.3 Conceptual corpus linguistics: A whole person model

With Bunge's theories of ontology and systemism as our theoretical backbones, we are ready to give some meat to the tripartite division of labour introduced in 5.1 above. First, the tripartite division presupposes oneness of the overall task, viz. conceptualizing corpus linguistics as a systemic system in Bunge's sense of the term. Only in this way do we hope to avoid the fragmentation and mutual incomprehensibility found in the current branches of linguistics.

Conceptual corpus linguistics abides by the postulates of factual sciences. Accordingly, it is concerned with concrete/material objects with real-life existence, viz. individual speakers alive and kicking. The individual speaker is seen as a system which is part of a bigger system, viz. a society where s/he makes a living. Like any other systems, the individual speaker system is understood basically through four stages of human knowledge development: (1) intuition, (2) fact-finding, (3) analysis and (4) synthesis (Skyttner, 2001: 30). Corpus compilers are by nature individual speakers, hence they have naturally grown intuitions about themselves, about what to speak, and about many other things besides. Fact-finding is delineated by one's ontological view about what constitutes a fact. "Wittgenstein started his famous Tractatus of 1921 asserting that 'the world is the totality of facts, not of things'"(requoted from Bunge, 2016: 249). Bunge counterargues that

> "in the factual sciences 'fact' denotes either a state of a thing or a sequence of states of a thing: there are *no facts in themselves, without material things*. For example, there would be no car collisions without cars, or changes of government without rulers and their subjects. Remember Aristotle's criticism of Plato's idea of movement in itself rather than moving things." (Bunge, 2016: 249; italics added)

Following Bunge, there is no such linguistic fact in itself: *There are only linguistic facts as states of a speaker*. This statement is taken as our fundamental commitment to the ontological foundation of conceptual corpus linguistics. What is recorded when a corpus compiler switches on a tape recorder? Is he recording a linguistic fact? *It is not unless it is taken as a state or a sequence of states of the speaker being recorded*. We shall come to this later in section 8 below.

Let us return to the third and fourth stages above: analysis and synthesis. Analysis examines the components of a system, while synthesis expounds the synergetic properties of a system as a whole. In other words, analysis decomposes the system and synthesis elucidates how the system functions as a whole in a larger system.

Both analysis and synthesis are aided by a variety of tools. The tools we employ for conceptual corpus linguistics is systems theory, and in particular Bunge's systematerialist ontology expounded in Bunge's *Treatise on Basic Philosophy* (Volumes 3 and 4, 1977, 1979). Bunge's systematerialist ontology is a formidable general theory dealing with the "furniture of the world" and "a world of systems". Accurately elucidated key concepts include, among others, substance, property, state, process, thing, system, emergence, submergence, space, time, causality, randomness, space, time, chemism, life, evolution, mind, society, social structure, participation, marginality, social cohesion, and history. The metalanguage employed for this giant task is

> "such elementary formal tools as set and function. In addition, the exposition follows the axiomatic format: primitive (undefined) concepts and defined ones, axioms (or postulates), theorems, and comments. However, the motivation and justification of my principles (axioms) originated in the sciences." (Bunge, 2016, pp. 260-261)

What I have benefited most is his way of dealing with a very complex system in the simplest possible but accurate way. His systemism is concisely captured by himself in his *Memoir* as follows:

> "I have proposed ... that any system σ may be schematically modeled by the ordered quadruple composition-environment-structure mechanism, or
>
> $$\mu(\sigma) = \langle C(\sigma), E(\sigma), S(\sigma), M(\sigma) \rangle,$$
>
> where

C (σ) = Set of constituents of σ at the given level of analysis;

E (σ) = Immediate environment of σ;

S (σ) = Set of bonding and non-bonding relations among the system's constituents;

M (σ) = Set of processes that keep σ going."

The quadruple model is illustrated thus:

> "For example, the composition of an organism at the cellular level is the set of its cells, whereas at the organ level it is the set of its organs; the structure of an organism is the totality of its bonding relations, such as connecting tissues and hormonal fluxes, and nonbonding relations, such as those of position and succession; and the mechanism of a system is composed of the processes that keep it alive, in the first place metabolism and interaction with its environment." (Bunge, 2016: 252-3)

Now the basic living-speaker scenario conceptual corpus linguistics faces is this:

> "A baby, once born, engages itself in postnatal experience of all kinds non-stop until death. During this "from womb to tomb" lifespan, it undergoes stages from pre-speech, to babbling, to talking freely, to writing (if educated) and finally to speechless death."

To conceptualize such lifespan development of speaker in such a way that it can be dealt with in terms of corpus linguistics, Gu and Xu (2013) and Gu (2016) have proposed a lifespan scheme of data development. In a nutshell, living is coextensive with experiencing which is coextensive with meaning-making that is equivalent to building a lifespan big data store. The big data formula reads like this:

The multi-dimensional self {...{...}...}

1. "Self", a technical term here, is a special data folder as it were that maintains the dynamic fluid data from the living-experiencing-meaning-making scheme;
2. "Dimension", also a technical term, stands for a specific aspect or property of the speaker under investigation. "Multi-dimensional" indicates our ontological commitment to the postulate that the human speaker is a system with multi-complexity.
3. {...{...}...} stands for a lifespan data set with many sub-sets.

The big data formula gives rise to a whole person model of speaker, as shown in Figure 4.

Figure 4 - *A whole person model of speaker (Gu, 2016: 488)*

The living being consists of the experiencing self, the meta-self, and the institutionalized general self. The experiencing self refers to the ever-going, moment-by-moment, multimodal interactions with the world, while the meta-self refers to the online or offline reflections on the experiencing self. The developing infant does not have a meta-self until about year 3. The institutionalized general self, on the other hand, is a set of identities the self has co-constructed with the community while making his living in it. In China, not until about the late 1980s, did the newborn baby was simultaneously given an institutionalized general self, i.e., national ID number.

Each self is modeled from a range of perspectives. Take the experiencing self for example. It has dimensions such as physiological, psychological, linguistic, learning, familial, working, socializing, spiritual/religious, etc. These dimensions of the experiencing-self undergo various phases of development. The meta-self, the ability of formulating second-order beliefs, desires, evaluations, etc., has the dimension of online reflection, and the I-me reflection.

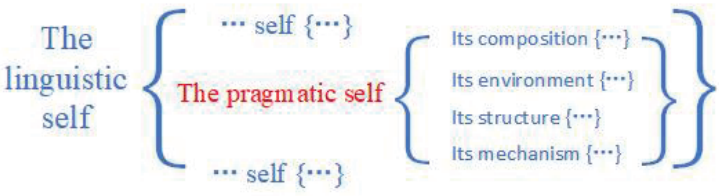
5.4 Critique of whole person model

The whole person model is in accordance with Bunge's systemic world view mentioned in 5.2 above. The five postulates – everything being a system, having emergent property, systemic not sectoral, all ideas into systems, and coherent validity for all – are upheld by the model. We cannot demonstrate them here for lack of space.

The model, against the yardstick of Bunge's composition-environment-structure-mechanism quadruple, suffers from many loopholes. Each dimensional self {...{...}...} should be further specified in terms of composition, environment, structure, and mechanism separately. Take the experiencing self for example. In view of the environment, it should be further specified into the experience-expectant growth self and the experience-dependent growth self (for this distinction see Greenough et al., 1987). The former refers to the fact that the infant's certain functions require basic experiences in order to develop, e.g., visual development requires visual stimulations of light from the environment, rooting behavior requires the stimulation from the breast nipple or the mock-up one, and most importantly, the development of speech requires the stimulation of speech sounds from caregivers. The latter, on the other hand, refers to the fact that some brain functions depend upon particular experiences, e.g., acquisition of a specific speech/language, e.g., a baby born and grown in Rome will acquire Italian due to specific input of speech, whereas in Beijing Chinese (i.e., Putonghua, a common language).

Take the pragmatic self for another example. It is a sub-set of the linguistic self, meanwhile it has its own sub-sets, which specify the quadruple, as shown in Figure 5.

Figure 5 - *The pragmatic self and its quadruple*

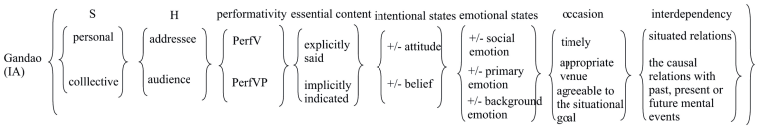


The pragmatic self shown in the figure can be further fine-tuned in terms of sub-sub-sets e.g., the illocutionary act self {...}. Gu (2013: 317) proposes an octet scheme as follows:

> "The Illocutionary act/force {S{ }, H{}, performativity {}, the essential content {}, the intentional states{}, the emotional states{}, the occasion {}, the interdependency {}}"

Adopting this scheme, the Chinese illocutionary act of *gandao* (feel) is conceptually analyzed as follows (see Figure 6):

Figure 6 - *Illocutionary act of gandao: a conceptual model (Gu 2013: 318)*



As the sample indicates it, Bunge's quadruple is substantiated in the octet. The bonding/non-bonding relations are handled in the sub-set of interdependency.

Up to this point one may wonder if such conceptual corpus linguistics leads to a jumble and a dead end. Our answer is NO. Because the whole enterprise is built on the basic set theory and Bunge's factual sciences. All sorts of selves {...{...}...} make a consistent and coherent whole through such logical operations as relations by union, and/or disjoint. As shown in Figure 6, members of the octet set hold logical union relations between them, while the values of each member, viz. the sub-sub-set members, are primarily disjoint between them, only under rare conditions are union. We cannot go to details here. Admittedly all we have achieved so far about conceptual corpus linguistics is only a tip of iceberg. It only points to a direction that may be fruitfully explored.

6. *Ideal corpus linguistics*

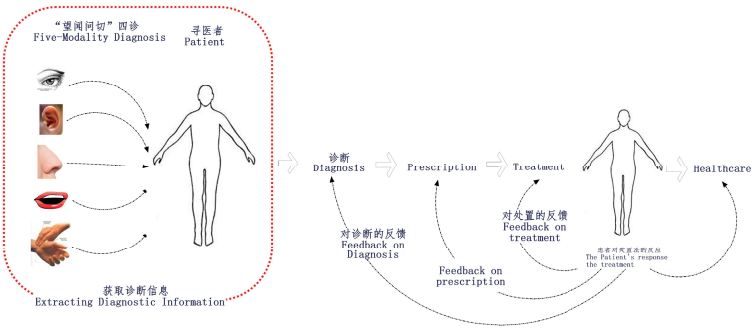
The conceptual corpus linguistics sketched above centers itself around the speaker as a whole person, which is the only "concrete, real thing" (in Bunge's terminology) that factual sciences deal with, thus treating speech as facts derived from real thing, and language as a construct.

Our position is that if corpus linguistics wants to be part of factual sciences, it must abide by this fundamental ontology.

Assuming that we endorse the ontology, we proceed to our next task, viz. conceptualizing ideal corpus linguistics. Ideal corpus linguistics, admitting the inadequacy of technology and limitations of human knowledge, attempts to think big and try to reach out to the bottom of iceberg. The central concern is how to compile a corpus that approximates, as faithfully, as accurately, as fully as possible, the human speaker's total saturated experience with total saturated signification. Eating a Roasted Peking Duck in a restaurant is a sequenced series of TSE-TSS states. An ideal corpus linguist, attempting to compile a corpus capturing this type of activity, faces a twofold task: (1) Make use of all sorts of data-capturing tools, currently available ones as well as imagined ones, (2) design a scheme for segmenting and annotating the collected data, manually or automatically or both. The twofold task is examined critically in depth in Gu (2009).

The practice of traditional Chinese medicine (TCM for short) provides us a real-life case study. A feeling-ill person comes to see a TCM doctor. The two are engaged in a TSE-TSS interaction empowered by natural multimodalities. The encounter however is unequal in terms of knowledge power. What is pertinent to our discussion here is the diagnostic method and the tools of data collection. The diagnostic method is called in Chinese sizhen (四诊), i.e., four ways of diagnosing, which are in our terminology five ways: by looking at, hearing, smelling, touching, and questioning, as visualized in Figure 7.

Figure 7 - *Diagnosing in TCM*



In view of ideal corpus linguistics, the practicing doctor is collecting multimodal data through natural multimodalities. Over three thousand years, such data collection and reflected practices have yielded thousand volumes of documents – TCM knowledge for short, and have saved a countless number of lives.

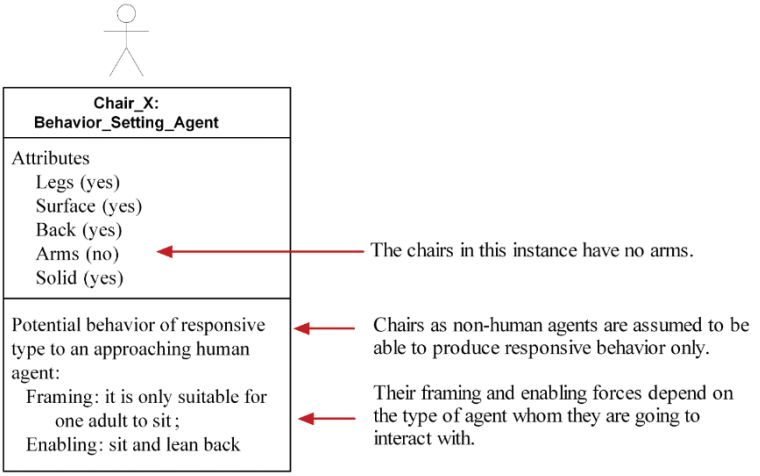
An agent-oriented modeling language (AOML) has been proposed in Gu (2006b, 2009), and can be used as an empowering tool for ideal corpus linguistics. As practicing corpus linguists well know, tools currently available, are technology-driven, i.e., empowered by the techniques, algorithms available to tool developers. There are programming languages (e.g. Java), mark-up languages (e.g. XML, TML, RDF), and modeling languages (e.g. UML). These languages are designed to talk to machines, and are not designed for corpus linguists to deal with raw data. XML, TML and RDF are in a way designed to deal with raw data, to mark them up so as to give them a machine-processable structure. But they do not offer much help to corpus linguists who face the problem of how to get the raw data first. AOML, in contrast, is designed to empower the practicing corpus linguist so that s/he can formalize human-interpreted TSE-TSS interactions. A telling advantage of doing so is that the formalization is quite intuitive to the corpus linguist on the one hand, and is easily convertible to software programming on the other.

The notion of agent has recently been given a great deal of attention in artificial intelligence (see, e.g., Hexmoor et al. 2003, Alonso et al. 2003, Ye & Churchill 2003, Russell & Norvig 2003, Wagner 2004). The AI agent is an automaton, and it is a part of a programming metalanguage. The agent of AOM is intended to be part of a modeling metalanguage.

The AOM conceives of the world, real or non-real, as consisting of agents of various kinds who interact with one another by way of (i) exchanging attributes, and (ii) exercising framing and enabling behaviors. This way of conception works quite naturally with modeling human agents. It takes some imagination to apply it to the interaction between, say, a human agent and a chair. In AOM, a chair, once engaged in interaction with an approaching human agent, becomes an agent as well. It has its own attributes, and exerts framing and enabling behaviors on the human agent. For instance, it enables the human agent to sit on it, and at the same time frames the human

agent's behavior, e.g. it does not allow the latter to stand on its arm (see Figure 8 for demonstration). When the same chair interacts with a fly, on the other hand, the framing behavior it exercises on the human agent disappears, and instead it enables the fly to perch on its arm very happily.[4]

Figure 8 - *Chair_Agent construct (quoted from Gu 2009: 447)*



It is obvious to corpus linguists that this notion of agent is quite intuitive. It is based on Gibson's (1986) ecological approach to perception, and is closely associated with his notion of affordance.

7. *Practical corpus linguistics*

To pre-empt some potential misunderstanding, it is important to bear it in mind that the term practical corpus linguistics is not meant to characterize some current practices. It is a relative concept circumscribed by conceptual corpus linguistics, and ideal corpus linguistics, as discussed above. Viewed from the perspective of software engineering (Chang, 2001, 2002, 2005), conceptual corpus linguistics represents the *stage of conceptual design*, and ideal corpus linguistics the *stage of data modeling*. Practical corpus linguistics, on the other hand,

[4] This paragraph is quoted from Gu (2009: 445-446).

deals with the third *stage of implementation/ application/evaluation*. The three-stage design and development ensure consistency, coherence, efficiency and effectiveness. This practice is rarely followed in corpus linguistics for several reasons. One is that it often takes years to compile a corpus; another is that it often takes cross-disciplinary, cross-department teams to complete a whole project, and thirdly, it is left to end users to decide how to use a compiled corpus. Thanks for these reasons, corpus linguistics may result in an incoherent, disintegrated state of affairs, as McEnery and Hardie observe:

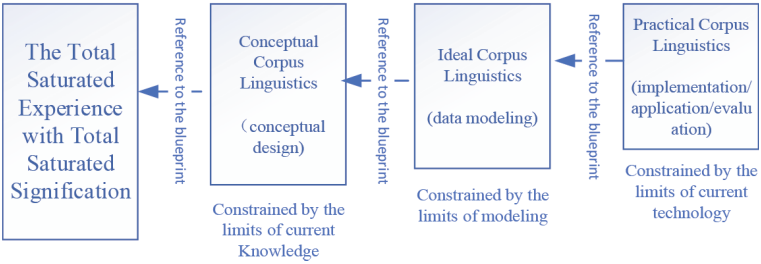
> "it is very important to realise that corpus linguistics is a *heterogeneous* field. Differences exist within corpus linguistics which separate out and subcategorise varying approaches to the use of corpus data." (McEnery and Hardie, 2012: 1; italics added)

Our proposal for separating practical corpus linguistics from the other two, meanwhile acknowledging the constraints imposed by practicalities such as the three reasons above, is intended to send a message that corpus linguistics is not merely practical and methodological, as it is held by many practitioners. McEnery and Hardie (2012: 1) capture this general spirit quite well:

> "What is corpus linguistics? It is certainly quite distinct from most other topics you might study in linguistics, as it is not directly about the study of any particular aspect of language. Rather, it is an area which focuses upon *a set of procedures, or methods*, for studying language". (italics added)

Practical corpus linguistics, in our conception, is both practical in the sense of constructing actual corpora for various purposes and theory-motivated in the sense of checking if it meets the blueprints laid down by ideal corpus linguistics and conceptual corpus linguistics, the blueprint of which is the TSE-TSS. Figure 9 visually demonstrates these relations.

Figure 9 - *Blueprint and reference relations in corpus linguistics*



The visualization makes it transparent the self-evaluation mechanism as an intrinsic design feature. The real-life TSE-TSS checks and evaluates conceptual corpus linguistics which in turn checks and evaluates ideal corpus linguistics, which in turn checks and evaluates practical corpus linguistics. Practical corpus linguistics, on the other hand, has an extra dimension of check-evaluation, namely by practical uses of corpus as well as by the demands of stakeholders. Such check-evaluation, surely very important, is an external one in our conceptualization. McEnery and Hardie (2012: 14-16) address the issues of "total accountability", "falsifiability" and "replicability". These three assessments are mainly targeted at the use or misuse of a corpus. It is also external to our design.

One may note that rectangles in Figure 9 are different in size and arranged in an order of from the largest on the right to the smallest on the left. This is intended to signal the underlying systems thinking as advocated by Bunge (synopsized in 5.2 above).

8. *Final summative reflections*

It is time to round up our journey and take stock. We started with a review of four positions represented respectively by Leech, Chafe, Sinclair and Gu: We witnessed a broadening scope of theorizing corpus linguistics, from about the *language performance*, to about the *mind* of language user, to about the *experiencing participant*, and finally to about the *whole person*.

The whole person, in Bunge's theory of ontology, is the only concrete materialistic thing that can be studied according to the principles and methodologies of factual sciences. As Bunge has argued, logically, linguistics aiming at idealized and abstract phenomenon violates the postulates of factual sciences, hence it is pseudoscientific at best.

Consequently, this paper argues that the fundamental ontological foundation of corpus linguistics is the postulate that there is no such linguistic fact in itself, and that there are only linguistic facts as states of a speaker. This point is worth reiterating: *What is recorded when a corpus compiler switches on a tape recorder is not a linguistic fact unless it is taken as a state or a sequence of states of the speaker being recorded*.

What are the logical conclusions to be drawn from the ontological foundation of corpus linguistics above?

8.1 The Postulate of Speaker Changeability or the Postulate of Language Pseudo-Changeability

The statement that language always changes is pseudo-scientifically valid and misleading. To construct a corpus to map out the state of the art of a language is equally pseudo-scientific and misleading. It is the live speaker that always changes from one state to another.

> "A change is an event or a process, whether quantitative or qualitative or both. Whatever its nature, a change is a modification in or of some thing or things: more precisely, it consists in a variation of the state of an entity. To put it negatively, there is no change separate from things – nor, indeed, are there changeless things even though some change slowly or only in certain limited respects. The world, then, consists of things that do not remain in the same state forever. This metaphysical hypothesis is an extrapolation from both ordinary experience and scientific knowledge." (Bunge, 1977: 215)

This speaker changeability postulate is not difficult to comprehend. Corpus linguists however would find it hard to buy in practice. For it requires that the speaker's state of affairs, physical and mental, be recorded as the primary data source. One may counterargue that this is practically infeasible and unnecessary for most purposes. The counterargument holds water but at its own peril: The corpus data is pseudo-scientific, incomplete and limited in use.

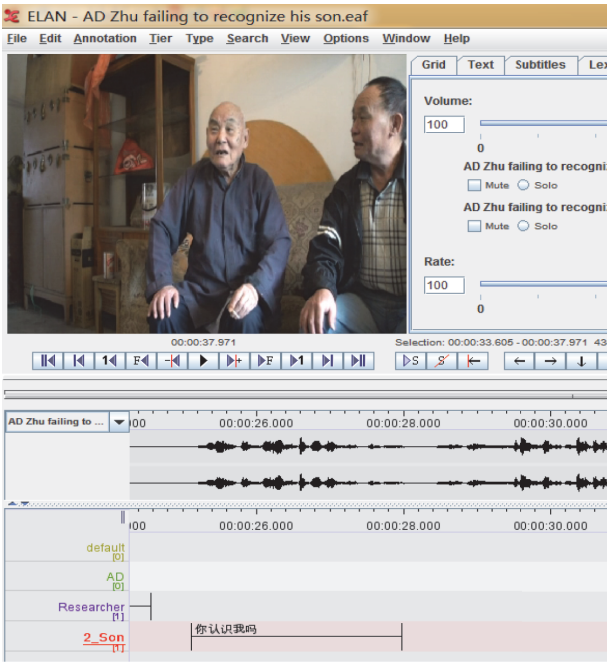
The present author used to hold this counter-argument himself until he embarked upon compiling two specialized corpora, one for autism spectrum disorder children, viz. multimodal corpus of ASD child discourse (MC-ASD-CD), and the other for ageing population, viz. multimodal corpus of gerontological discourse (MCGD). Both corpora have been under construction for a decade now. One big challenge is the issue of privacy protection. Permission to record

is hard to come by. Our solution to the challenge is to form research partnerships with ASD therapy centres, neurology departments of hospitals. In this way trust from patients as well as caregivers becomes much easier to attain. Furthermore, ethics concerning data collection is also secured since it has to gain approval from the institutional ethics committee before the recording starts.

Note that to segment and annotate audio-video streams of two discourse types, one crucial information that must be annotated is the contemporary states of the child or the ageing person. Without this crucial information the corpus linguist would study everything but the child or the ageing person, which is the very purpose of compiling the corpora. To illustrate the point, here is a talk exchange between an eighty-one year old AD (AD), the AD's second son (SS) and a corpus compiler (CC).

– CC: (pointing at SS) Do you know him?
– AD: (looking at his second son) No.
– CC: (pointing at himself) Do you know me?
– AD: (looking at CC) No.

Figure 10 - *A talk exchange between AD and others (quoted from Gu 2015: 466-7)*



One can no doubt do conversation analysis, pragmatic analysis, etc. All these descriptive analyses will only have some bearing on the theories of CA or pragmatics unless we treat talk exchanges substantiating properties of real-life speakers, and ongoing changing states of the mind in particular. Perkins, studying ASD child discourse samples, rightly points out:

> "the child could be described as breaking Grice's Maxims of Quantity, Relevance and possibly Manner ('be brief'), but such descriptive labels do not get us very far when trying to design a remedial programme. One can hardly tell the child to 'stop breaking Grice's maxims'!" (2007: 31)

To tackle the speaker changeability problem, we have collaborated with two major hospitals and turned a walk-in clinic into a research lab. In this way the states of the speaker also include biodata and medical images data.

Finally, we must pre-empt a mis-construal of our conceptualization of corpus linguistics: we may be charged for methodological individualism. This charge bears no tooth, for Bunge's theory of systemism is consistently followed in our whole enterprise. The charge however is understandable for we have not dealt with a crucial property of a system, namely emergence for lack of space.

In a word, the tripartite scheme of conceptualizing corpus linguistics, as outlined above, is hopefully able to produce a linguistic theory bearing its own trade mark. Let us wait and see. Fingers across!

Acknowledge

This paper is part of the projects (Codes 20AYY011, 21&ZD294) funded by the National Social Science Foundation of China.

References

Alonso, Eduardo & Kudenko, Daniel & Kazakov, Dimitar. 2003. *Adaptive Agents and Multi-Agent Systems*. Berlin: Springer.

Biber, Douglas & Johansson, Stig & Leech, Geoffrey & Conrad, Susan & Fingegan, Edward. 2000. *Longman Grammar of Spoken and Written English*. Beijing: Foreign Language Teaching and Research Press.

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

Bunge, Mario Augusto. 1977. *Treatise on Basic Philosophy Volume 3: Ontology I: The Furniture of the World*. Dordrecht: D. Reidel Publishing Company.

Bunge, Mario Augusto. 1979. *Treatise on Basic Philosophy Volume 4: Ontology II: A World of Systems*. Dordrecht: D. Reidel Publishing Company.

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

Bunge, Mario Augusto. 2016. *Between Two Worlds: Memoirs of a Philosopher-Scientist*. Switzerland: Springer International Publishing.

Carter, Ronald & McCarthy, Michael. 2006. *Cambridge Grammar of English*. Cambridge: Cambridge University Press

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.

Chafe, Wallace. 1994. *Discourse, Consciousness, and Time*. Chicago: The University of Chicago Press.

Chafe, Wallace. 2018. *Thought-Based Linguistics*. Cambridge: Cambridge University Press.

Chang, Shi-kuo. 2001. *Handbook of Software Engineering & Knowledge Engineering, Vol. 1: Fundamentals*. New Jersey: World Scientific Publishing Co. Pte. Ltd.

Chang, Shi-kuo. 2002. *Handbook of Software Engineering & Knowledge Engineering, Vol. 2: Emerging Technologies*. New Jersey: World Scientific Publishing Co. Pte. Ltd.

Chang, Shi-kuo. 2005. *Handbook of Software Engineering & Knowledge Engineering, Vol. 3: Recent Advances.* New Jersey: World Scientific Publishing Co. Pte. Ltd.

Chomsky, Noam. 1986. *Knowledge of Language: Its Nature, Origin, and Use*. Westport: Praeger.

Dillinger, Mike. 1990. On the concept of 'a language'. In Weingartner, Paul & Dorn, Georg J.W. (eds.), *Studies on Mario Bunger's Treatise*, 10-38. Amsterdam: Rodopi.

Fodor, Jerry A.. 1983. *The Modularity of Mind: An Essay on Faculty Psychology*. Cambridge, MA: A Bradford Book.

Gibson, J.J.. 1986. *The Ecological Approach to Visual Perception*. Hillsdale, New Jersey: Lawrence Erlbaum Associations, Inc., Publishers.

Goffman, E.. 1963. *Behavior in Public Places*. New York: The Free Press.

Greenough, William T. & Black, James E. & Wallace, Christopher S.. 1987. Experience and Brain Development. *Child Development*, 58. 539-559.

Gu, Yueguo. 2006a. Multimodal text analysis: a corpus linguistic approach to situated discourse. *Text and Talk*, 26-2. 127-167.

Gu, Yueguo. 2006b. "Agent-oriented modeling language, Part 1: Modeling dynamic behavior" (基于角色的建模语言(AML)一：动态行为建模). In *Proceedings of the 20th International CODATA Conference, Beijing*, 21–47. Published by the Information Centre, the Chinese Academy of Social Sciences.

Gu, Yueguo. 2009. From real-life situation to video stream data-mining. *International Journal of Corpus Linguistics* 14:4, 433-466.

Gu, Yueguo. 2013. A conceptual model of Chinese illocution, emotion and prosody. In Tseng, Chiu-yu (ed.), *Human Language Resources and Linguistic Typology*, 309-362. Taibei: Academia Sinica. Pp.

Gu, Yueguo. 2015. Multimodality and linguistic research. (多模态感官系统与语言研究). *Journal of Contemporary Linguistics*《当代语言学》Vol. 17, No. 4, 448-469.

Gu, Yueguo. 2016. Multimodal experiencing, situated cognition and big data with a demonstrative analysis of a newborn baby. 当下亲历与认知、多模态感官系统与大数据研究模型。*Contemporary Linguistics*,《当代语言学》第18 卷第4 期475－513 页。

Gu, Yueguo & Xu, Xunfeng. 2013. Alzheimer's disease patient discourse: A multimodal corpus linguistics approach. *Plenary speech at the 5th Symposium on Functional Linguistics and Multimodality*. The Polytechnic University of Hong Kong.

Hexmoor, Henry & Castelfranchi, Cristiano & Falcone, Rino. 2003. *Agent Autonomy*. Boston: Kluwer Academic Publishers.

Leech, G.N.. 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.

Legge, James. 2008 [1891]. *Tao Te Ching* by Lao Tse. The Floating Press.

Linell, Per. 2005. *The Written Language Bias in Linguistics: Its Nature, origins and Transformations*. London: Routledge Taylor & Francis Group.

McEnery, Tony & Hardie, Andrew. 2012. *Corpus Linguistics: Method, Theory and Practice*. Cambridge: Cambridge University Press.

Perkins, Michael. 2007. *Pragmatic Impairment*. Cambridge: Cambridge University Press.

Russell, Stuart J. & Norvig, Peter. 2003. *Artificial Intelligence*. New Jersey: Pearson Education, Inc.

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

Skyttner, Lars. 2001. *General Systems Theory: Ideas and Applications*. Singapore: World Scientific Publishing Co. Pte. Ltd.

Wagner, Thomas A.. 2004. *Application Science for Multi-Agent Systems*. Boston: Kluwer Academic Publishers.

Ye, Yiming & Churchill, Elizabeth. 2003. *Agent Supported Cooperative Work*. Dordrecht: Kluwer Academic Publishers Group.

DOI: 10.17469/O2106SLI000003

TAKEHIKO MARUYAMA

Designs and Analyses of Japanese Speech Corpora

Since 2000 a series of Japanese speech corpora has been under development in NINJAL. In this paper, three of them, the CSJ (Corpus of Spontaneous Japanese), the CEJC (Corpus of Everyday Japanese Conversation) and the SSC (Showa Speech Corpus) will be introduced. These corpora are transcribed and morphologically analysed in a unified format so as to be mutually comparable. Users can access them through the web site "*Chunagon*" and search the data with part of speech information. Also, all the sound files, transcription and related data are distributed individually.

Keywords: speech corpora, CSJ, CEJC, SSC, filled pauses.

1. *Introduction*

A "corpus" can be defined as "a collection of written or spoken material stored on a computer and used to find out how language is used" (*Cambridge Dictionary*). The Survey of English Usage project, initiated in 1959 under Randolph Quirk at University College London, gathered a total of 100 million words of written and spoken British English and used it for the survey. In 1964 the first electric corpus, the "Brown Corpus", was created at Brown University; it consists of 100 million words of written American English. These two corpora opened up corpus linguistics as a new field of linguistic research.

The term "speech corpus" used here refers to a systematic and large collection of speech recorded in real-world situations. It contains digital audio files of various types of speech, such as monologues, conversations, and historical recordings. Additionally, various linguistic annotations have been added to speech corpora, including transcriptions, part of speech tags, utterance units, and meta-data. Using large speech corpora with rich annotation, linguists can empirically analyse how spoken language previously has been and presently is used in real life.

Since 2000 a series of Japanese speech corpora has been under development in NINJAL. In this paper, three of them, the CSJ (Corpus

of Spontaneous Japanese), the CEJC (Corpus of Everyday Japanese Conversation) and the SSC (Showa Speech Corpus) will be introduced. The distributions of filled pauses and first-person pronouns across these speech corpora will be compared to illustrate the utility of this kind of corpus design.

2. *Japanese speech corpora in NINJAL*

2.1 NINJAL

Established in 1948, NINJAL (National Institute for Japanese Language and Linguistics) has conducted scientific research on Japanese language for more than 70 years. In 1952 NINJAL started recording daily conversations in various situations in order to describe the intonation patterns, vocabulary, sentence length and structures, and types of words used in colloquial Japanese. The results were published in the report *Danwago no Jittai* ('Research in Colloquial Japanese') in 1955. Approximately 40 hours of speech were recorded and approximately 30 hours were analysed, making this a world-pioneering work in corpus-based research into colloquial speech.

Recently NINJAL has been designated as the Centre for Excellence of corpus creation in Japan. NINJAL has designed and constructed a series of Japanese corpora: the Corpus of Spontaneous Japanese (CSJ), Balanced Corpus of Contemporary Written Japanese (BCCWJ), Corpus of Historical Japanese (CHJ), International Corpus of Japanese as a Second Language (I-JAS), Corpus of Everyday Japanese Conversation (CEJC), Corpus of Japanese Dialects (COJADS), and Showa Speech Corpus (SSC). These corpora cover a wide range of Japanese language data, both written and spoken (monologue and dialogue), a span of 1300 years of Japanese history, a wide variety of dialects, and both learners and native speakers. Users may access the corpora through the web application "*Chunagon*" (https://chunagon.ninjal.ac.jp/).

2.2 Corpus of Spontaneous Japanese (CSJ)

The CSJ is the first of these corpora, released to the public in 2004 (NINJAL 2006). It includes 651 hours and 7.52 million words of spontaneous speech (mainly monologue) recorded from 1999 to 2001. This provided a new language resource to the field of linguis-

tics, especially for Japanese phonetics, phonology and syntax. It also contributed to the development of techniques for speech processing systems in areas such as automatic speech recognition and natural language processing (Maekawa 2004; NINJAL 2006).

The speech recorded in the CSJ can be classified into two categories: Academic Presentation Speech (APS) and Simulated Public Speaking (SPS). APS is composed of live recordings of presentations to various academic societies. SPS contains general speeches and comments by laypeople on everyday topics, given to small audiences. A relatively formal speaking style is observed in APS, while a casual speaking style is observed in SPS. Most monologues in the and SPS are 10-15 minutes long.

The CSJ is characterised by its rich annotations. It contains very precise transcription, segment labels and intonation labels for phonetics and phonology, part of speech tags for morphology and lexicology, clause-boundary labels, dependency structure, and discourse structure for syntax and discourse analysis. Also, speaker's information is provided for use in sociolinguistic analysis. All annotation is stored in a relational database which users can manipulate to retrieve data efficiently.

Table 1 shows the distribution of filled pauses appearing in the SPS core data set, which consists of 226,902 words. Let's compare two groups of speakers: male and female. The ratio of filled pauses used by male speakers is 7.7% of the total number of words, whereas the ratio of filled pauses used by female speakers is 5.0%. This indicates that male speakers produce filled pauses more frequently than female speakers.

Looking at the distribution of each form, clear differences can be observed between male and female speakers. Of their filled pauses, male speakers' use of *ma* and its elongated form *ma:* comprises 29%, while female speakers' use of those forms comprises only 17%. On the other hand, female speakers' use of *ano:* and *an*o comprises 37%, while male speakers' use of these forms makes up only 15%. This distribution shows that male and female speakers choose different types of filled pauses: male speakers using *ma:* and female speakers using *ano:*. The result shows a sociolinguistic pattern of behaviour in Japanese monologue speech.

Table 1 - *Filled pauses observed in the CSJ (SPS core)*

Male (117,411 words)			*Female (109,491 words)*		
2,080	22.9%	*e:*	1,076	19.7%	*ano:*
1,361	15.0%	*ma:*	955	17.4%	*ano*
1,276	14.1%	*ma*	732	13.4%	*e:*
767	8.5%	*ano:*	549	10.0%	*ma*
569	6.3%	*ano*	419	7.7%	*ma:*
433	4.8%	*e*	256	4.7%	*e*
315	3.5%	*e:to*	181	3.3%	*sono*
263	2.9%	*sono*	149	2.7%	*n*
203	2.2%	*n*	147	2.7%	*e:to*
9,072	100%	total	5,473	100%	total

2.3 Corpus of Everyday Japanese Conversation (CEJC)

The CEJC consists various conversations in daily situations, such as conversations during dinner with the family at home, meetings with colleagues at work, and conversations while driving (Koiso *et al.* 2016, 2022). It includes 200 hours and 2.4 million words of daily conversations in a balanced selection. To estimate distributions of various daily conversations, a survey of everyday conversational behaviour had been conducted previously with about 250 adult Japanese informants. The survey asked when, where, how long, with whom, and during what kind of activity informants was engaged in conversations in their daily life. Based on the results, conversation forms (chat, business, meeting), places (home, workplace, public, indoor, outdoor, transport), and activities (housework, work, eating, private, communal, transfer, other) were defined as a measure of the design of a balanced corpus (Koiso *et al.*, 2016, 2022).

The most distinctive feature of the CEJC is that it provides multi-directional video files, as shown in Figure 1. The entire speech situation were captured by the three cameras, thus speakers' gaze movement, nodding, and gestures can be observed. The CEJC also serves audio files, transcription, TextGrid files for Praat, eaf files for ELAN, part-of-speech tags, and speaker's information.

Figure 1 - *An example of video file in the CEJC*

Table 2 - *Filled pauses observed in the CEJC (ver. 2018)*

Male (260,183 words)			*Female (349,144 words)*		
2,160	65%	*ano*	2,042	63%	*ano*
412	12%	*sono*	320	10%	*sono*
142	4%	*e:*	186	6%	*etto*
113	3%	*e:to*	147	5%	*e:to*
75	2%	*n*	115	4%	*a:no*
73	2%	*n:*	92	3%	*n*
71	2%	*etto*	66	2%	*a*
52	2%	*a*	61	2%	*e:tto*
3,305	100%	total	3,263	100%	total

Table 2 shows the distribution of filled pauses appearing in the CEJC (ver. 2018), which consists of 609,327 words. The ratio of filled pauses used by male speakers is 1.2% of the total number of words, whereas the ratio of use by female speakers is 0.9%. Compared to the result from the CSJ, the proportion of filled pauses appearing in conversation is considerably lower than in monologue.

Comparing male and female speakers, the distribution is almost identical. For both groups, use of *ano* is just over 60% and use of *sono* is around 10% of all filled pauses. This result implies that differences in style – monologue vs. conversation – produce different linguistic

behaviours according to gender. Male and female speakers behave differently in monologue situations and behave similarly in conversation.

2.4 Showa Speech Corpus (SSC)

The SSC is a collection of conversations and monologues recorded from the early 1950s to the early 1970s (a span included in the Showa Era) by NINJAL (Maruyama 2020, 2021). The earliest of these recordings were collected in 1952 in the survey of colloquial speech mentioned above. Various types of colloquial speech were sampled according to the sampling frame, including region (uptown, downtown, outskirts), place (home, neighbourhood, school, place of work), gender, age, educational background, and number of speakers. Also, some monologues such as academic presentations and congratulatory addresses were recorded.

Figure 2 is a picture of recording a conversation in the 1950s.

Figure 2 - *Recording conversation in the 1950s*

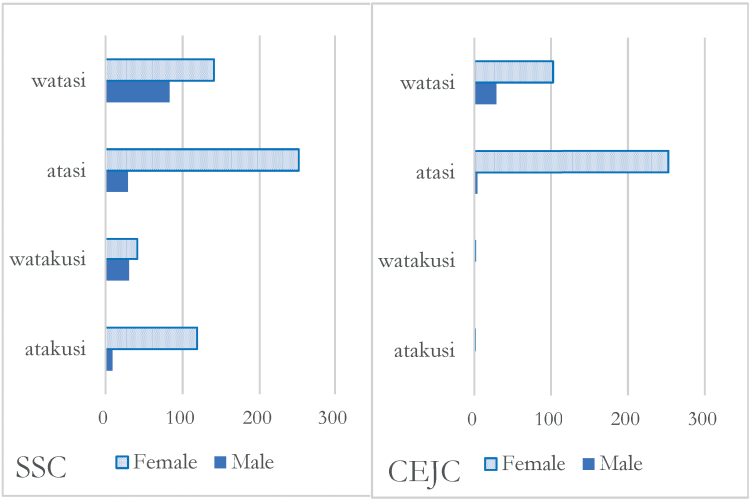
The original reel-to-reel materials, which had been stocked in the NINJAL archive for decades, were digitized by the 1990s. The author of this paper collected the sound files in the archive, newly transcribed them, morphologically analysed them for part of speech, and annotated them with meta-data. A series of this data was compiled into the SSC, which was released in 2022. The SSC includes 73 conversations

(in total approximately 27 hours) and 50 monologues (approximately 17 hours). Most of the conversations are casual chats involving men and women of all ages, whereas all the monologues were lectures or talks held at NINJAL.

One of the characteristic features of the SSC is that it provides a basis for diachronic analysis of spoken Japanese. As expected, some "old-fashioned" linguistic expressions can be observed in the SSC, including aspects of pronunciation, intonation, vocabulary, and grammar. Connecting the SSC to the contemporary speech corpora like the CSJ and CEJC creates a diachronic speech corpus, covering the late 20th century and early 21st century (Maruyama, 2020, 2021).

In Japanese, the form of the first-person pronoun *watasi* has some morphological variations, *watasi*, *atasi, watakusi,* and *atakusi*. The forms *watakusi* and *atakusi* are more polite than *watasi* and *atasi*. Comparing the SSC conversation part to the CEJC, the distributions of first pronoun variants (per 0.1M words) can be illustrated as in Figure 3.

Figure 3 - *Distribution of first-person pronoun in the SSC and CEJC*



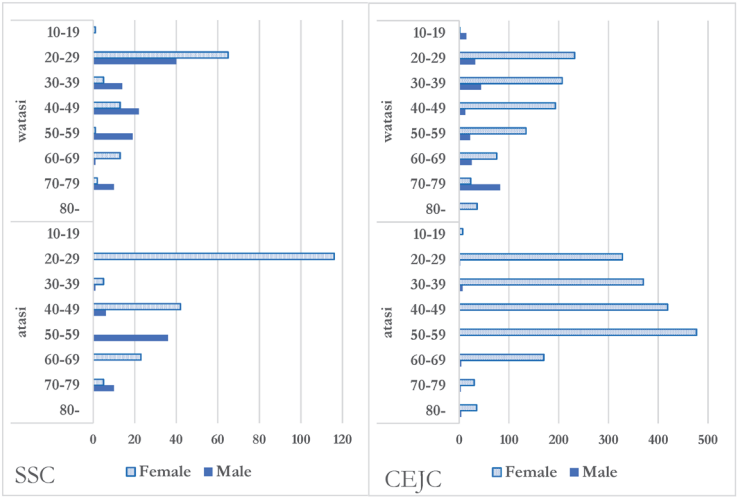
The numbers of uses of *watasi* and *atasi* by female speakers are almost the same between SSC and CEJC, so there has been no change during the 60 years. On the other hand, uses of *watasi* and *atasi* by male

speakers has declined significantly. The more polite forms *watakusi* and *atakusi* have completely disapeared in male and female speakers. These are diachronic changes in the use of first-person pronouns.

Figure 4 shows the number of uses of first-person pronoun variants broken down by speakers' age. In the SSC on the left side, *atasi* is shown to be used more by younger female speakers. On the other hand, in the CEJC on the right, *atasi* is used more by older female speakers. This can be considered to indicate that female speakers who were young in the 1950s are now in the older age group, and the tendency of using *atasi* is still present even today.

Looking at the distribution of *watasi* in the CEJC, the younger age groups use it more than the older age groups. From the past to the present day, it can be interpreted that use of the first-person pronoun by female speakers is in the process of transitioning from *atasi* to *watasi*.

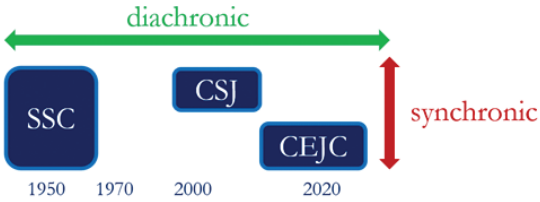
Figure 4 - *Distribution of first-person pronoun by speakers' age*



3. *Concluding remarks*

This paper outlines the CSJ, CEJC and SSC, three of the speech corpora designed and constructed in NINJAL. The positioning of the three corpora may be illustrated as follows.

Figure 5 - *Relationship of the CSJ, CEJC and SSC*



By comparing the CSJ, a corpus of monologue, with the CEJC, a corpus of daily conversation, it is possible to analyse the diversity of the synchronic spoken language. On the other hand, by linking the historical SSC corpus with the contemporary CSJ and CEJC corpora, it is possible to analyse diachronic changes in spoken language.

In the future, the development of more diverse speech corpora will enable comprehensive analysis and description of spoken Japanese in modern and contemporary times.

Acknowledgments

This study is supported by Grant-in-Aid for Collaborative Research Project of NINJAL "A multifaceted study of spoken language using a large-scale corpus of everyday Japanese conversation", and JSPS KAKENHI Grant Number 20H05630 and 16H03426. I would like to thank Dr Stephen Wright Horn for useful comments and help.

References

Koiso, Hanae & Tsuchiya, Tomoyuki & Watanabe, Ryoko & Yokomori, Daisuke & Aizawa, Masao & Den, Yasuharu. 2016. Survey of Conversational Behavior: Towards the Design of a Balanced Corpus of Everyday Japanese Conversation. In *Proceedings of LREC 2016*, 4434-4439. https://aclanthology.org/L16-1702/

Koiso, Hanae & Amatani, Haruka & Den, Yasuharu & Iseki, Yuriko & Ishimoto, Yuichi & Kashino, Wakako & Kawabata, Yoshiko & Nishikawa, Ken'ya & Tanaka, Yayoi & Watanabe, Yuka & Usuda, Yasuyuki. 2022.

Design and Evaluation of the Corpus of Everyday Japanese Conversation, In *Proceedings of LREC 2022*, 5587-5594.
http://www.lrec-conf.org/proceedings/lrec2022/pdf/2022.lrec-1.599.pdf

NINJAL 1955. *Danwago no jittai (Research in Colloquial Japanese)*. NINJAL Report 8. Tokyo: NINJAL.

NINJAL 2006. *Nihongo Hanasi Kotoba Kopasu no Kotikuho (Construction of the Corpus of Spontaneous Japanese)*. NINJAL Report 124. Tokyo: NINJAL.

Maekawa, Kikuo. 2004. Design, Compilation, and Some Preliminary Analyses of the Corpus of Spontaneous Japanese, In Yoneyama, Kiyoko & Maekawa, Kikuo (Ed.), *Spontaneous Speech: Data and Analysis*, 87-108. Tokyo: NINJAL.

Maruyama, Takehiko. 2020. On the Possibility of a Diachronic Speech Corpus of Japanese. In Bekeš, Andrej & Srdanović, Irena (Ed.), *Japanese Language from Empirical Perspective: Corpus-based studies and studies on discourse*, 219-234. Ljubljana: Znanstvena založba FF.
https://ebooks.uni-lj.si/zalozbaul//catalog/book/187

Maruyama, Takehiko. 2021. Diachronic Change of Spoken Japanese in the 20th Century: A Corpus Study. In Suzuki, Seiko & Cordereix, Pascal & Bergounioux, Gabriel (Ed.), *Mémoire sonore du Japon : le disque, la musique et la langue*, 17-32. Paris: Bibliothèque nationale de France.

DOI: 10.17469/O2106SLI000004

ANNE LACHERET-DUJOUR, PAOLA PIETRANDREA

Rhapsodie: Un treebank prosodico-sintattico per il francese parlato

Il corpus di francese parlato *Rhapsodie* presenta produzioni di 89 soggetti, uomini e donne, per una grande varietà di generi discorsivi (± spontaneo, ± pianificato, dialoghi faccia a faccia, interviste, trasmissioni radiofoniche e televisive). Il corpus è composto da 57 estratti brevi (lunghi in media 5 minuti ciascuno) per un totale di 3 ore di parlato trascritto ortograficamente (33000 parole) e fonologicamente, allineato al suono al livello di fonema, sillaba, parola, turno, pausa e sovrapposizione. L'annotazione sintattica e quella prosodica sono organizzate ciascuna in più livelli. In sintassi, oltre ad un'annotazione delle dipendenze sono annotate le liste e la struttura macro-sintattica. In prosodia sono annotate le unità prosodiche maggiori (periodi intonativi), le prominenze, le disfluenze, le unità prosodiche interne al periodo e i contorni intonativi.

Parole chiave: interfaccia prosodia-sintassi, prominenze, disfluenze, liste, francese.

1. *Introduzione*

Il treebank *Rhapsodie* è un corpus di francese parlato annotato sintatticamente e prosodicamente con l'obiettivo di modellizzare il ruolo svolto dall'intonazione e la sintassi nella segmentazione del discorso in unità elementari e condurre studi funzionali sulla struttura prosodico-sintattica del francese in diversi generi discorsivi. Per quanto riguarda la sintassi, combinando il lavoro iniziato con il corpus C-Oral-Rom (Cresti & Moneglia 2005) con la proposta di analisi elaborata a Aix en Provence da Blanche-Benveniste *et al.* (1990), si propongono un'annotazione morfosintattica, microsintattica e macrosintattica. Per quanto riguarda la prosodia, *Rhapsodie* costituisce il primo esempio di corpus annotato contemporaneamente al livello delle frasi prosodiche derivate dalle prominenze accentuali e al livello dei profili intonativi degli enunciati che lo compongono. Per quanto riguarda l'interfaccia prosodia-sintassi, l'allineamento temporale delle unità prosodiche e sintattiche permette di esplorare simultaneamente que-

sti due livelli di annotazione, studiare il loro allineamento, che è variabile nel francese parlato, e proporre un'interpretazione funzionale di queste variazioni (Lacheret-Dujour *et al*. 2019).

2. *Struttura del corpus e metadati*

Il corpus *Rhpasodie* è stato costituito per verificare l'ipotesi di una correlazione tra le caratteristiche prosodico-sintattiche di un discorso e le caratteristiche della situazione in cui è prodotto, cioè il suo genere[1] (v. tabella 1).

Tabella 1 - *Variabili linguistiche, situazione e genere del discorso: una relazione causale da modellizzare nella lingua parlata*

Caratteristiche della situazione
→ Caratteristiche prosodico-sintattiche
→ Genere discorsivo

Il corpus *Rhapsodie* è organizzato in 56 brevi testi (di 5 minuti in media ciascuno) rappresentativi di vari generi discorsivi, prodotti da un numero di parlanti sufficientemente ampio da evitare idiosincrasie individuali. I dati sono stati tratti da corpora già esistenti (Durand *et al.* 2009; Eshkol-Taravella *et al.* 2011; Branca-Rosoff *et al*. 2012), e completati da corpora multimediali raccolti specificamente per *Rhapsodie*.

Per strutturare i metadati sono selezionate sei caratteristiche situazionali maggiori. Innanzitutto, si distinguono monologhi, prodotti da un solo parlante, e dialoghi, prodotti da almeno due parlanti, in situazioni più o meno interattive. Si distinguono poi discorsi privati (faccia a faccia) e discorsi pubblici (conferenze, trasmissioni televisive o radiofoniche). Gli argomenti e i compiti (racconti personali, lezioni, interviste, sermoni, allocuzioni...) affrontati nei vari estratti sono molto vari e per quanto riguarda la parola pubblica, gli estratti rappresentano esempi di parlato su canali diversi: interviste politiche, dibattiti, trasmissioni di divulgazione scientifica, ecc... Infine, ogni testo è classificato secondo cinque variabili: (i) livello di pianificazione;

[1] Cfr. la nozione di «registro» di Biber & Conrad (2009), secondo i quali un registro è caratterizzato da una serie di tratti lessicali e grammaticali ricorrenti tipici di una varietà linguistica e associati a determinate funzioni comunicative generali.

(ii) grado di interattività; (iii) compito svolto dai parlanti; (iv) canale di comunicazione; (v) tipo di sequenza discorsiva maggiormente rappresentato nel campione. Per codificare e interrogare i metadati di Rhapsodie è stato utilizzato il formato IMDI-CMDI, sviluppato al Max Planck di Nimega (CMDI, http://www.clarin.eu/cmdi, Broeder *et al*. 2012). La tabella 2 presenta sinteticamente i metadati che possono essere interrogati nel corpus.

Tabella 2 - *Variabili situazionali nel corpus Rhapsodie*

Tipo di discorso	*Privato/Pubblico*	*Monologo*
		Dialogo
	<Pianificazione>	Spontaneo, semi-spontaneo, pianificato
	<Interattività>	Interattivo, semi-interattivo, non interattivo
	<Canale di comunicazione>	Faccia a faccia vs. Conferenza, trasmissione radiofonica, trasmissione televisiva
	<Tipo di sequenza>	Argomentativa, descrittiva, procedurale, oratoria

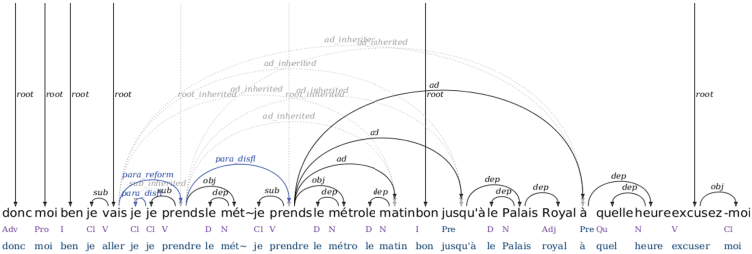
3. *L'annotazione sintattica*

L'annotazione sintattica del corpus *Rhapsodie* consiste in un'annotazione morfosintattica, un'annotazione microsintattica e un'annotazione macrosintattica.

L'annotazione morfosintattica è piuttosto classica: si identificano le parole, cioè token dotati di una valenza teorica; si assegnano dei lemmi a questi token; e per ogni lemma si precisa la parte del discorso.

L'annotazione microsintattica permette di precisare la struttura di dipendenza dei dati, ma anche le strutture di lista (figura 1). Per lista intendiamo quel dispositivo sintattico per il quale due o più parole occupano la stessa posizione in un albero di dipendenza (Blanche-Benveniste 1990; Kahane & Pietrandrea 2012; Masini *et al* 2018). È importante notare che molti fenomeni tipici del parlato come le riformulazioni, correzioni, richieste di chiarimento, di conferma, che vengono solitamente considerate come interruzioni della coesione sintattica, sono in realtà per la gran parte realizzate attraverso liste sintattiche, cioè enunciando parole che si trovano in una posizione sintattica già aperta nel discorso (1): questa caratteristica permette di assicurare nel parlato un certo livello di coesione sintattica.

Figura 1 - *L'annotazione morfosintattica e microsintattica in Rhapsodie*



(1) *alors < { { j'a~ | j'avais } beaucoup | j'avais beaucoup } trop peur de m'installer (comme ça) seule { d~ | dans } la brousse //* [Rhap-D2004, Rhapsodie]

'allora < { { io a~ | io avevo } davvero | io avevo davvero } troppa paura di andare a vivere (così) sola { ne~ | nella } foresta//'

L'annotazione macrosintattica del corpus Rhapsodie combina il modello di annotazione macrosintattica proposto a Firenze (Cresti & Moneglia 2005) con il modello di Aix en Provence sviluppato da Blanche Benveniste *et al.* (1990). Gli estratti sono innanzitutto segmentati in Unità Illocutorie (UI). All'interno di ogni UI è identificato un nucleo, cioè una sequenza portatrice della forza illocutoria dell'intera UI ed eventualmente altre unità accessorie prive di forza illocutoria che possono precedere, interrompere o seguire il nucleo (dette rispettivamente pre-nuclei, in-nuclei, post-nuclei). Sono inoltre annotati i marcatori del discorso. In (2) e (4), i pre-nuclei sono delimitati dal simbolo <, in (3) il post-nucleo è preceduto dal simbolo >, in (4) il marcatore del discorso è indicato tra virgolette "". Le UI sono delimitate dal simbolo //.

(2) *alors < là < la psychiatrie < c'est autre chose //* [Rhap-D0006, CFPP2000]
allora < da questo punto di vista < la psichiatria < quella è un'altra cosa //

(3) *ça a duré dix ans > le silence autour de moi //* [Rhap-D2001, Mertens corpus]
è durato dieci anni > il silenzio intorno a me

(4) *donc < moi < "ben" { je vais | { je|je} prends le mét~ | je prends} le métro le matin "bon" jusqu' au Palais Royal //*
quindi < io < "insomma" { vado | { io|io} prendo la met~ | prendo} la metro di mattina "insomma" fino a Palais Royal //

4. *L'annotazione prosodica*

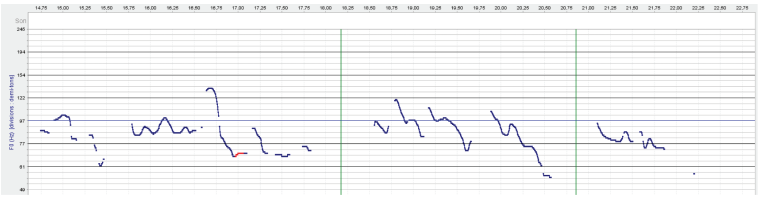
L'annotazione prosodica è stata condotta su due piani: sono stati identificati le frasi melodiche e i confini tra i segmenti, e sono stati poi identificati i profili interni alle unità. L'annotazione prosodica *Rhapsodie* ha tre specificità: (1) è un'annotazione bottom-up percettiva ('t Hart *et al.* 1990; Wightman 2002). E questo ha per conseguenza che le unità generate dal modello non sono necessariamente corrispondenti alle unità funzionali. Così il periodo intonativo, identificato da un confine terminale, non si allinea necessariamente alla clausola o all'atto linguistico. (2) Sono annotate tutte le disfluenze e sovrapposizioni. (3) Il dominio di proiezione dei contorni intonativi non è fissato a priori, ma definito dal ricercatore in funzione del suo obiettivo d'osservazione.

Ogni enunciato (o turno di parola per gli estratti dialogici) è segmentato automaticamente con il software Analor (Lacheret & Victorri 2002) in una successione di periodi intonativi (PI) sulla base di criteri essenzialmente acustici, (tavola 3, figura 2).

Tabella 3 - *Criteri di segmentazione in periodi intonativi*

Identificazione di una pausa di almeno 300 ms
Identificazione di un movimento di F0 ampio: il tratto [±ampio] è fissato in funzione dell'intervallo melodico, misurato in semitoni, tra l'ultimo estremo di F0 prima della pausa e la media di F0 su tutta la porzione che precede la pausa.
Identificazione di una reinizializzazione melodica dopo la pausa.

Figura 2 - *Segmentazione in 3 PI dell'enunciato* je pense aux nombreuses victimes de la tempête $_{PI}$ et à toute leur famille endeuillé e_{PI} dont nous partageons la peine $_{PI}$ *'penso alle numerose vittime della tempesta e a tutte le loro famiglie in lutto, di cui condividiamo il dolore' [D2004, Corpus Rhapsodie, tratto da un discorso presidenziale di Nicolas Sarkozy]*



La prominenza, cioè la proprietà di una sillaba di distaccarsi percettivamente dal suo contesto fonetico (Terken 1991; Wagner *et al*. 2015), è un indice cruciale che il ricevente usa per processare la prosodia on-line. L'annotazione manuale è stata effettuata con PRAAT (Boersma & Weenink 2010) da cinque annotatori non esperti e rivista da tre esperti. Sono stati distinti tre gradi di prominenza: le sillabe non prominenti (annotate con '0'), le sillabe debolmente prominenti (annotate con 'W' per *weak*), e le sillabe fortemente prominenti (annotate con 'S' per *strong*).

Una sillaba può essere al tempo stesso prominente e disfluente nel qual caso è annotata con (H'), e le marche di disfluenza possono essere variabili e combinabili: allungamento, pausa d'esitazione, ripetizioni, eccetera (Tabella 4).

Tabella 4 - *Etichettatura delle prominenze nell'enunciato* c'était assez assez terrible et les les maisons ont brûlé *'era davvero terribile e le case sono bruciate' [Rhap-D0003, Corpus PFC]; con dall'alto in basso: il livello delle sillabe, il livello delle prominenze, il livello delle disfluenze*

se	tE	a	se	a	se	te	Ri	ble
0	0	W	S	0	W	0	S	0
		H	H					

le	le	mE	zo~	o~	bRy	le
H						

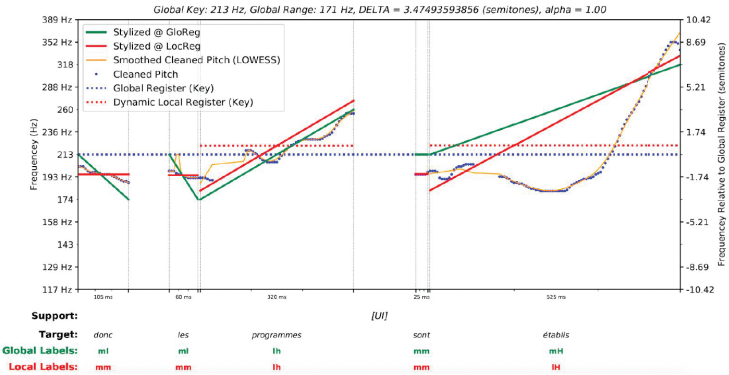
A partire dalle prominenze identificate (Prom), vengono applicate 3 regole che permettono di derivare la struttura prosodica di base (figura 3): (1) ogni sillaba prominente di un periodo marca il confine destro e la testa di un piede metrico (PM); (2) la testa di un piede metrico che si allinea con il confine destro di una parola ortografica marca la fine di un gruppo ritmico (GR); (3) la prima testa di un PM associata a una prominenza ('S') indica la frontiera destra di un gruppo intonativo (GI).

Figura 3 - *Generazione dell'albero prosodico del sintagma* le sherpa du président *'lo sherpa del presidente'*

Syl	l@	SER	pa	dy	pRe	zi	da~
Prom			W		W		S
PM		l@SERpa			dypRe		zida~
GR		le sherpa			du président		
GI				le sherpa du président			

Il tool SLAM+ è stato sviluppato per generare i contorni intonativi stilizzati di qualunque unità del corpus Rhapsodie (sintattica o prosodica). Il tool codifica tre informazioni principali (Yu Chen *et al.* 2019): i livelli intonativi attraversati da un contorno (da sovra-acuto a infra-grave), la direzione del contorno (ascendente o discendente), la sua forma (concava o convessa). Un insieme di etichette tonali permette di dare conto dei punti chiave d'un contorno (figura 4).

Figura 4 - *Generazione dei contorni delle parole e etichette tonali associati, estratto di [Rhap-M1001]*



4. *Conclusioni*

Il corpus Rhapsodie è scaricabile liberamente secondo i termini della licenza Creative Commons Attribution – Non commercial – Share Alike 3.0 France sul sito https://rhapsodie.modyco.fr/. Le registrazioni (wav/mp3), le analisi acustiche (F0 grezze, pulite manualmente e F0 stilizzate automaticamente, formato pitch), la trascrizione ortografica (txt), l'annotazione macrosintattica (txt e formato tabulare),

l'annotazione microsintattica (formato tabulare), l'annotazione e la segmentazione prosodica (textgrid, formato tabulare), e i metadati (xml, html) sono accessibili sia sotto forma di archivi sia campione per campione. I metadati possono essere interrogati *on line*. Le prime analisi funzionali del corpus (strutture micro e macrosintattiche, tratti prosodici, unità intonosintattiche maggiori) sono presentati in Lacheret *et al*. 2019.

Riferimenti bibliografici

Biber, Douglas & Conrad, Susan. 2009. *Register, Genre and Style*. Cambridge: Cambridge University Press.

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.

Boersma, Paul & Weenink, David. 2010. *Praat: doing phonetics by computer (Version 5.3)*. www.praat.org.

Branca-Rosoff, Sonia & Fleury, Serge & Lefeuvre, Florence & Pires, Matthew. 2012. *Discours sur la ville. Corpus de Français Parlé Parisien des années 2000 (CFPP2000).* <http://cfpp200.univ-paris3.fr>

Broeder, Daan & van Uytvanck, Dieter & Gavrilidou, Maria & Trippel, Thorsten & Windhouwer, Menzo. 2012. Standardizing a component metadata infrastructure. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, 1387-1390.

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

Durand, Jacques & Laks, Bernard & Lyche, Chantal. 2009. Le projet PFC (phonologie du français contemporain) : une source de données primaires structurées. In Durand, Jacques & Laks, Bernard & Lyche, Chantal (a cura di), *Phonologie, variation et accents du français*, 19-61. Paris: Hermès.

Eshkol-Taravella, Iris & Baude, Oliver & Maurel, Denis & Hriba, Linda & Dugua, Céline & Tellier, Isabelle. 2011. Un grand corpus oral «disponible»: le corpus d'Orléans 1968-2012. *Ressources linguistiques libres TAL 52*(3): 17-46.

Kahane, Sylvain & Pietrandrea, Paola. 2012. Types d'entassement en français. In *Actes du Congrès Mondial de Linguistique Française* (CMLF 2012), 1809-1828.

Lacheret, Anne & Victorri, Bernard. 2002. La période intonative comme unité d'analyse pour l'étude du français parlé: modélisation prosodique et enjeux linguistiques. *Verbum*, 24/1-2, 55-73.

Lacheret-Dujour, Anne & Kahane, Sylvain & Pietrandrea, Paola. 2019. *Rhapsodie, a prosodic and syntactic treebank of spoken French*. Amsterdam: John Benjamins.

Masini Francesca & Mauri Caterina & Pietrandrea Paola. 2018. List constructions: Towards a unified account. *Italian Journal of Linguistics,* 30(1), 49-94.

Terken, J. 1991. Fundamental Frequency and Perceived Prominence. *Journal of the Acoustical Society of America*, 89, 1768-1776.

't Hart, J., Collier, R. & Cohen, A. 1990. Perceptual Study of Intonation: An Experimental-Phonetic Approach to Speech Melody. Cambridge: Cambridge University Press.

Wagner, Petra & Origlia, Antonio & Avesani, Cinzia & Christodoulides, George & Cutugno, Francesco & D'Imperio, Mariapaola & Escudero, David & Gili Fivela, Barbara & Lacheret, Anne & Ludusan, Bogdan & Moniz, Helena & Ní Chasaide, Ailbhe & Niebuhr, Oliver & Rousier-vercruyssen, Lucie & Simon, Anne-Catherine & Simko, Juraj & Tesser, Fabio & Vainio, Martti. 2015. Different parts of the same elephant : a roadmap to disentangle and connect different perspectives on prosodic prominence. In *Proceedings of the 18th International Congress of Phonetic Sciences*, 0202.1-5.

Wightman, Colin W. 2002. Tobi Or Not Tobi?. In *Proceedings of Speech Prosody*, Aix-en-Provence, France, http://sprosig.isle.illinois.edu/sp2002/.

Yu Chen, Liu & Lacheret-Dujour, Anne & Obin, Nicolas. 2019. Automatic Modelling and Labelling of Speech Prosody: What's new with SLAM+. In Calhoun, Sasha & Escudero, Paola & Tabain, Marija & Warren, Paul (a cura di), *Proceedings of the 19th International Congress of Phonetic Sciences*, Melbourne, Australia 2019. https://assta.org/proceedings/ICPhS2019/.

DOI: 10.17469/O2106SLI000005

EMANUELA CRESTI, LORENZO GREGORI, MASSIMO MONEGLIA, CARLOTA NICOLÁS, ALESSANDRO PANUNZI

The LABLITA Speech Resources

The LABLITA lab of the University of Florence makes available on the web three main spoken corpora: the LABLITA reference corpus of spoken Italian, the IPIC cross-linguistic database of information structure, the C-Or-DiAL Spoken Spanish corpus for teaching Spanish L2. These resources have been annotated following the Language into Act Theory (Cresti 2000) for what regards prosody and its relationship with pragmatics and information structure, and present the speech flow segmented into utterances and information units in correspondence with perceptively relevant prosodic breaks. The LABLITA corpus gives an account of the diaphasic variation of the Italian language spoken in Tuscany according to a detailed corpus design. DB-IPIC, based on a heavily annotated sub-corpus of the LABLITA corpus and comparable Spanish and Brazilian Portuguese corpora, allows the user to retrieve from corpora how information is structured in spontaneous speech, observing how information structure can vary cross-linguistically. C-Or-DiAL proposes to teachers and learners of Spanish L2 a dedicated resource for integrating speech into the learning activities.

Keywords: speech corpora, prosody, information structure, second language acquisition.

1. *Introduction*

In this paper, we present the main multilingual speech resources developed by the LABLITA lab of the DILEF Department of the University of Florence that are presently available on the web:

a. the LABLITA reference corpus of spoken Italian
b. the IPIC database of information structure in spoken Romance corpora (Italian, Brazilian Portuguese, and Spanish)
c. the C-Or-DiAL collection for teaching Spoken Spanish L2.

Paragraphs 2., 3., 4., respectively present the three resources. Despite their different focus, these corpora share a common perspective on the study of spontaneous speech. They refer to the Language into Act

Theory (L-AcT) which can be summarized as follows. The speech activity finds its origins in a mental/affective representation which is the speaker's reaction to an external input (Fagioli 1971[1]). The mental image triggers a linguistic action schema directed to the addressee, according to an embodiment process (Arbib 2012), and is conventionally codified in a pragmatic activity: *illocutionary act*, according to Speech act theory (Austin 1962). Crucially, *prosody* constitutes the interface between the pragmatic activity and the linguistic content (*locutionary act*). This frame results in a pragmatic approach centered on the speaker's activity and focusing on illocutionary force and information structure, both performed through prosodic means (Cresti 2020; Cresti & Moneglia 2018). From this premises follow a set of requirements on how corpora should be compiled and annotated for grounding corpus based linguistic research and applications.

2. *The LABLITA corpus*

2.1 General Presentation

The LABLITA corpus, which is now available online in its first public release, gathers in a single resource, published under the ORFEO platform,[1] a collection of three Italian speech corpora recorded in Tuscany between 1965 and the present days. These resources have already been partially delivered in various occasions, starting from the Corpus of Spoken Italian published by the Accademia della Crusca (Cresti 2000) (approx. 100.000 words). Additional files collected for different purposes by researchers at LABLITA has been joined to this early resource and constitutes the GRIT sub-corpus of the LABLITA collection. Then, the Italian section of the C-ORAL-ROM corpus[2] (approx. 300.000 words each; Cresti & Moneglia 2005) has been added. The above sub-corpora are integrated by the Stammerjohann corpus (Stammerjohan 1971), whose original acoustic source was made avail-

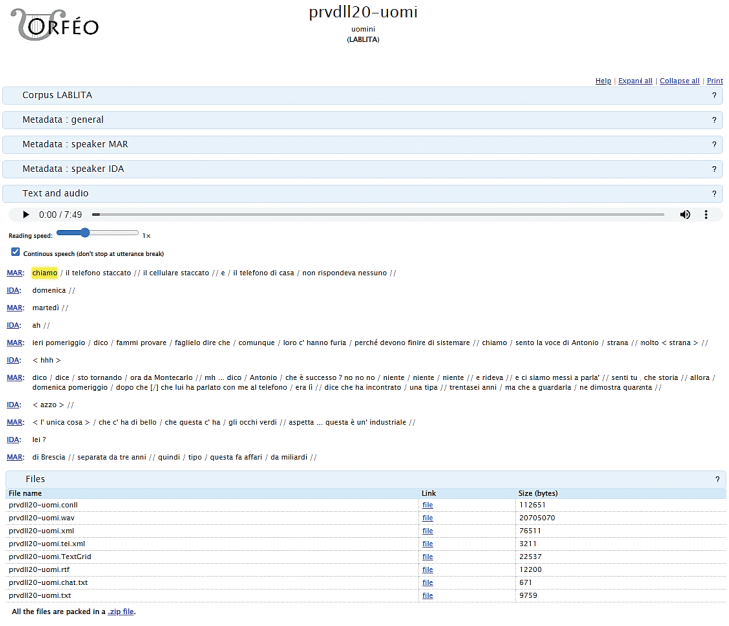
[1] We wish to thank Jeanne-Marie Debaisieux for making the Orfeo platform available to us.

[2] C-ORAL-ROM was achieved in a Project co-ordinated by LABLITA within in FP/5 of the EU. A choice of this section has been also published as a resource to be compared with the written variety for teaching Corpus Linguistic techniques (Cresti & Panunzi 2013).

able to us by the author within the FIRB project *Archivi dell'italiano orale in diacronia*. The Stammerjohan corpus, collected in Florence in 1965, is the first corpus of spoken Italian. It has been derived from 40 hours of original recordings, sampled according to the LABLITA corpus design strategy (see below). The corpus was already distributed online together with a sampling of Italian spoken in Florence during the 90' for comparative analysis of language change (approx. 100.000 words each) (Moneglia & Scarano 2008; Moneglia & Panunzi 2022).

Nowadays the LABLITA corpus comprehends approx. 700.000 token, is open and continuously updated. The acoustic signal is transcribed in CHAT/LABLITA format, which enriches the traditional CHAT transcripts with the annotation of *terminal* and *non-terminal prosodic breaks* (Cresti & Moneglia 1997). Transcripts are aligned to the acoustic source by *terminated sequences* (see below). The corpora are delivered to the users online and for downloading as a collection of files comprising Metadata, Transcription and PoS tagging in TXT fies, aliment in xml files (Winpitch and Praat formats). Figure 1 shows how sessions are presented.

Figure 1 - *Screenshot of a session of the LABLITA Corpus in Orfeo*



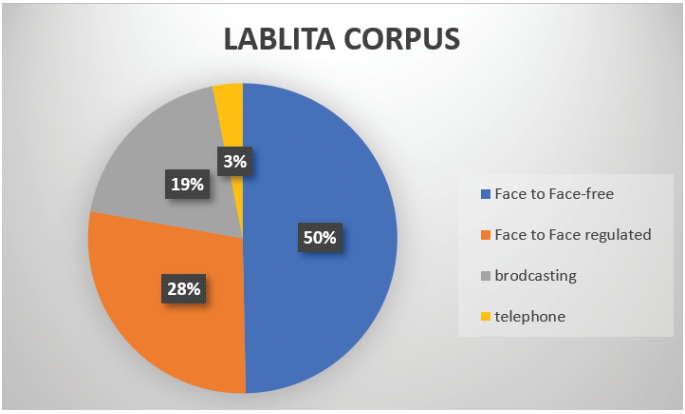
2.2 Sampling criteria and representativeness

The overall purpose of the Corpus is to provide an adequate data set for the study of "spontaneous speech", documenting its main variations. We consider "spontaneous speech" any speech production conceived and performed within a direct interactive relationship between speakers, in which language conception and language performance occur simultaneously.

The LABLITA corpus collects 422 samples of continuous speech (between 1500 and 3000 words). The corpus is proposed as a reference corpus of spoken Italian. To this end it is intended to provide a significant representation of the linguistic choices (at the lexical, syntactic, prosodic, and pragmatic levels) that characterize spoken Italian. Its specific goal is to be representative of the diaphasic variation, which is reflected in the corpus design.[3]

The variation of the corpus is limited from the diatopic point of view to speakers located in Florence and Tuscany. Diachronic variation is documented since 1965. Diastratic variation is not balanced but includes more than 1000 different speakers whose main metadata are delivered (education, sex, age, profession, and origin). Diamesic variation includes, in addition to face-to-face interactions, also a telephone sampling and a collection of broadcasting.[4] Figure 2 illustrates that more then 2/3 of the corpus document the direct interactive language usage (face to face).

Figure 2 - *Distribution of samples by Channel in the LABLITA corpus*



[3] Gestural aspects are also studied at LABLITA on the basis of new collections of audio-video recordings (Cantalini & Moneglia 2020; Cantalini, in this volume).
[4] Provided by Teche-RAI for the C-ORAL-ROM corpus.

Face-to-face speech constitutes the most relevant part of the collection, and is sampled on three dependent levels, which constraint the intersubjective interaction governing the speech activities:

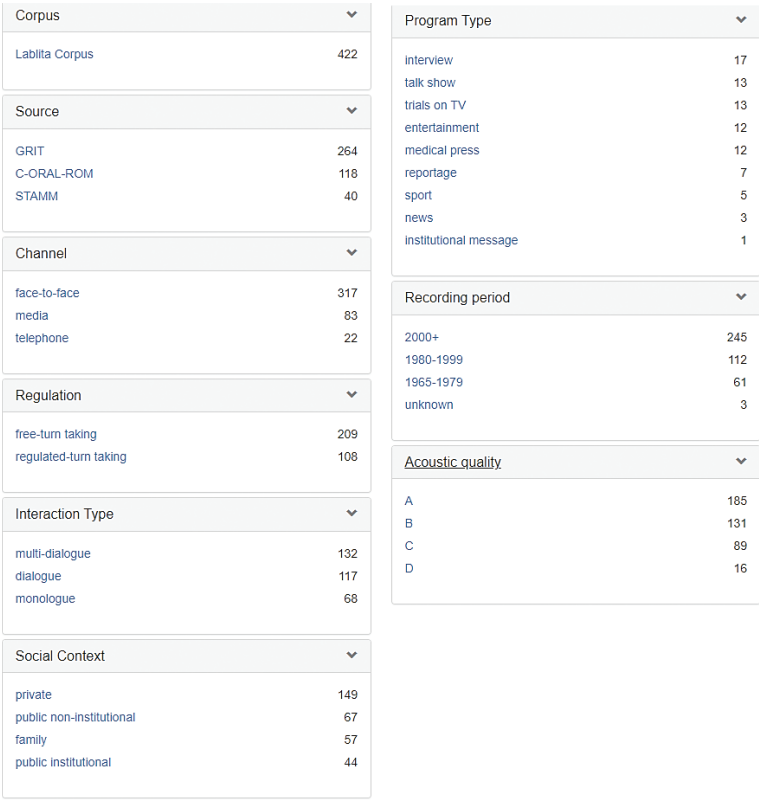
1. the top level distinguishes contexts in which the turn-taking is free or on the contrary, undergoes explicit or implicit rules. This is a way of selecting samples characterized by supervised and formal speech, although no judgment regarding the linguistic style has been given. Free turn-taking contexts (informal) record 2/3 of samples, so ensuring representativeness to the basic spontaneous speech usage which is informal and unsupervised.
2. the second level regards the relationships between speakers that are determined by the social context in which the interaction takes place. The corpus distinguishes four contexts which can in principle influence the affective relation among speakers, and, therefore, their linguistic behaviour: a) family, b) private life, c) public life, d) public life in institutional contexts. This social variation in the corpus design is intended also to give the probability of occurrence to many different possible activities of everyday life in which the linguistic interaction takes place, varying for task, relevance, and goals. The overall strategy is to privilege, from a quantitative point of view, sampling in family and private life.
3. For each of the above social context three types of linguistic interactions are documented: a) monologue;[5] b) dialogue between two speakers; c) multi-dialogue.

   The broadcasting contains a variation of formats, covering a large typology of the programs which occupy the public space on the TV, which despite the emergence of the new media, still represents a context where speech has a high impact in contemporary society.

Figure 3 illustrates the fields of the Corpus design as they are presented in the Orfeo platform. Corpus design fields can be the object of selection by the user.

[5] Many sessions that have been classified as "monologue", record the participation of more than one speaker, but in this case, there is always a dominant actor occupying the turn-taking with a long stretch of speech.

Figure 3 - *Number of samples of the LABLITA Corpus within the corpus design structure*



The user can access the corpus and make searches restricting to the type of context of his interest, compare the language usage of a given context with others, or according to the type of research, can select only files of a given period or acoustic quality.[6]

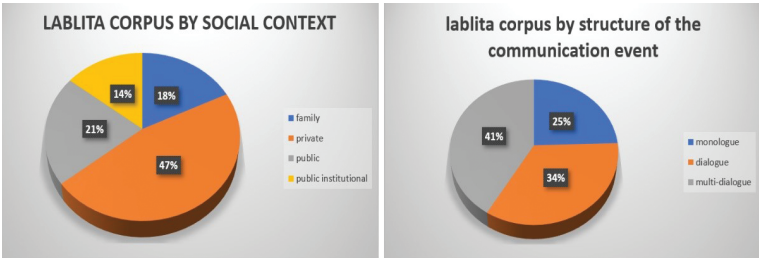
The representativeness of the corpus follows from the intersection of the above parameters. Corpus sampling gives a reasonable proportion to each field in the variation according to the probability of occurrence in the everyday life of the population (Maruyama, in this volume).

[6] Although speaker metadata are available, they cannot be used for selecting speakers of a certain type in the present release.

As regards the social domain of the interaction, half of the face-to-face sub-corpus records events occurring in the private life and about 20% in family interactions, so giving peer relationships larger space with respect to the public and institutional contexts, which occur less frequently and in which the language activity might be more controlled.

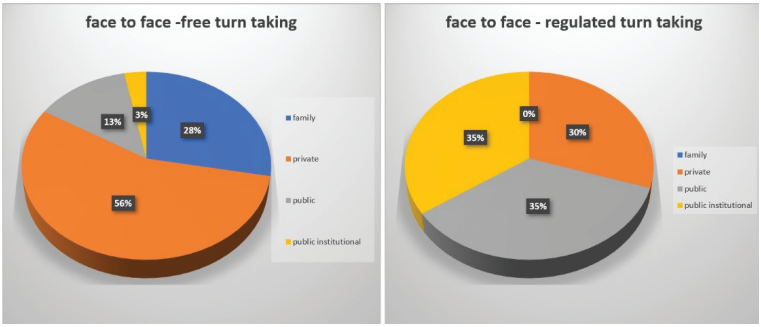
The parameter "structure of the communicative event" allows recording of dialogues between two speakers and multi-dialogues where many speakers interact within a given situation. These are sampled in a similar percentage and cover most of the corpus, being by far the most frequent contexts of usage. However, the corpus also gives a large space (1/4 of the corpus) to monologic performances, so also documenting the linguistic structures that can have some probability of occurrence only in narratives, conferences, explanations, political speech, preaching, etc.

Figure 4 - Social *and* Structural *variations of speech events in the LABLITA corpus*



Finally, when considered face to the social domain, the distinction between contexts in which the spoken interaction foresees free or regulated turn taking shows interesting feature of the Italian socio-linguistic situation. In family life the turn taking is always free and regulation does not occur, while in public institutional context regulation is almost a requirement.

Figure 5 - *The intersection between* Regulation of turn taking *and* social context *in the LABLITA corpus*



The representativeness of the universe of spontaneous speech can be ensured when in principle all event types have at least some probabilities of occurrence, and for this reason, each parameter in the corpus design should be represented by speech events of many different types (Cresti & Panunzi 2013; Maruyama, in this volume). This objective is reached as a function of an implementation strategy that on one side tries to fulfil the various parameters of the corpus design in adequate proportion and on the other to avoid over-representation of a single event type, ensuring variation of task, goal, and genre in the samples gathered within every single field. The following is a picture of the event type variation which can be found in the LABLITA corpus.

When filling the *public-institutional* parameter we sampled monologues, dialogues, and multi dialogues in various contexts of the public life: in administrative and political contexts, in religious life, in cultural life, in economic life. For instance, sampling regards meetings in public companies, interventions at the district council, sermons during a Sunday Mass, harangue, and interrogations by the Public Prosecutor in the trial, university lectures, papers in a conference, professional explanations, narratives by children or teachers and oral test at school, rallies or political meetings during the election campaign, and many other contexts.

When considering *family life*, the corpus records typical monologues or dialogues regarding whatever topic occurring in this context, for instance: memories of old family members, life stories, conversations at dinner, interactions while preparing food, chat between relatives, planning of feast events, hated discussions, problems with

the neighbour, reproach to children, play sessions with children, storytelling to children, plans for vacations or future activities and explanations on how to manage the house affairs, driving school to young relatives, chat in the car, discussions regarding shows, politics or while browsing a photo album, criticisms to teen-agers, etc.

When looking to the *private life* the corpus also comprises a large variety of context and topics: honeymoon stories, travel stories and adventures said to friends, interviews to many professionals on their work (doctor, ceramist, knitter, projectionist, professors, actors, nurse, railway worker, soldiers, partisans, politicians, film directors, etc), psychological interview, professional explanations in various trades, chat at lunch with colleagues, chat among bartenders, instructions to domestic workers, sail of financial products by promoters, professional interactions between mechanics, plumbers, masons, electricians, chefs etc., private lessons, work discussions with colleagues in the office, dialogues among musicians of a band, chat with the beautician while doing a depilation, with the hairdresser while she is styling, discussions among friends on various topics, dinners with friend, chat among teenagers, plans for participating to a competition, chat while driving, in the train, in the park.

There is no internal balance among context types. The collection strategy is to get random samples according to the recording opportunities.

2.3 Corpus annotation

The overall size of the corpus is conspicuous but not enormous (about 700,000 words), its main added value is that transcripts are aligned to the acoustic source by pragmatic units (utterance). The utterance is the linguistic counterpart of a speech act (Austin 1962); i.e. the minimal linguistic entity that can be pragmatically interpreted (Biber *et al.* 1999; Cresti 2000) and the utterance boundaries in the speech flow identify the domain of relevance of linguistic relations (*reference units* according to Izre'el *et al.* 2020). For this reason, the relevance of the corpus for linguistic studies can be evaluated not only in terms of the number or words but also considering the number of pragmatic units achieved by speakers in their language performance (105,000 reference units). Within this set, 86,000 reference units belongs to face-to-face interactions. According to Language into Act Theory (L-AcT,

Cresti 2000), speech comprehends two types of reference units: *utterance*, as the counterpart of a simple speech act, representing approximately 90% of entries, and *stanza*, the linguistic counterpart of a flow of thought (Chafe 1994), that can be composed of many short utterances linked by prosody the one to the other. Stanzas are most frequent in monologues and when the spoken interaction is formal (Saccone 2020).

The segmentation of speech into reference units is not a trivial matter (Raso et al. 2020). The LABLITA corpus adopts the L-AcT framework which stresses the strict correlation between pragmatic activities and prosodic cues. More specifically L-AcT provides an operative method for speech segmentation based on the perception of prosodic cues ('t Hart *et al.* 1990). The continuous flow of speech is primarily segmented by perceptually relevant prosodic interruptions to which competent speakers of a language assign *terminal value* (Izre'el *et al.* 2020). The identification of the terminated sequences (TS) allows the parsing of speech into autonomous speech activities which can be a reasonable object of linguistic analysis.[7] Terminated sequences can be in turn segmented into prosodic units (PU) by prosodic breaks, which are perceived with a *non-terminal* value. PUs corresponds in the L-AcT framework to *information units* (IUs) (Cresti 2000; Cresti & Moneglia 2018; Moneglia & Raso 2014).

For instance, the following is the word-by-word transcription of the first dialogic turn of the conversation among girls presented in Figure 1:

*MAR: chiamo il telefono staccato il cellulare staccato e il telefono di casa non rispondeva nessuno

Considering the possible syntactic relations among words in this sequence, the structure turns out highly undetermined. Some of the possible interpretations are suggested below through punctuation and capital letters: it may be a unique structure where constituents stand in asyndetic coordination relation (a), it can be parsed in three utterances (b), or in five utterances, as in (c):

[7] The agreement regarding perception of terminal breaks has been evaluated in C-ORAL-ROM from the L-AcT perspective (Danieli et al 2004) and also independently of this theoretical background (Panunzi et al 2020).

(a) Chiamo, il telefono staccato, il cellulare staccato e il telefono di casa non rispondeva nessuno.
(b) Chiamo. Il telefono staccato. Il cellulare staccato e il telefono di casa non rispondeva nessuno.
(c) Chiamo il telefono: staccato! E il cellulare: staccato! E il telefono di casa non rispondeva nessuno.

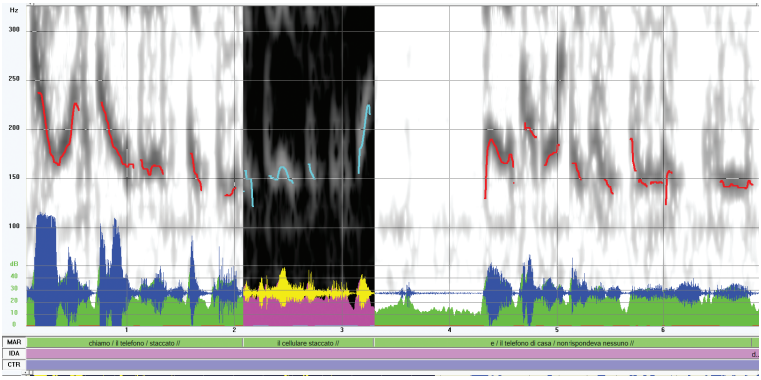
However, when considering prosody, the structure is not underdetermined, since the perception of terminal prosodic breaks ("//") defines the boundaries of the sequence, and restricts the possible interpretations of three specific speech acts corresponding to three TSs:

*MAR: chiamo / il telefono / staccato // il cellulare staccato // e il telefono di casa / non rispondeva nessuno //

The annotation files of the LABLITA corpus are delivered in Praat and Winpitch files where the transcript is shown aligned to the acoustic source. The user can verify that each stretch of speech marked with a terminal break is an independent utterance and can study the linguistic relations established within each of them and their prosodic correlations. For instance, the second utterance of the turn is a *nominal utterance* performed in one PU corresponding to one IU, while the third utterance is an *anacoluthon* performed in two PUs conveying a *Topic* / *Comment* information pattern.

In Figure 6 the text / sound alignment is displayed in a screenshot of the Winpitch software, which distributes the transcription in independent layers for each speaker. The program displays F_0 track, spectrogram, and intensity of each utterance, ensuring full and direct exploitation of corpus data for phonetic and linguistic analysis.

Figure 6 - *Text sound alignment in Winpitch*



3. *DB-IPIC: a Database for Information Patterning Interlinguistic Comparison*

DB-IPIC is an online resource, that allows to browse and perform complex searches on spoken corpora annotated following the L-AcT principles. It is designed to host corpora with prosodic segmentation and information structure and allows to study linear relations between PUs and IUs within each reference unit (TS) type marked by a terminal prosodic break (*utterances* and *stanzas*). In addition to it, DB-IPIC supports token-level annotation of parts of speech and lemmas.

The resource is composed of an XML database, and a web interface for corpus querying. The database contains speech transcriptions, each one enriched with metadata and all the levels of annotation; data related to each session is embedded in a single XML file, according to a specific XML model, exemplified by the excerpt in Figure 7:

Figure 7 - *XML model for linguistic annotation in DB-IPIC*

```
<turn speak="EDO">
        <term_seq num="1" type="utt">
          <tone_unit inf="COM">
                    <word lemma="guardare" pos="VER:fin">guarda</word>
                    <word lemma="chi" pos="WH">chi</word>
                    <word lemma="ci" pos="ADV">c'</word>
                    <word lemma="essere" pos="VER:fin">è</word>
                    <break type="nonterminal">/</break>
          </tone_unit>
          <tone_unit inf="ALL">
                    <word lemma="nonna" pos="NOUN">nonna</word>
                    <break type="terminal">//</break>
          </tone_unit>
        </term_seq>
</turn>
```

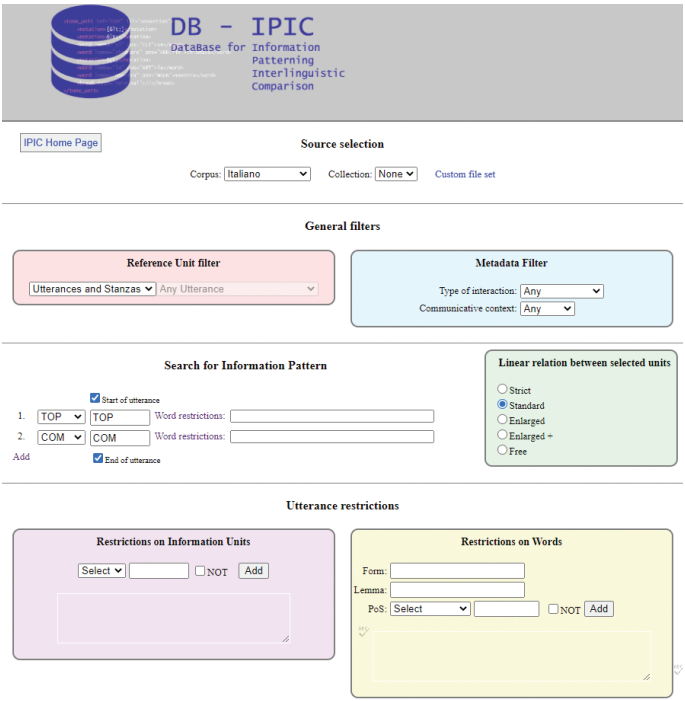
DB-IPIC online interface (Figure 8) is a PHP web application that provides a user-friendly way to extract information from the database. With this tool it's possible to query a corpus at different levels, according to the logical structure of the data set. DB-IPIC can operate on five levels:

1. data source: it is possible to query a whole corpus or to specify a subset of sessions; different corpora can be managed in DB-IPIC
2. metadata: sessions can be filtered by their properties, specifying the communicative context (familiar or public) and the interaction type (monologue, dialogue, or conversation)
3. informational patterns: the user can select the TSs by specifying their IU pattern
4. information units: it's possible to search TSs containing or not containing specific IUs independently of their informational pattern
5. words: users can refine their search by including or excluding words with a specific form, PoS, or lemma.

DB-IPIC currently contains an Italian corpus of 74 texts (124,735 total words) chosen from the Informal section of Italian C-ORAL-ROM (Cresti & Moneglia 2005), and 3 small comparable corpora (mini-corpora) of Italian (IT), Spanish (ES) and Brazilian Portuguese (BP). These last two are derived from C-Or-DiAL (see

§ 4) and from C-ORAL-BRASIL (Raso & Mello 2010), while the small Italian is a subset of the main Italian corpus. These mini-corpora have a similar size (30,000 to 40,000 words) and the same design; all of them are manually annotated with prosodic and information structure annotation according to the L-AcT tagset and definitions of IU types (Moneglia & Raso 2014)[8]. This allows the use of DB-IPIC for studying how information is structured by prosody in each language corpus and to perform cross-linguistic comparisons between spoken Italian, Spanish and Brazilian Portuguese. For instance, Figure 8 is a screenshot of the IPIC query interface presenting the search for *Topic* / *Comment* utterances in the Italian sub-corpus.

Figure 8 - *DB-IPIC search interface*



Query results page (Figure 9) displays the list of utterances that match the search criteria. In the corpus 687 TS (on 5117 utterances) show

[8] PoS-tagging and lemmatization are derived through automatic tools.

a *Topic* / *Comment* structure. Each line contains utterances transcription, prosodic and information structure annotation, PoS-tagging and lemmatization. Moreover, a direct access to audio source is available and downloadable in MP3 format. Results can be also exported in CSV format to be analysed through a spreadsheet software.

Figure 9 - *Screenshot of data returned by DB-IPIC*

Beside specific queries on information patterning and morpho-lexical fillings, DB-IPIC can be used for deriving comparative distributional data with respect to the represented languages. One of the main comparative data is reported in Table 1., which contains the percentage of the different type of reference units of speech according to L-AcT theory (Utterances Vs Stanzas) in the three mini-corpora. Data show that the percentage of Utterances and Stanzas is almost constant in the three languages, given that stanzas are about 9% of the total TSs.

Table 1 - *Utterance vs Stanzas in IT, ES and BP mini-corpora*

	IT	%	ES	%	BP	%
Utterances	5117	90,36%	5898	91,51%	5046	91,55%
Stanzas	546	9,64%	547	8,49%	466	8,45%
Total units	5663		6445		5512	

If we consider only the Utterances, we can derive data regarding the complexity of the information patterns. In Table 2, for instance, we distinguished Simple Utterances (see examples 1a-1c), in which there is only the Comment unit, and Complex Utterances (see examples 2a-2c), in which other information units occur.

(1a) *ART: le quattro componenti son queste //COM (ifamdl04, 47)
the four components are these //

(1b) *PAC: no sé muy bien que hay //COM (efamdl04, 43)
I don't know very well what is there //

(1c) *FLA: seu dinheiro tá caindo hhh //COM (bfamdl01, 510)
your money is falling [out of your pocket]' //

(2a) *MAR: più di così /TOP 'un arriva /COM caro //ALL (ifamdl19, 339)
more than that / it doesn't arrive / dear //

(2b) *PIL: no /PHA es que ese /TOP era una estrategia de la defensa //COM
no / it is that this / was a defence strategy

(2c) *BEL: uhn /TMT talvez na parte maior /COM não //PHA (bfamdl02, 194)
hm / maybe in the bigger part / no //

Table 2. shows that, with respect to this measure, Italian and Spanish present similar values, while Brazilian Portuguese records a higher percentage of Simple Utterances. This can be interpreted as a different overall strategy in structuring the spoken production, as already observed in specific comparative studies that exploited the DB-IPIC data (Panunzi & Mittmann 2014).

Table 2 - *Simple vs Complex Utterance in IT, ES and BP mini-corpora*

	IT	%	ES	%	BP	%
Simple Utterances	3491	68,22%	4073	69,06%	3840	76,10%
Complex Utterances	1394	27,24%	1627	27,59%	1089	21,58%
Uncomplete Utterances[9]	232	4,53%	198	3,36%	117	2,32%
Total Utterances	5117		5898		5046	

4. *C-Or-DiAL*

C-Or-DiAL is a spoken corpus of Spanish created for teaching purposes and published online (Nicolás Martínez 2012). The aim of

[9] The class of Uncomplete Utterance contains mostly interrupted sequences in which there is not any Comment unit.

C-Or-DiAL is to provide learners of Spanish L2 with a didactic resource freely accessible on the web that allows the study of the language through an explicit reflection on its spoken variety.

The corpus is integrated with linguistic annotation and metadata, and is delivered as a didactic material. Teachers can select texts by difficulty level, according to learning needs of their classes. The corpus includes 240 short audio recordings (approximately 3 minutes each) with the corresponding transcripts for 120,000 words and is presented in Spanish through a user-friendly web interface.

Figure 10 shows in a screenshot the file index of C-Or-DiAL. From this page, the user can select the sessions of his interest and access the audio and text files. Each field contains higher-level metadata orienting the user to select the sessions to be used in a class.

Figure 10 - *Index of the C-Or-DiAL corpus*

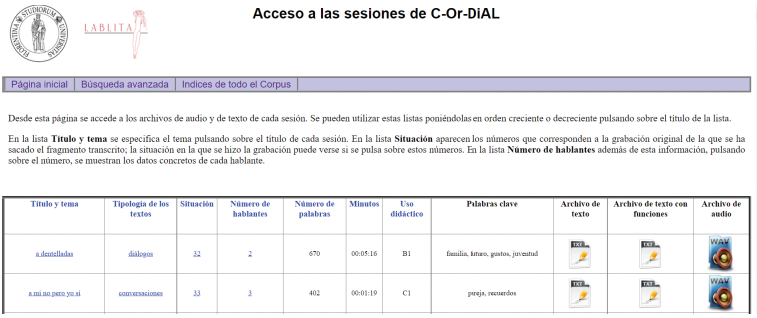


Figure 11 - *Higher-levels metadata of C-Or-DiAL files*

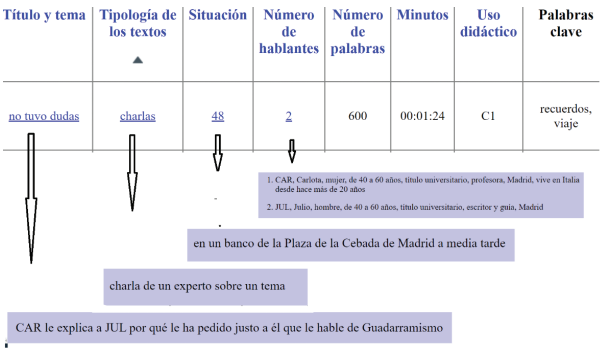
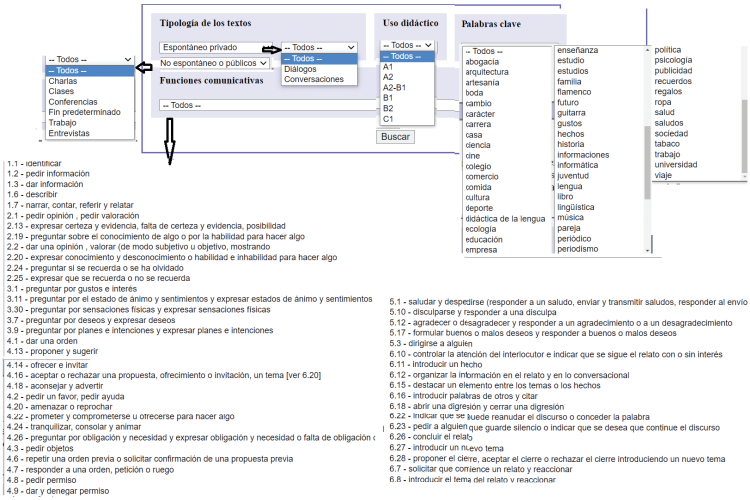


Figure 11 illustrates in detail the kind of information available for each session. Each recording has a *Title*, a *Genre* of text (chat, conference, dialogue, multi-dialogue, interview, talks at work), a *Situation* where the communicative event occurs, and *Speakers* metadata. The last four fields show the length and keywords of the recording, and the suggested level according to the European frame.

An advanced search interface is also available to teachers and learners to make decisions on what might be the better oral source for their activities. To this end, the user can select any of the fields present on the page. The system will return all sessions that have the required characteristics. Figure 12 illustrates the parameters governing the possible choices. For instance, the choice can be oriented to *informal private* sessions, which allow selecting among *dialogues* or *multi-dialogues,* alternatively to *formal public* speech, where a good lot of professionals have been recorded. In this case, the user can select sessions among: Speeches, Classes, Conferences, Interviews, etc. The user can also explore the database by restricting his search by level of difficulty of the text or selecting a topic through the keywords.

Moreover, the main annotation of the corpus, which is specifically significant for learning activities, regards the communicative functions which are instantiated in the spoken interactions. These functions (listed in Figure 12) have been identified and annotated in the transcripts according to the repertory published within the *Plan Curricular Cervantes*, which is the main background framework for Spanish L2.

Figure 12 - *Advanced search interface of C-Or-DiAL*



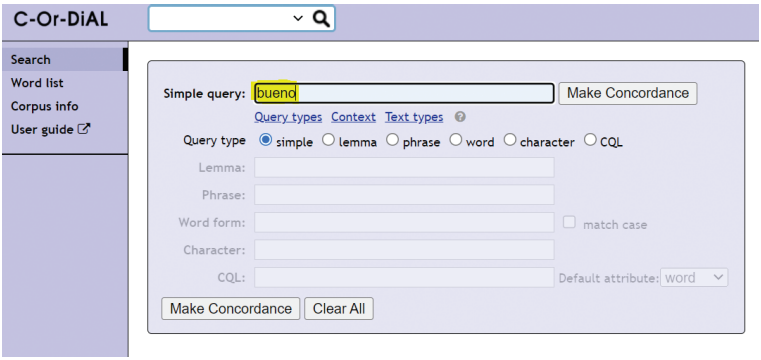
Transcripts are in the CHAT/LABLITA format and bear the systematic annotation of the terminal and non-terminal prosodic breaks marking the utterance boundaries. Transcripts are headed by the full set of session metadata. Files are in txt format; they can be downloaded also in a richer version, comprising the annotation of communicative functions codes, as in the stretch of dialogue in Figure 13:

Figure 13 - *Transcripts in C-Or-DiAL*

```
*BRI: ¿ 1.2 pero ponemos Madrid Cuneo o Madrid el destino que [/] qué queréis ver el xxx 1.2 ?
*PED: no tú [/] tú / 3.9 para que nos vuelva [/] para que nos vuelva a dar esto mismo y ahora vamos por el mapa y lo vamos viendo ya 3.9 ¿ entiendes ? 4.1
pones Madrid / dirección Purchena / es lo mismo / para que sea exactamente / para que nos salga lo mismo que esto / Purchena 19 Madrid 4.1 //
*CAR: 2.2 es que Arles / < que es más bonito >2.2 /
*PED: 4.1 < y abajo > pones Cuneo / Italia 4.1 // < xxx >
*CAR:/ 1.3 < lo más bonito > de esta zona Chavela es la [//] es el paisaje de la Camargue 1.3 //
*PED: 1.3 no Cuneo es abajo / Cuneo [/] Cuneo es abajo / eso es la &esta / la dirección nos va a llevar a Cuneo centro pero Cuneo hhh 1.3 ...
*CAR: 4.13 Cuneo 4.13 //
```

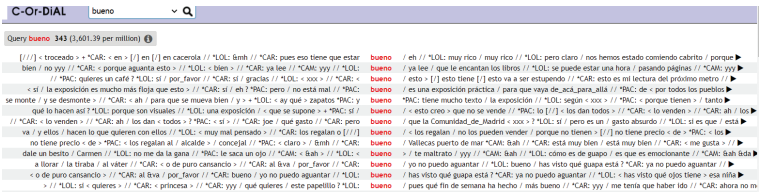
Although the alignment files of the corpus were not delivered, the acoustic source of each utterance is available to the user through a concordances search, exploiting standard corpus linguistics methods. The corpus has been implemented within the NoSketchEngine interface, where keywords can be searched through a standard form (Figure 14). Occurrences are returned (Figure 15) and can be listened in the context of each utterance in which they occur. The user can search for tokens, lemmas, or phrases in all corpus or in sub-corpora.

Figure 14 - *Query interface in C-Or-DiAL*



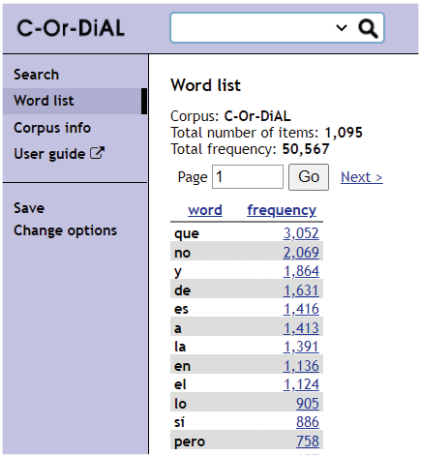
He can appreciate for instance what is the acoustic and distributional difference between one expression used as a Discourse Marker, (*bueno*) in the second row of the screenshot in Figure 15, and the same expression used as an Adjective (in the first row).

Figure 15 - *Keyword in context with the audio file of the utterance in which they occur*



The user can also generate frequency lists by word, lemmas, and PoS (Figure 16).

Figure 16 - *Screenshot of the Frequency lists*



C-Or-DiAL can contribute to the development of teaching materials and practices in several ways: it may implement classical teaching methods with examples derived from real speech but can also promote the development of new teaching method, where speech is at the core of learning activities (Nicolás Martínez et al. 2016). For instance, after careful listening to the audio and the control of comprehension with the help of transcription we can move on to listening and speaking exercises, and only afterward do we pass to the study of vocabulary and grammar. The practice of transcribing small audio fragments from C-Or-DiAL is fruitful at the A1-A2 levels. Transcription with tags can be done at intermediate levels, to learn how to analyze speech, focusing on the relevance of prosodic patterning, and dialogic strategies. The specific lexicon used in spontaneous speech can be the object of focused study with students of medium and high level.

References

Arbib, Michael. 2012. *How the brain got language*. Oxford: Oxford University Press.

Austin, John. 1962. *How to do things with words*. Oxford, Oxford University Press.

Biber, Douglas & Johansson, Stig & Leech, Goffrey & Conrad, Susan & Finegan, Edward. 1999. *The Longman Grammar of Spoken and Written English*. London and New York: Longman.

Cantalini, Giorgina. 2022. Corpus multimodale annotato per lo studio della gestualità co-verbale nel «parlato-parlato» e nel «parlato-recitato». (in this volume)

Cantalini, Giorgina & Massimo Moneglia. 2020. The Annotation of Gesture and Gesture / Prosody Syncronization in Multimodal Speech Corpora. *JOSS Journal of Speech Science*, 9. 1-24.

Chafe, Wallace. 1994. *Discourse, Consciousness, and Time: The Flow and Displacement of Conscious Experience in Speaking and Writing*. Chicago: The University of Chicago Press.

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

Cresti, Emanuela. 2020. The pragmatic analysis of speech and its illocutionary classification according to the Language into Act Theory. In Izre´el, Shlomo & Mello, Heliana & Panunzi, Alessandro & Raso, Tommaso (eds), *In search of basic units of spoken language: A corpus-driven approach*, 181-219. Amsterdam: John Benjamins.

Cresti, Emanuela & Moneglia, Massimo. 1997. L'intonazione e i criteri di trascrizione del parlato adulto e infantile. In Bortolini, Umberta & Pizzuto, Elena (a cura di), *Il progetto CHILDES: strumenti per l'analisi del linguaggio parlato*, vol. II, 57-90. Pisa: Edizioni del Cerro.

Cresti, Emanuela & Moneglia, Massimo (eds). 2005. *C-ORAL-ROM. Integrated reference corpora for spoken romance languages*. Amsterdam: John Benjamins.

Cresti, Emanuela & Moneglia, Massimo. 2018. The illocutionary basis of Information Structure. Language into Act Theory. In Adamou, Evangelia & Haude, Katharina & Vanhove, Martine (eds.), *Information structure in lesser-described languages: Studies in prosody and syntax*, 359-401. Amsterdam: Benjamins.

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

Danieli, Morena & Garrido, Juan Maria & Moneglia, Massimo & Panizza, Andrea & Quazza, Sivia & Swerts, Marc. 2004. Evaluation of Consensus on the Annotation of Prosodic Breaks in the Romance Corpus of Spontaneous Speech "C-ORAL-ROM". In Lino, Maria Teresa & Xavier, Maria Francisca & Ferreira, Fátima & Costa, Rute & Silva, Raquel (eds.), *Proceedings of the 4th LREC Conference*, vol. 4, 1513-1516, Paris.

Fagioli, Massimo. 1972[1]. *Istinto di morte e conoscenza*. Roma: L'Asino d'oro. (Trad. Eng. *Instinct of death and knowledge*, L'Asino d'oro 2021).

't Hart, Johan & Collier, René & Cohen, Antonie. 1990. *A Perceptual Study on Intonation. An Experimental Approach to Speech Melody*. Cambridge: Cambridge University Press

Izre'el, Shlomo & Mello, Heliana & Panunzi, Alessandro & Raso, Tommaso (eds.). 2020. *In Search of Basic Units of Spoken Language: A Corpus-Driven Approach*. Amsterdam: Benjamins.

MacWhinney, Brian.2000. *The CHILDES Project: Tools for Analyzing Talk. 3rd Edition*. Mahwah, NJ: Lawrence Erlbaum Associates.

Maruyama, Takehiko. 2022. Designs and Analyses of Japanese Speech Corpora. (in this volume)

Moneglia, Massimo. 2005. The C-Oral-Rom resource. In Cresti, Emanuela & Moneglia, Massimo (eds), *C-ORAL-ROM. Integrated Reference Corpora for Spoken Romance Languages*, 1-70. Amsterdam: Benjamins.

Moneglia, Massimo & Raso, Tommaso. 2014. Notes on Language into Act Theory (L-AcT). In Raso, Tommaso & Mello, Heliana (eds), *Spoken corpora and linguistic studies*, 468- 495. Amsterdam: Benjamins.

Moneglia, Massimo & Scarano Antonietta. 2008. Il Corpus Stammerjohann. Il primo corpus di italiano parlato, in rete nella base dati di LABLITA. In Pettorino, Massimo (ed.), *Atti del Congresso Internazionale "La comunicazione Parlata", Napoli 2006*, 1699-1739. Napoli: Liguori.

Moneglia, Massimo & Panunzi, Alessando. 2022. Micro-Diachronic Corpora for Measuring the Lexical Change of Spontaneous Speech in Florence Compared to Standard Italian. *Langages*, 226. 41-54.

Nicolás Martínez, Carlota. 2012. *C-Or-DiAL Corpus Oral Didáctico Anotado Lingüísticamente*. Madrid: Liceus.

Nicolás Martínez, Carlota & Hernández Toribio, María Isabel. 2016. *Del oído al habla*. Barcelona: Octaedro.

Panunzi, Alessandro & Gregori, Lorenzo. 2012. DB-IPIC. An XML database for the representation of information structure in spoken language. In Panunzi, Alessandro & Raso, Tommaso & Mello, Heliana (eds), *Pragmatics and prosody. Illocution, modality, attitude, information patterning and speech annotation*, 121-127. Firenze: Firenze University Press.

Panunzi, Alessandro & Mittmann, Maryualê. 2014. The IPIC resource and a cross-linguistic analysis of information structure in Italian and Brazilian Portuguese. In Raso, Tommaso & Mello, Heliana (eds.), *Spoken Corpora and Linguistic Studies*, 129-151. Amsterdam: Benjamins.

Panunzi, Alessandro & Gregori, Lorenzo & Rocha, Bruno. 2020. Comparing annotations for the prosodic segmentation of spontaneous speech: Focus on reference units. In Izre'el, Shlomo & Mello, Heliana & Panunzi, Alessandro & Raso, Tommaso (eds.), *Search of Basic Units of Spoken Language. A corpus-driven approach*, 403-431. Amsterdam: Benjamins.

Raso, Tommaso & Mello, Heliana. 2012. *C-ORAL-Brazil I, Corpus de referência do português brasileiro falado informal*. Belo Horizonte: EDITORAufmg.

Raso, Tommaso & Teixeira, Bárbara & Barbosa, Plinio. 2020. Modelling Automatic Detection of Prosodic Boundaries for Brazilian Portuguese Spontaneous Speech. *JOSS Journal of Speech Science*, 9. 105-128.

Saccone, Valentina. 2020. La Stanza nella Teoria della Lingua in Atto: Un'analisi sintattica. *CHIMERA: Revista De Corpus De Lenguas Romances Y Estudios Lingüísticos*, *7*. 55–68.

Stammerjohann, Harro. 1970. Strukturen der Rede. Beobachtungen an der Umgangssprache von Florenz. *Studi di Filologia Italiana* XXVIII. 295-397.

C-ORDIAL< http://lablita.it/app/C-Or-DiAL/>

C-ORDIAL <http://lablita.it/C-Or-DiAL/run.cgi/first_form>

IPIC: <http://www.lablita.it/app/dbipic/>

LABLITA Corpus, http://corpus.lablita.it/>

Praat <http://www.fon.hum.uva.nl/praat/>

sketchengiNE <https://www.sketchengine.eu/>

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

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

DOI: 10.17469/O2106SLI000006

CATERINA MAURI, SILVIA BALLARÉ, EUGENIO GORIA, MASSIMO CERRUTI

Il corpus KIParla

In questo articolo presentiamo le principali caratteristiche del CORPUS KIParla, una nuova risorsa per lo studio dell'italiano parlato, liberamente accessibile al sito www.kiparla.it. Il corpus è stato progettato per essere gratuitamente consultato attraverso l'interfaccia NoSketch Engine e per essere espanso nel tempo tramite l'aggiunta di nuovi moduli. Il corpus KIParla fornisce l'accesso a una vasta gamma di metadati che caratterizzano sia i partecipanti che le interazioni, utilizzabili come filtri di ricerca. Al momento il KIParla consiste di due moduli (KIP e ParlaTO), che permettono di effettuare ricerche sulla variazione diafasica, diatopica e diastratica dell'italiano contemporaneo.

Parole chiave: corpus, italiano parlato, sociolinguistica.

1. *Introduzione: il corpus KIParla*

Il corpus KIParla è una risorsa elettronica per lo studio dell'italiano parlato di recente pubblicazione, frutto della collaborazione tra l'Università di Torino e l'Università di Bologna, e aperto a futuri contributi provenienti da altri gruppi di ricerca. Il video della DEMO è accessibile online: https://underline.io/lecture/33036-d12---il-corpus-kiparla

Il KIParla si distingue da altre risorse attualmente disponibili per lo studio dell'italiano parlato per alcune proprietà; fra le altre, la possibilità di avvalersi di una serie di metadati relativi alle caratteristiche socio-demografiche dei parlanti e al tipo di interazione in cui essi sono coinvolti, e l'opportunità di consultare i dati sia in formato audio sia in formato testuale. La risorsa è costruita, inoltre, in maniera tale da rendere possibili futuri ampliamenti, sotto forma di nuovi moduli parzialmente indipendenti ma che condividano uno stesso nucleo di metadati e lo stesso sistema di raccolta e gestione dei dati. Infine, il KIParla è una risorsa di libero accesso che si avvale della piattaforma di interrogazione NoSketch Engine (Rychlý 2007).

2. *Progettazione del corpus*

Il corpus KIParla è costituito da materiali linguistici registrati, fino ad ora, nelle città di Bologna e di Torino. Le due città presentano una situazione sociolinguistica per certi versi analoga, caratterizzata dalla compresenza non soltanto delle varietà locali di italiano e dialetto ma anche di altri italiani regionali e dialetti italiani, oltre che di italiano di non nativi e lingue di recente immigrazione; entrambe le città, infatti, sono e sono state meta di mobilità interna e migrazione.

In fase di raccolta dati sono state registrate diverse informazioni relative ai parlanti, come ad es. luogo di origine, età, titolo di studio, occupazione. Il corpus comprende poi vari tipi di interazione verbale, corrispondenti a diverse situazioni comunicative, classificate essenzialmente secondo i seguenti parametri:

– relazione simmetrica/asimmetrica tra i partecipanti;
– presenza/assenza di un argomento predefinito;
– presenza/assenza di norme per la presa dei turni di parola.

3. *La costruzione del corpus: raccolta dei dati, trascrizione e accessibilità*

La raccolta dati è stata effettuata da ricercatori, ricercatrici, studenti e studentesse delle Università di Bologna e di Torino. Tutte le interazioni sono state registrate a microfono palese e i soggetti coinvolti hanno firmato un consenso informato (conforme alle norme europee di protezione dati – v. G.D.P.R.), che autorizza il gruppo di lavoro a utilizzare i dati raccolti per finalità di ricerca, ad archiviarli e condividerli in forma parzialmente anonimizzata. Per questo motivo, prima della pubblicazione, i materiali linguistici (sia i file audio sia le trascrizioni) sono stati anonimizzati: l'unico dato sensibile direttamente accessibile è la voce stessa del parlante.

Le trascrizioni sono state effettuate utilizzando il software ELAN (Sloetjes & Wittenburg 2008), che permette l'allineamento delle trascrizioni alle relative tracce audio; inoltre, per dare conto di alcune caratteristiche intrinseche della comunicazione parlata (ad esempio l'uso dell'intonazione e la sovrapposizione tra turni di diversi parlanti), si è scelto di seguire una versione semplificata del sistema Jefferson (Jefferson 2004), frequentemente impiegato nell'analisi della conversazione.

Una volta ultimata la raccolta e la trascrizione dei dati, è stato elaborato uno script in python che permette di consultare i dati sulla piattaforma NoSketch Engine, consentendo all'utente di:

- utilizzare i metadati (relativi ai parlanti e alle conversazioni) sia come filtri di ricerca sia come informazioni relative alle singole registrazioni;
- collegare l'occorrenza ricercata con l'unità intonativa in cui si trova;
- avere accesso all'intera trascrizione (ortografica e secondo il sistema Jefferson) della conversazione in cui si trova l'occorrenza cercata;
- effettuare ricerche considerando la semplice trascrizione ortografica;
- consultare separatamente ogni modulo.

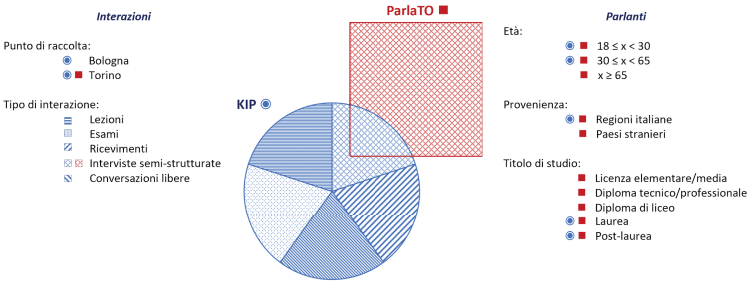
4. *Struttura modulare e incrementale del corpus KIParla*

Il corpus KIParla è caratterizzato da una modularità incrementale, ovvero è organizzato al suo interno in moduli fra loro indipendenti: è dunque possibile aggiungere progressivamente nuovi moduli a quelli esistenti. I moduli sono da intendere come (sotto)corpora di parlato che condividono (almeno) un core set di metadati, presentano una trascrizione effettuata originariamente tramite ELAN, e offrono la consultazione attraverso NoSketch Engine. I vari (sotto)corpora possono concentrarsi su diverse varietà di lingua e/o diversi punti di inchiesta; il disporre di una procedura condivisa per la raccolta e il trattamento dei dati garantisce del resto un alto livello di comparabilità tra i moduli.

Ad oggi, il corpus KIParla è costituito da due (sotto)corpora (v. Fig. 1), il KIP e il ParlaTO.

Il KIP offre innanzitutto la possibilità di indagare fenomeni di variazione diafasica, specialmente di registro, dell'italiano nel parlato di soggetti colti; mentre il ParlaTO offre in primo luogo l'opportunità di esplorare aspetti di diversificazione diastratica dell'italiano parlato. Entrambi i corpora, poi, includono produzioni di parlanti con provenienza geografica diversa; consentono perciò di osservare almeno alcune manifestazioni della variazione diatopica dell'italiano. Con il KIParla, nel complesso, si ha quindi la possibilità di indagare aspetti di diversificazione geografica (KIP e ParlaTO), sociale (ParlaTO) e, limitatamente a parlanti colti, situazionale (KIP) dell'italiano parlato.

Figura 1 - *I moduli attuali del corpus KIParla*



4.1 Il modulo KIP

Il modulo KIP, concepito inizialmente come unità autosufficiente, è stato allestito nell'ambito del progetto *LEAdhoC – Linguistic expression of ad hoc categories* (2015-2019, SIR n. RBSI14IIG0) e rappresenta il nucleo originario del corpus KIParla. La sua costruzione è iniziata nel 2016 e si è conclusa nel 2019.

La risorsa è costituita da circa 70 ore di parlato raccolte a Bologna e a Torino in contesto universitario; le interazioni sono state registrate in diverse situazioni comunicative (v. Fig. 1) e hanno coinvolto studenti e professori universitari. In virtù della gamma dei contesti interazionali considerati, il corpus KIP consente in primo luogo di condurre ricerche su aspetti e fenomeni di variazione diafasica nel parlato di soggetti colti. Nella Tab. 1 si riportano la struttura del KIP, le ore registrate per ciascun contesto, il numero di informatori, le dimensioni e i metadati raccolti.

Tabella 1 - *Il modulo KIP*

Attività	*Bologna*	*Torino*	*TOT*
Conversazioni libere	10:00:37	06:22:24	16:23:01
Esami	03:09:34	03:10:48	6:20:22
Lezioni	12:19:39	13:25:33	25:45:12
Interviste semistrutturate	06:18:37	07:47:38	14:06:15
Ricevimento studenti	02:59:11	03:49:08	6:48:19
TOT	34:47:38	34:35:30	**69:23:08**
Informatori	150	123	273

Metadati	*Parlanti:* classe d'età, sesso, regione in cui si sono frequentate le scuole superiori, occupazione (studente/professore). *Interazioni:* tipo di interazione, relazione tra i partecipanti (simmetrica/asimmetrica), presenza di un moderatore (sì/no), argomento (libero o no), numero di partecipanti.
Dimensioni	661.175 tokens

Con conversazioni libere si intende la registrazione di parlato conversazionale spontaneo tra studenti, raccolto senza alcuna indicazione da parte del gruppo di ricerca. Per ottenere un tipo di evento comunicativo che si avvicinasse il più possibile alla situazione desiderata, la registrazione delle conversazioni è stata svolta nella maggior parte dei casi coinvolgendo direttamente studenti e studentesse delle Università coinvolte: in particolare è stato chiesto loro di registrare autonomamente momenti ricreativi di vario tipo (ad esempio pause durante lo studio, cene, ...) in cui fossero coinvolti altri membri della loro rete sociale, anch'essi studenti universitari.

Esami e ricevimento studenti fanno riferimento ai due contesti più tipici in cui è possibile osservare uno scambio comunicativo a due tra studente e docente. Si tratta in entrambi i casi di attività con un grado maggiore di strutturazione, che prevedono il raggiungimento di scopi concreti da entrambe le parti. Per il ricevimento studenti cfr. ad esempio il lavoro in prospettiva conversazionale di Limberg (2010). Da un punto di vista sociolinguistico si considerano questi eventi come caratterizzati da un maggiore grado di formalità rispetto al parlato spontaneo colloquiale, e da una maggiore presenza dei sottocodici legati alle discipline trattate di volta in volta.

Le lezioni costituiscono il tipo di interazione in cui sono maggiormente coinvolti i docenti, e sono presenti pochissimi interventi da parte degli studenti presenti a lezione, che, in questo caso, sono stati anche esclusi dalla raccolta dei metadati. Il parlato è perlopiù monologico, e dunque caratterizzato da una maggiore pianificazione del turno di parola, e da un maggiore apporto della varietà scritta, ad esempio nel caso di brani letti e commentati o della presenza delle slides.

Infine, l'intervista semi-strutturata è stata inserita in modo da poter includere nel corpus anche interazioni studente-studente caratterizzate da un maggiore grado di strutturazione e da ruoli esplicitamente formalizzati, appunto l'intervistato e l'intervistatore. Infatti, è stato chiesto a studenti e studentesse del tirocinio di intervistare membri della propria rete sociale sulla base di una traccia prestabi-

lita di temi e questioni. In una prima parte, l'intervistatore chiede di descrivere la propria casa facendo un paragone con le altre case in cui gli intervistati hanno vissuto. Successivamente, viene chiesto agli intervistati di raccontare un evento particolarmente significativo legato a una delle case in cui hanno vissuto, in modo da ottenere anche una parte di parlato monologico. Nella Tab. 2 vengono mostrati i metadati che possono essere usati come filtri di ricerca:

Tabella 2 - *KIP: i metadati usabili come filtri di ricerca*

Parlanti	Classe d'età:	under25, 26-30, 31-35, 36-40, 41-45, 46-50, 51-55, 56-60, over60
	Sesso:	M, F
	Regione delle scuole superiori:	Abruzzo, Basilicata, Calabria, Campania, Emilia-Romagna, Estero, Friuli-Venezia Giulia, Lazio, Liguria, Lombardia, Marche, Molise, Piemonte, Puglia, Sardegna, Sicilia, Toscana, Trentino-Alto Adige, Umbria, Valle d'Aosta, Veneto
	Occupazione:	p (professore), s (studente)
Interazioni	Tipo:	conversazione libera, esami, interviste semistrutturate, lezioni, ricevimento studenti
	Luogo:	Bologna, Torino
	Relazione:	Simmetrica, asimmetrica
	Partecipanti:	1, 2, 3, 4, 5, 6
	Moderatore:	Sì, no
	Topic:	Fisso, libero

4.2 Il modulo ParlaTO

Il corpus ParlaTO è stato allestito nell'ambito di un progetto omonimo (*ParlaTO – Corpus plurilingue del parlato di Torino*, Fondazione CRT, E.O. 2018, ID63411) ed è confluito nel più ampio KIParla a settembre 2020. Il ParlaTO è composto essenzialmente da una serie di conversazioni registrate a Torino per mezzo di interviste semi-strutturate. Le interazioni hanno coinvolto parlanti d'età diversa, alcuni di origine italiana (piemontese e non), altri di origine straniera, e con livelli d'istruzione e tipi di occupazione differenti. Le interviste hanno affrontato esperienze personali di vita in città (studio, lavoro, attività nel tempo libero o in pensione, ricordi del passato, ecc.) e hanno visto all'opera più intervistatori, ciascuno dei quali quasi sempre apparte-

nente alla rete sociale dell'intervistato; a una stessa intervista, inoltre, hanno spesso partecipato più intervistati. Il che ha favorito modalità di conversazione tipiche del comportamento *in-group*, caratterizzate dall'impiego di varietà spontanee (e, in certi casi, dall'uso alternato di lingue diverse).

Il corpus comprende due sezioni, contenenti l'una le interazioni con parlanti di origine italiana e l'altra le interazioni con parlanti di origine straniera. Al momento è interrogabile online soltanto la prima sezione, che è costituita di circa 50 ore di parlato e include produzioni in italiano, piemontese e altri dialetti (specialmente di area meridionale). L'uso dell'italiano è largamente prevalente in tutte le conversazioni, ed esclusivo nella maggior parte di esse; il piemontese e gli altri dialetti compaiono invece soltanto occasionalmente, secondo la fenomenologia del discorso bilingue. La seconda sezione, che è in via di allestimento, ammonterà anch'essa a circa 50 ore di parlato e verterà sull'italiano di nativi e non nativi e sulle lingue di recente immigrazione. Le Tabelle 3 e 4 forniscono alcune informazioni essenziali riferite alla prima sezione del corpus.

Tabella 3 - *La prima sezione del corpus ParlaTO*

Informatori	88 parlanti	
Raccolta dati	Torino, 2018-2020	
Metodo	Interviste semi-strutturate (individuali e di gruppo)	
Lingue	Italiano, dialetto piemontese, altri dialetti	
Metadati	Parlanti: classe d'età, sesso, regione di nascita, titolo di studio, occupazione, lingua materna, competenza (attiva o passiva) di dialetto/i italo-romanzo/i, competenza (attiva o passiva) di altre lingue, città e quartiere di residenza, città di nascita del padre e della madre. Interazioni: numero di partecipanti, lingue impiegate	
Ore di parlato	Giovani ($18 \leq x \leq 30$):	17:33:20
	Adulti ($30 < x \leq 60$):	14:49:53
	Anziani ($60 < x \leq 89$):	16:15:31
Dimensioni	552.461 tokens	

Il corpus è dotato di metadati relativi alle caratteristiche sociodemografiche dei parlanti e ad aspetti dell'interazione (v. Tabella 3). Alcuni metadati, quali ad es. la classe d'età, il sesso, la regione di nascita, il titolo di studio e l'occupazione del parlante, sono usabili a tutti gli effetti come filtri di ricerca (v. Tabella 4); altri, quali il quartiere di

residenza dell'informatore o la città di nascita del padre e della madre, sono accessibili soltanto come informazioni supplementari (contenute in tabelle Excel scaricabili dal sito web del corpus). La seconda sezione del corpus avrà inoltre alcuni metadati pertinenti ai parlanti di origine straniera, quali il tempo di permanenza e gli anni di studio in Italia.

Tabella 4 - *La prima sezione del corpus ParlaTO: i metadati usabili come filtri di ricerca*

Parlanti	Classe d'età:	16-20, 21-25, 26-30, 31-35, 36-40, 41-45, 46-50, 51-55, 56-60, 61-65, 66-70, 71-75, 76-80, 81-85, 85-90
	Sesso:	M, F
	Regione di nascita:	Abruzzo, Basilicata, Friuli-Venezia Giulia, Lazio, Lombardia, Piemonte, Puglia, Sardegna, Sicilia, Trentino-Alto Adige, Veneto
	Titolo di studio:	*elem* (diploma di scuola elementare), *medie* (licenza media), *it* (diploma di istituto tecnico o professionale), *lic* (diploma di liceo), *laurea* (laurea triennale, magistrale e a ciclo unico), *phd* (dottorato di ricerca)
	Occupazione:	*artig* (artigiani, operai specializzati e agricoltori), *comm* (professioni qualificate nelle attività commerciali e nei servizi), *disocc* (disoccupati), *impr* (legislatori, imprenditori e alta dirigenza), *intell* (professioni intellettuali, scientifiche e di elevata specializzazione), *nonq* (professioni non qualificate), *oper* (conduttori di impianti, operai di macchinari fissi e mobili e conducenti di veicoli), *pens* (pensionati), *stud* (studenti)
Interazioni	Partecipanti:	2, 3, 4, 5, 6
	Lingue:	italiano, italiano e dialetto

4.3 Nuovi moduli

Attualmente, grazie alla collaborazione degli studenti e delle studentesse dell'Università di Bologna che hanno preso (e stanno prendendo) parte al tirocinio curricolare KIParla, sono in fase di allestimento due nuovi moduli.

Il primo, denominato KIPPasti, è costituito da registrazioni di parlato spontaneo effettuate durante il corso di pranzi e cene. L'ideazione del modulo è avvenuta nella primavera 2020 quando, a causa della situazione pandemica, non era possibile raccogliere dati linguistici di parlato che prevedessero il contatto con membri esterni al proprio nucleo ristretto di conviventi. Il quadro complessivo, dunque, ci ha

portati a pensare ad un tipo di raccolta dati che potesse essere svolto in sicurezza dagli studenti coinvolti nel progetto e, parallelamente, che potesse essere di interesse per ricerche future.

Inoltre, nella larga maggioranza dei casi, i tirocinanti coinvolti nella raccolta dati si trovavano nei loro luoghi di origine. Per questa ragione, si è deciso di cercare di bilanciare il corpus in base all'area geografica di registrazione (nord, centro, sud e isole). In questo momento, sono in via di raccolta le ultime registrazioni necessarie per giungere al bilanciamento e sono in fase avanzata le trascrizioni del materiale raccolto. Ad oggi, il modulo è composto da oltre 50 registrazioni per un totale di oltre 30 ore.

Il secondo modulo, per cui attualmente si sta procedendo alla raccolta dati, è denominato ParlaBO e mira ad avere una struttura analoga a quella del ParlaTO ma ha come unico punto di inchiesta la città metropolitana di Bologna. Fino ad oggi sono state raccolte oltre 20 interviste per un totale di circa 19 ore di registrazione.

4.4 Prospettive future

Al momento, sono in corso collaborazioni con colleghe e colleghi di altri atenei, e dunque si prevede la prossima pubblicazione di ulteriori moduli con dati raccolti in diverse città italiane. Le dimensioni e la rappresentatività del corpus KIParla sono destinate a crescere nel tempo.

In futuro, inoltre, è prevista la lemmatizzazione e il pos-tagging dei dati del corpus (v. già Bosco *et al.* 2020).

Riferimenti bibliografici

Bosco, Cristina & Ballarè, Silvia & Cerruti, Massimo & Goria Eugenio & Mauri Caterina. 2020. "KIPoS @ EVALITA2020: Overview of the Task on KIParla Part of Speech Tagging". In Basile, Valerio & Croce, Danilo & Maro, Maria & Passaro Lucia C. (eds.), *Proceedings of the 7th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian*, 489-495. CEUR.org, http://ceur-ws.org.

Jefferson, Gail. 2004. Glossary of transcript symbols with an introduction. In Lerner, Gene H. (ed.), *Conversation Analysis: studies from the first generation*, 13-31. Amsterdam, John Benjamins.

Limberg, Holger. 2010. *The interactional organization of academic talk: office hour consultations*. Amsterdam , John Benjamins.

Rychlý, Pavel. 2007. Manatee/Bonito – A Modular Corpus Manager. In *1st Workshop on Recent Ad-vances in Slavonic Natural Language Processing*, Brno, Masaryk University, 65-70.

Sloetjes, Han & Wittenburg, Peter. 2008. Annotation by category – ELAN and ISO DCR. In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008)*, 816-820.

DOI: 10.17469/O2106SLI000007

MARCO BIFFI, FRANCESCA CIALDINI

Banche dati per il trasmesso: il LIR e il LIT*

Il contributo ha lo scopo di descrivere due corpora dell'italiano, il LIR – *Lessico Italiano Radiofonico*, banca dati testuale e audio progettata per lo studio della lingua radiofonica, e il LIT – *Lessico Italiano Televisivo*, corpus testuale e audiovisivo sul web, che raccoglie un campione rappresentativo dell'italiano televisivo. I due corpora costituiscono un importante nucleo di base per rappresentare l'italiano trasmesso. Dopo aver ricostruito lo stato dell'arte, a partire dalla definizione di *trasmesso* dagli anni Ottanta, vengono illustrati nel dettaglio i due corpora, il metodo usato per la loro realizzazione, le modalità di interrogazione consentite dal motore di ricerca, le nuove prospettive di studio legate anche al concetto di sostenibilità.

Parole chiave: trasmesso orale, italiano radiofonico, italiano televisivo, diacronia, sostenibilità.

1. *Due corpora per l'italiano trasmesso: progetti, realizzazioni, sostenibilità*

Con il LIR (*Lessico italiano radiofonico*) e il LIT (*Lessico italiano televisivo*) la linguistica dei corpora ha rivolto la propria attenzione all'italiano trasmesso. Il LIR è stato concepito proprio quando Francesco Sabatini mise definitivamente a fuoco l'etichetta di *trasmesso*; dopo averla usata nel 1982 in un volume collettaneo dedicato all'educazione linguistica[1] e averla recuperata nella sua grammatica *La comunicazione e gli usi della lingua* nel 1984 (sottolineandone peraltro la non ortodossia con l'uso delle virgolette)[2], è nel 1994 che ne definisce i contorni e le caratteristiche, sdoganandola così per gli studi linguistici, in una relazione dal titolo, appunto, *Prove per l'italiano «trasmesso»*, al Convegno *Gli italiani trasmessi: la radio*, tenutosi a Firenze,

* Il contributo è frutto dell'elaborazione comune dei due autori, tuttavia si deve a Marco Biffi la stesura del paragrafo 1 e a Francesca Cialdini la stesura del paragrafo 2.

[1] Sabatini 1982.

[2] Sabatini 1984.

all'Accademia della Crusca, nel maggio del 1994, relazione poi pubblicata negli atti usciti nel 1997[3].

Fu proprio a seguito dei lavori di quel convegno che Nicoletta Maraschio ebbe l'idea di compilare un lessico della lingua radiofonica. E secondo la tradizione cruscante si pensò a un corpus informatizzato di partenza; un corpus che doveva però avere caratteristiche diverse da quelli visti fino a quel momento e a cui si era abituati[4]. Innanzi tutto, doveva dare accesso ai materiali autentici trascritti e informatizzati, caratteristica – si badi bene – che mancava anche al LIP (*Lessico di frequenza dell'italiano parlato*), che presentava gli indici lessicali in formato cartaceo, a stampa, e dava accesso soltanto alla trascrizione dei testi orali usati come base di partenza, peraltro a disposizione unicamente con un disco floppy da 3 pollici e mezzo allegato al volume[5]. Come è ben noto, sono successive le versioni consultabili attraverso il web, che prima hanno dato accesso diretto alla trascrizione dei testi e poi anche alla voce[6]. L'accesso al materiale autentico per il LIR era centrale: era infatti l'unica garanzia per lo studioso di poter analizzare tutte le specificità della lingua della radio, anche quelle soprasegmentali di intonazione e marcature espressive.

Quando si cominciò a progettare il LIR, a partire appunto dal 1994, gli ostacoli tecnologici erano enormi, a cominciare dalla possibilità di archiviare i file audio di un corpus di 108 ore di registrazione, che dovevano tradursi in circa 60 ore di parlato al netto della musica (un ordine di grandezza voluto perché il corpus fosse comparabile con il LIP). Si sfruttarono al massimo le tecnologie disponibili ed Eugenio Picchi realizzò una versione apposita del suo DBT (il DBT LIR), che consentiva – una volta agganciati i contesti allargati – di allineare in perfetto sincrono il file audio autentico: si cercava una forma e si arrivava ad ascoltarla nel punto esatto in cui era stata pronunciata, con

[3] Sabatini 1997.

[4] Sulla realizzazione del corpus LIR e sulla metodologia di ricerca in questa sede per motivi di spazio ci limitiamo a segnalare solo alcuni studi pubblicati nel corso degli anni. Si vedano almeno Maraschio *et al*. 1997; Maraschio *et al*. 2004; Alfieri & Stefanelli 2005; Biffi & Setti 2008.

[5] *LIP Lessico di frequenza dell'italiano parlato*, a cura di Tullio De Mauro, Federico Mancini, Massimo Vedovelli, Miriam Voghera, Milano, Etas, 1993.

[6] I testi sono disponibili nella *Banca Dati dell'Italiano Parlato*, consultabile all'indirizzo http://badip.uni-graz.at/it/; le registrazioni sono disponibili nel *VOLIP -Voce del LIP*: https://parlaritaliano.studiumdipsum.it/index.php/it/volip.

l'indicazione della trasmissione, del genere, della tipologia comunicativa e dello speaker (interno/esterno; maschio/femmina)[7].

Erano tempi pionieristici in cui non era ben chiara l'importanza strategica della sostenibilità. Eppure il LIR ha involontariamente precorso i tempi anche da questo punto di vista. La prima versione consultabile in locale divenne presto sempre più complicata da gestire e così è stata necessaria una prima implementazione e il trasferimento della banca dati sul web, con un nuovo software, che non garantiva l'elasticità e la potenza di ricerca della consultazione della versione in DBT, ma che però lo rendeva disponibile a tutti per scopi scientifici ma anche didattici. Tanto più che nel 2005 si erano avviati i lavori per realizzare un analogo corpus per la lingua della televisione, il LIT, che doveva avere caratteristiche analoghe al LIR con i dovuti adattamenti legati al mezzo[8]: in questo caso era necessario risalire nuovamente a un file audio-video autentico, perché per la lingua della televisione il contesto doveva essere allargato all'immagine per recuperare prossemica, postura, espressioni, combinazioni con altri codici.

Il nuovo software per il LIT, nativo web, fu adattato al LIR, tanto più che in questo modo i due corpora sarebbero stati omogenei e integrabili tra loro, come avviene effettivamente nel portale *Vivit – Vivi italiano*, all'interno della sezione *Archivi digitali*[9].

Il LIT è anche consultabile a partire dal *Portale dell'italiano televisivo*, realizzato all'interno di un PRIN a cui hanno collaborato vari gruppi di ricerca nazionali[10].

LIR e LIT si sono però dovuti nuovamente confrontare con la sostenibilità: la scelta fatta dagli informatici di appoggiarsi ad Adobe Flash Player si è dimostrata infatti esiziale quando – dopo gli annun-

[7] Cfr. Maraschio *et al.* 2004, pp. 21-34.

[8] Anche per il LIT ci limitiamo a segnalare alcuni studi pubblicati negli anni: si vedano almeno Mauroni & Piotti 2010 e Alfieri *et al.* 2016. Per la parte specificatamente informatico-linguistica cfr. Biffi 2010.

[9] *Vivit – Vivi italiano. Il portale dell'italiano nel mondo* è un archivio informatico di materiali e strumenti rivolti agli italiani all'estero, in particolare a quelli di seconda e terza generazione: https://www.viv-it.org/. Per la sezione degli *Archivi digitali* l'indirizzo diretto è il seguente: https://www.viv-it.org/schede/archivi-digitali.

[10] Il portale è il frutto del progetto PRIN 2008 *Il portale dell'italiano televisivo: corpora, generi e stili comunicativi* (unità di ricerca: Università di Firenze, Università di Catania, Università di Genova, Università di Milano, Università della Tuscia). È consultabile all'indirizzo www.italianotelevisivo.org, al quale si rimanda per approfondimenti.

ci di dismissione e la dichiarata dismissione – a partire dal primo gennaio 2021 tutto ciò che si basava su questo software ha cessato di funzionare.

Complice l'emergenza Coronavirus, quando abbiamo fatto la nostra proposta nel 2019 per il poster, LIR e LIT erano consultabili, ma quando si stava definendo il programma non lo erano già più.

Il LIR e il LIT attualmente consultabili si basano quindi su una nuova versione del programma di interrogazione (transitoria), sviluppata in questi ultimi mesi per traghettare questi due strumenti verso una nuova piattaforma che finalmente sarà realizzata con la massima attenzione alla sostenibilità.

Passiamo ora una descrizione più articolata dei due strumenti e delle loro funzionalità, illustrate anche nel divenire delle varie versioni descritte.

2. *I due corpora: descrizione e funzionalità*

2.1 Il LIR

Il LIR è una banca dati testuale e audio progettata con l'obiettivo di studiare la varietà radiofonica di trasmesso. Il primo corpus, rappresentativo delle principali emittenti nazionali, è stato raccolto nel 1995, secondo un prelievo a scacchiera su una settimana di maggio e comprende 108 ore di trasmissioni radiofoniche di 9 emittenti nazionali (Radio1, Radio2, Radio3, Radio Deejay, Rete 105, RTL 102.5, Italia Radio, Radio Radicale, Radio Vaticana)[11]. La banca dati raccoglie, dunque, 64 ore di parlato, 650.000 occorrenze, 86.000 forme.

Il secondo corpus, relativo al 2003 e limitato alle reti RAI, comprende 32 ore di parlato, 310.000 occorrenze, 39.000 forme; come per il primo corpus, i prelievi sono stati effettuati sulla settimana di maggio, meno caratterizzata da eventi esterni come Natale, Pasqua e periodi festivi[12].

Il lavoro di costituzione del corpus si è articolato in quattro fasi principali: 1) trascrizione del materiale audio; 2) inserimento manua-

[11] Per motivi tecnici sono state aggiunte alcune registrazioni di trasmissioni di Radio Deejay e RTL 102.5 in onda in una settimana del febbraio 1996.

[12] Su questo aspetto si veda lo studio di Biffi & Setti 2008: 351.

le dei testi nella piattaforma e creazione delle trasmissioni; 3) allineamento testuale con i file mp3 forniti dalle emittenti; 4) marcatura del testo con DBT di Eugenio Picchi, che – come spiegato sopra – ha poi realizzato una specifica versione del motore di ricerca (DBT LIR) per la consultazione integrata di testi e file audio.

Il materiale è stato trascritto interamente secondo criteri che, tenendo conto di una serie di elementi, permettono di individuare i principali tratti linguistici del trasmesso radiofonico. Ne riportiamo alcuni, a titolo esemplificativo:

- si segnala la fine di enunciato dichiarativo con la doppia barra obliqua //;
- la barra semplice / indica la scansione interna dell'enunciato, come eventuali pause o il cambio di intonazione;
- le parentesi uncinate < > indicano le sovrapposizioni di turno;
- {T} indica un troncamento;
- { indica l'inizio di una esitazione e } ne indica la fine.

Nello specifico, per quanto riguarda la marcatura del corpus, sono state individuate le quattro categorie di *Radio*, *Genere*, *Tipologia comunicativa*, *Speaker*. All'interno di queste categorie sono presenti alcune sottocategorie interessanti dal punto di vista della ricerca sociolinguistica (per esempio, oltre alle emittenti radiofoniche, *Tipologia comunicativa*: monologo, dialogo, telefonata, monologo a più voci, turno frammentato, esecutivo, programmato, semi-improvvisato, spontaneo; *Genere*: pubblicità, annunci, letteratura, notizie, intrattenimento culturale, intrattenimento leggero; *Speaker*: professionista, esterno, maschio, femmina). Con la presenza di categorie e sottocategorie possiamo sia fare la ricerca di singoli brani sia individuare specifici sub-corpora linguistici. Inoltre, è possibile interrogare sia separatamente il LIR 1995 (definito LIR/1) e il LIR 2003 (definito LIR/2) sia in contemporanea, in modo da permettere, almeno per le reti RAI, una prima valutazione dei cambiamenti in diacronia nella lingua della radio[13].

[13] È anche presente la sezione *Sala d'ascolto*, una stanza virtuale in cui è possibile selezionare una trasmissione e allo stesso tempo ascoltare la registrazione e seguire la trascrizione.

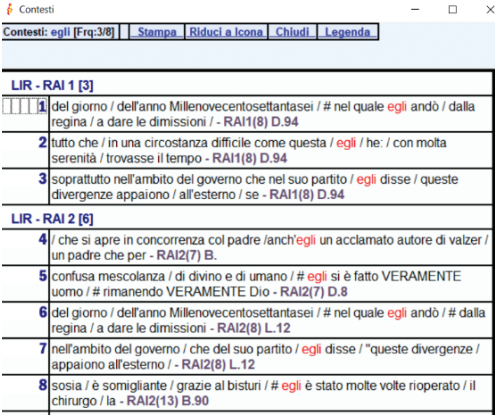
Figura 1 - *La distribuzione di* egli *alla radio nel 1995 e nel 2003*

Nel LIR è possibile effettuare ricerche di forme di vario tipo, a partire da quelle che riguardano la veste fonica e che consentono riflessioni sui tratti della pronuncia, fino a quelle relative a fenomeni morfosintattici tipici dell'italiano contemporaneo. Per esempio, come è stato messo in evidenza anche in altri studi, è interessante osservare la distribuzione dei pronomi di terza persona *egli*/*lui* e analizzare in particolare l'uso di *egli* tra 1995 e 2003 (36 nel LIR/1 e 6 nel LIR/2, per un totale di 42 occorrenze)[14]. Per ciascun risultato ottenuto è possibile allargare il contesto, ascoltare la porzione di testo che interessa, ricevere informazioni sui metadati e avere a disposizione l'intera trascrizione della trasmissione. Tra le ricerche possibili ricordiamo anche quella dei forestierismi, interessante anche in una prospettiva diacronica: nel LIR/1 1995, per quanto riguarda la RAI, risultano 509 forestierismi; nel LIR/2 2003 (solo RAI, come abbiamo detto) se ne contano 1050[15].

[14] Si vedano Maraschio 2005: 140-141 e Biffi & Setti 2008: 355.

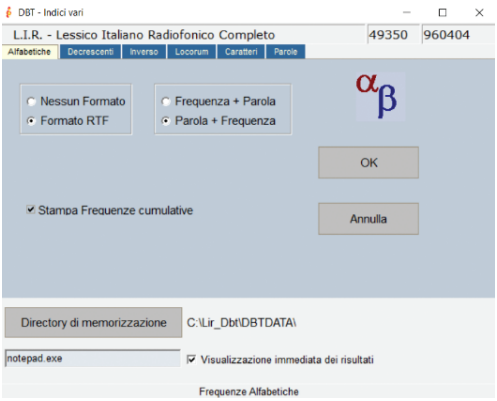
[15] Sui forestierismi nel LIR si veda gli studi di Fanfani 1997a e Fanfani 1997b. Come nota anche Maraschio 2016: 77, dagli anni Novanta a oggi è entrato in italiano un numero di forestierismi, in particolare anglismi, superiore al doppio di quello entrato nel decennio precedente.

Figura 2 - *Il contesto allargato*



Infine, la sezione "Indici" consente la generazione di indici statistici e di frequenza.

Figura 3 - *Indici vari*



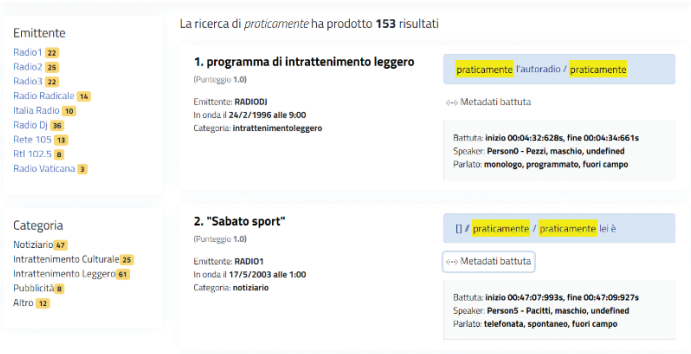
2.1.1 Il LIR web

La versione web del LIR[16] presenta un'interfaccia di interrogazione in cui sono possibili diverse tipologie di ricerche, a partire da una ricerca semplice e da una avanzata. La più immediata è quella che riguarda

[16] http://lir.accademiadellacrusca.org/lir2/.

una singola forma; per esempio, se cerchiamo *praticamente*, avverbio diffuso alla radio, otteniamo i seguenti risultati:

Figura 4 - *I risultati di ricerca*



I dati ottenuti, in primo luogo, sono relativi alle occorrenze totali (153 risultati), con le frequenze distribuite per emittente e per categoria; sono presenti inoltre i contesti immediati, che contengono le informazioni sulla trasmissione (titolo, emittente, data, categoria). È possibile anche consultare i metadati della battuta relativi a parlante, tipologia comunicativa, presenza in campo o fuori campo. Cliccando su uno dei contesti si accede al contesto allargato della trascrizione e alla registrazione audio digitale parallelizzata.

Figura 5 - *La registrazione audio parallelizzata*

Con la ricerca avanzata sono possibili ricerche più raffinate. Infatti, è possibile la ricerca di più forme attraverso l'operatore booleano AND ("Trova i risultati che contengano tutte le seguenti parole"), di una sequenza ("Trova i risultati che contengano la seguente sequenza libera", sia come "sequenza esatta", sia "sequenza ordinata" sia "sequenza non ordinata") e di forme attraverso gli operatori OR e AND NOT ("Trova i risultati che contengano una qualunque delle seguenti parole" e "Trova i risultati che non contengano le seguenti parole"). La distanza misurata in parole viene stabilita dall'utente. Per esempio, possiamo cercare la sequenza esatta *piuttosto che* (con distanza 1), per osservare la diffusione del suo uso con valore disgiuntivo in diacronia. Analizzando i contesti dei 36 risultati totali ottenuti, notiamo che le occorrenze di *piuttosto che* con valore disgiuntivo sono 5 nel corpus del 1995 e salgono a 7 nel 2003.

Figura 6 - *I risultati relativi a* piuttosto che

Inoltre, è possibile restringere la ricerca in base all'emittente, ai generi e alle tipologie e considerare le maiuscole/minuscole, ottenere l'elenco delle forme e le statistiche globali. Nella ricerca è consentito l'utilizzo dei caratteri *jolly*: il *?* sostituisce un carattere e l'asterisco *estende la ricerca a intere parti di parola. Se cerchiamo, per esempio, la forma *maxi** (*maxi-* è tra i prefissi più produttivi in LIR1/2), risulteranno 20 occorrenze. È possibile sia accedere ai contesti sia ottenere l'elenco delle forme:

Figura 7 - *L'elenco delle forme*


La ricerca di *maxi** ha prodotto **20** risultati

Forma	Num. occorrenze
maxi	15
maxiemendamento	1
maxiprocessi	1
maxirichiesta	1
maxitamponamento	1

2.2 Il LIT

Il LIT è una banca dati interrogabile che raccoglie 168 ore di trasmissioni delle reti Rai e Mediaset, prelevate nel corso del 2006 secondo una griglia statisticamente rappresentativa. Nel corpus è possibile ricercare una parola (o un gruppo di parole), accedere a dati quantitativi sulla frequenza, ai contesti immediati e al materiale autentico trascritto e marcato secondo vari parametri[17]. Come ricordato, il corpus LIT condivide le caratteristiche del LIR sia per quanto riguarda l'individuazione del campione rappresentativo di riferimento sia nelle modalità di interrogazione. Infatti, i prelievi sono stati effettuati nell'arco temporale dell'anno 2006 (dal 2 gennaio al 26 dicembre), interessante dal punto di vista linguistico per gli eventi che lo hanno caratterizzato – per esempio le Olimpiadi, i mondiali di calcio e le elezioni politiche –, nella fascia oraria serale che va dalle 19.00 alle 23.00, secondo una griglia rappresentativa dal punto di vista statistico. Per ciascuna rete è stato effettuato un prelievo di mezz'ora alla settimana a rotazione, in modo da coprire tutti i giorni della settimana nell'arco dell'anno[18].

[17] Per una prima descrizione del corpus si veda Biffi 2010. La banca dati è consultabile all'indirizzo http://lit.accademiadellacrusca.org/lit2/; come ricordato sopra, una versione precedente è consultabile nel *Portale dell'italiano televisivo* https://www.italianotelevisivo.org/ e negli *Archivi digitali* del *Vivit* https://www.viv-it.org/schede/archivi-digitali.

[18] Per esempio, lo schema di prelievo delle prime due settimane è il seguente: 1ª settimana, lunedì 19.00-19.30 Canale5 (02/01/06), martedì 19.30-20.00 Rete4 (03/01/06), mercoledì 20.00-20.30 Italia1 (04/01/06); 2ª settimana, giovedì 20.30-21.00 Canale5 (12/01/06), venerdì 21.00-21.30 Rete4 (13/01/06), sabato 21.30-22.00 Italia1 (14/01/06). Sui prelievi e sul metodo usato si veda Biffi 2010.

Dal punto di vista metodologico, il lavoro di costituzione del corpus si è articolato in quattro fasi principali: 1) trascrizione del materiale audiovisivo; 2) inserimento dei testi nella piattaforma; 3) allineamento testuale con i file video forniti da Rai e Mediaset; 4) marcatura del testo con annotazioni XML/TEI.

I criteri di trascrizione del materiale consistono in una versione semplificata dei criteri utilizzati per il LIR e permettono di individuare i principali tratti linguistici del trasmesso. Per esempio:

- si segnala la fine di enunciato dichiarativo con la doppia barra obliqua //;
- la barra semplice / indica la scansione interna dell'enunciato, come eventuali pause o il cambio di intonazione;
- le parentesi uncinate < > indicano le sovrapposizioni di turno;
- il trattino – segnala l'interruzione di parola (attaccato alla parola interrotta, es.: *veram-*);
- i due punti *:* indicano l'allungamento dovuto a esitazione;
- nel caso di parola o parte di parola incomprensibile, la forma linguistica è sostituita da *[xxx]*;
- il maiuscolo indica un fenomeno di enfasi[19].

Una volta immesse nella piattaforma le trascrizioni del materiale audiovisivo e create le trasmissioni (complete delle informazioni di *titolo*, *rete*, *data*, *ora*), sono state individuate alcune categorie funzionali alla ricerca. Oltre alle emittenti, sono presenti le seguenti categorie relative ai generi e sottogeneri[20]: 1) *Pubblicità*; 2) *Fiction*: Film TV, Miniserie, Serie, Serial; 3) *Intrattenimento*: Cartoni animati, Varietà, Game show, Reality show, Comico/satirico, Talk show Intrattenimento; 4) *Informazione*: TG, Reportage/inchiesta, Telecronaca in diretta, Approfondimento, Talk show Informazione; 5) *Divulgazione scientifica e culturale*: Documentari, Magazine, Talk show Divulgazione.

[19] Altri criteri di trascrizione sono: punto interrogativo *?* alla fine di enunciato interrogativo; punto esclamativo *!* alla fine di enunciato esclamativo; puntini di sospensione ... per l'intonazione sospensiva; apici doppi " " per indicare titoli, citazioni, discorso diretto riportato; parentesi quadre *[ ]* per commenti paralinguistici e situazionali. Le sigle uniformate sono le seguenti: *Rai1*, *Rai2*, *Rai3*, *Canale5*, *Italia1*, *Rete4*; *Tg1*, *Tg2*, *Tg3*, *Tg4*, *Tg5*, *Studio Aperto*.

[20] Come nota Biffi 2010: 45, la categoria generi/sottogeneri è stata quella più complicata da definire, dato l'ibridismo dei generi tipico della neotelevisione.

Inoltre, sono state individuate le categorie che fanno riferimento alla tipologia comunicativa (*monologo*, *dialogo*), al tipo di parlato legato al grado di spontaneità (*esecutivo*, *programmato*, *improvvisato*), al parlante, descritto in base ad alcuni parametri sociolinguistici (*interno* se si tratta del professionista della televisione, esterno se non fa parte del mondo della tv; *uomo*/*donna*; *in campo*/*fuori campo*).

Nella fase successiva del lavoro il testo trascritto è stato allineato con i file video forniti da Rai e Mediaset e marcato con con specifiche annotazioni XML/TEI. Per ciascuna porzione di battuta è stata definita un'annotazione con marcatori XML/TEI in base alle categorie sopra descritte. In particolare, la classificazione in base al parlante, al quale vengono associate le porzioni di testo trascritto (a ciascun parlante viene attribuito un colore di riferimento, con lo scopo di facilitare l'individuazione del cambio turno), costituisce un aspetto rilevante per la ricerca[21].

Come per il LIR, anche per il LIT è possibile sia una ricerca semplice sia una più avanzata. Se cerchiamo, per esempio, un aggettivo come *straordinario*, otteniamo i seguenti risultati:

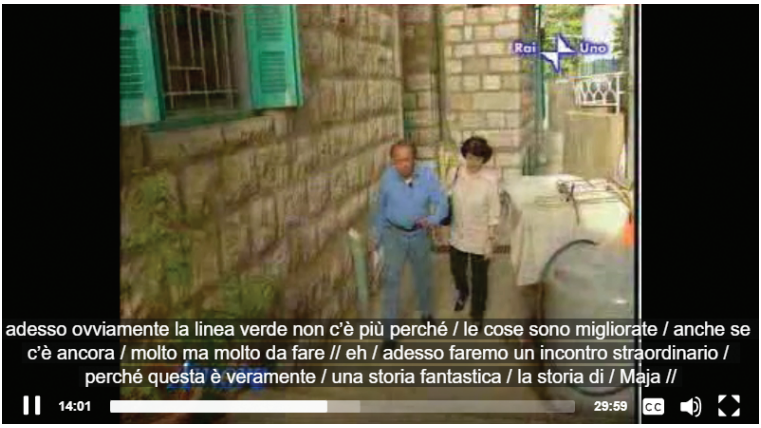
Figura 8 - *I risultati*

Le informazioni ottenute sono relative alle occorrenze totali (61 risultati), con le frequenze distribuite per emittente e per categoria. Sono presenti, inoltre, i contesti immediati, che contengono le informazioni sulla trasmissione: titolo, emittente, data, categoria. È possibile consultare anche i metadati della battuta relativi a parlante, tipologia

[21] Su questo aspetto si veda lo studio di Biffi 2010: 45-50.

comunicativa, presenza in campo o fuori campo; cliccando su uno dei contesti si accede al contesto allargato della trascrizione e alla registrazione audio digitale parallelizzata:

Figura 9 - *Il collegamento alla registrazione audio parallelizzata*



Come nel LIR, sono possibili le ricerche con i caratteri jolly: per esempio, *straordinari** consentirà di trovare le forme di maschile plurale e di femminile singolare e plurale e la forma dell'avverbio in *-mente*. Inoltre, la ricerca avanzata consente di ottenere risultati più raffinati: è possibile, infatti, la ricerca di più forme ("Trova i risultati che contengano tutte le seguenti parole"), di una sequenza ("Trova i risultati che contengano la seguente sequenza libera", sia come "sequenza esatta", sia "sequenza ordinata" sia "sequenza non ordinata") e di forme attraverso gli operatori OR e AND NOT ("Trova i risultati che contengano una qualunque delle seguenti parole" e "Trova i risultati che non contengano le seguenti parole"). La distanza misurata in parole viene stabilita dall'utente.

Infine, è possibile restringere la ricerca in base all'emittente, ai generi (e sottogeneri), alla tipologia comunicativa, alla tipologia di parlato, al parlante (maschio/femmina, interno/esterno, in campo/fuori campo) e considerare le maiuscole/minuscole, ottenere l'elenco delle forme e le statistiche globali.

Figura 10 - *La ricerca avanzata*

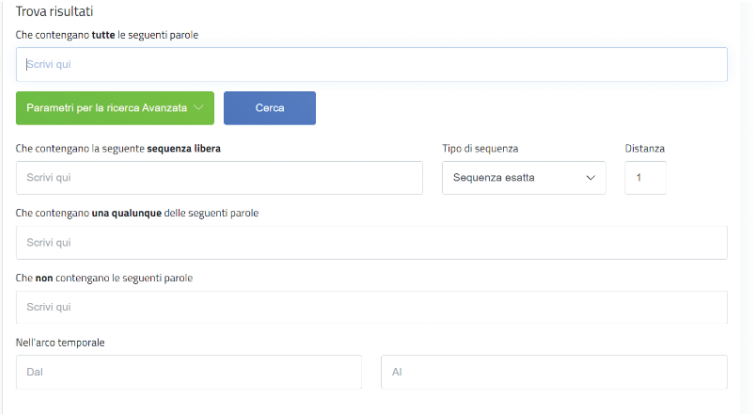
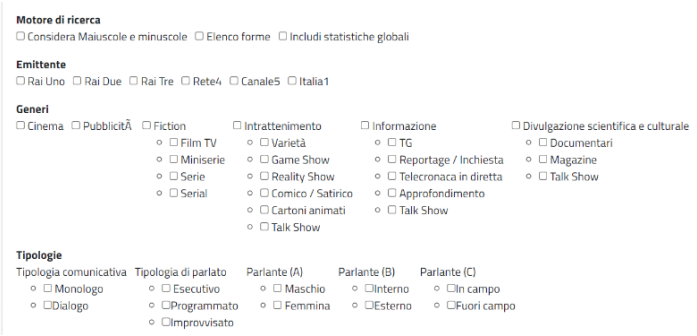


Figura 11 - *La ricerca per categorie*



Riferimenti bibliografici

Alfieri, Gabriella & Stefanelli, Stefania. 2005. Lessici dell'italiano radiofonico (LIR). In Burr, Elisabeth (a cura di), *Tradizione & innovazione. Il parlato: teoria – corpora – linguistica dei corpora. Atti del VI Convegno SILFI, Suisburg, 28 giugno-2 luglio 2000*. Firenze: Cesati. 397-412.

Alfieri, Gabriella & Biffi, Marco & Giuliano, Mariella & Motta, Daria (a cura di). 2016. *Il portale della tv, la tv dei portali, Atti del Convegno (Firenze, Accademia della Crusca, 8 marzo 2013)*. Acireale – Roma: Bonanno.

Biffi, Marco. 2010. Il LIT – Lessico Italiano Televisivo. In Mauroni, Elisabetta & Piotti, Mario (a cura di), *L'italiano televisivo 1976-2006*. Firenze: Accademia della Crusca. 35-70.

Biffi, Marco. 2016. Il portale dell'italiano televisivo: corpora, generi, stili comunicativi. In Alfieri, Gabriella & Biffi, Marco & Giuliano, Mariella & Motta, Daria (a cura di), *Il portale della tv, la tv dei portali. Atti del Convegno (Firenze, Accademia della Crusca, 8 marzo 2013)*, 11-30. Acireale – Roma: Bonanno.

Biffi, Marco & Setti, Raffaella. 2008. Dieci anni di italiano parlato alla radio: corpora LIR 1995/ LIR 2003 a confronto. In Pettorino, Massimo (a cura di), *La comunicazione parlata, Atti del Congresso Internazionale (Napoli 23-25 febbraio 2006)*, 361-398. Napoli: Liguori.

Bonomi, Ilaria & Maraschio, Nicoletta. 2016. *Giornali, radio e tv: la lingua dei media* (Collana "l'Italiano. Conoscere e usare una lingua formidabile", a cura dell'Accademia della Crusca e Repubblica, n. 8) [poi ristampato nel 2017]. Roma: Gruppo Editoriale L'Espresso.

Cialdini, Francesca. 2016. L'aggiornamento della banca dati LIT e il DIALIT. In Alfieri, Gabriella & Biffi, Marco & Giuliano, Mariella & Motta, Daria (a cura di), *Il portale della tv, la tv dei portali. Atti del Convegno (Firenze, Accademia della Crusca, 8 marzo 2013)*, 31-47. Acireale – Roma: Bonanno.

Fanfani, Massimo. 1997a. Forestierismi alla radio. In *Gli italiani trasmessi: la radio*, 729-788. Firenze: Accademia della Crusca.

Fanfani, Massimo. 1997b. I programmi radiofonici della RAI. *Bollettino d'informazioni* VII (1-2). 73-77.

LIP *Lessico di frequenza dell'italiano parlato*, a cura di Tullio De Mauro, Federico Mancini, Massimo Vedovelli, Miriam Voghera, Milano, Etas, 1993.

Maraschio, Nicoletta. 2005. La Radio. In Lo Piparo, Franco & Ruffino, Giovanni (a cura di), *Gli italiani e la lingua*, 135-146. Palermo: Sellerio Editore.

Maraschio, Nicoletta. 2016. La radio. In Bonomi, Ilaria & Maraschio, Nicoletta (a cura di), *Giornali, radio e tv: la lingua dei media* (Collana "l'Italiano. Conoscere e usare una lingua formidabile", a cura dell'Accademia della Crusca e Repubblica, n. 8) [poi ristampato nel 2017], 53-98. Roma: Gruppo Editoriale L'Espresso.

Maraschio, Nicoletta & Antonini, Anna & Bellucci, Patrizia & Fanfani, Massimo & Stefanelli, Stefania & Avesani, Cinzia & Pratesi, Monica. 1997. Il progetto LIR. I lessici di frequenza dell'italiano radiofonico. *Bollettino d'informazioni* VII (1-2). 53-94.

Maraschio, Nicoletta & Stefanelli, Stefania & Buccioni, Stefania & Biffi, Marco. 2004. Dal corpus LIR: prove e confronti lessicali. In Albano

Leoni, Federico & Cutugno, Francesco & Pettorino, Massimo & Savy Renata (a cura di), *Il Parlato Italiano, Atti del Convegno Nazionale "Il Parlato Italiano", Napoli 13-15 febbraio 2003*, 1-36. Napoli: M. D'Auria Editore.

Sabatini, Francesco. 1982. La comunicazione orale, scritta, trasmessa: la diversità del mezzo, della lingua e delle funzioni. In Boccafurni, Anna Maria & Serromani, Simonetta (a cura di), *Educazione linguistica nella scuola superiore. Sei argomenti per un curricolo*, 105-127. Roma: Istituto di Psicologia CNR.

Sabatini, Francesco. 1984. *La comunicazione e gli usi della lingua: pratica, analisi e storia della lingua italiana.* Torino: Loescher.

Sabatini, Francesco. 1997. *Prove per l'italiano «trasmesso»*, In *Gli italiani trasmessi: la radio*, 11-30. Firenze: Accademia della Crusca.

Setti, Raffaella. 2011. Interrogando il LIT. Il lessico televisivo contemporaneo tra spettacolarità e stereotipia. Lo spettacolo delle parole. In Caffarelli, Enzo & Fanfani, Massimo (a cura di), *Lo spettacolo delle parole. Studi di storia linguistica e onomastica in onore di Sergio Raffaelli*, 167-182. Roma: Società Editrice Romana.

DOI: 10.17469/O2106SLI000008

GIORGINA CANTALINI

Corpus multimodale annotato per lo studio della gestualità co-verbale nel «parlato-parlato» e nel «parlato-recitato»

Il corpus in oggetto studia le relazioni della gestualità co-verbale con i diversi livelli di organizzazione linguistica, in particolare la strutturazione prosodica. È composto da campioni comparabili di interviste fatte ad attori di prosa, parallelamente registrati nell'esecuzione di uno stesso brano teatrale. Offre un modello di metodologia di annotazione in cui i quadri teorici utilizzati, di approccio configurazionale, mostrano caratteristiche di compatibilità e sovrapponibilità funzionalmente efficaci; inoltre comparando due varietà (spontaneo e recitato) e due modalità (orale e gestuale) e sfruttando l'opportunità doppia di confronto restituisce risultati significativi sia riguardo alla sincronizzazione tra modalità che alla variazione di strategie di modellizzazione tra varietà. Il tipo di analisi proposta consente dunque di approfondire lo studio della relazione della gestualità con il parlato orale. Consente inoltre avanzamenti nello studio della semantica del gesto.

Parole chiave: annotazione multilivello, corpora multimodali. gesto, prosodia, parlato recitato, parlato spontaneo, variazione diamesica, variazione modale.

1. *Introduzione*

Il corpus multimodale di parlato spontaneo e parlato recitato, di cui al demo[1] presentato al LIV Convegno Internazionale di Studi della SLI e del quale questo contributo scritto costituisce un'integrazione, è dedicato allo studio della gestualità co-verbale (a partire da Kendon 1972, 1980, 2004; McNeill 1992, 2005) e alle sue relazioni con i diversi livelli di organizzazione linguistica, in particolare la strutturazione prosodica (Loehr 2004; Cresti 2000).

[1] https://doi.org/10.48448/3k4n-hg49

Alla base dello studio condotto vi sono state sia motivazioni individuali che di interesse più generale. In quanto didatta della recitazione e attrice io stessa, mi occupo da tempo di parlato-scritto (Nencioni 1976) con l'obiettivo di farlo risultare parlato-parlato: una pagina scritta, in cui il linguaggio sia già stato espresso e depositato su un supporto cartaceo, deve essere eseguita simulandone condizioni di pronunciamento affatto differenti e invece proprie del parlato spontaneo (su tutto: il processo di ideazione deve svolgersi contemporaneo a quello di locuzione sotto la spinta di una reale interazione comunicativa). Numerose osservazioni empiriche negli anni mi hanno portato a constatare come tale risultato fosse ottenibile solo grazie a uno specifico coinvolgimento fisico-dinamico di tutto il corpo, che è cominciato ad apparirmi, negli anni, sistematico e necessario non solo per l'*espressione*, con riferimento alle performance attorali o oratorie, ma anche per la *comprensione* del testo in quanto pagina scritta. Dunque, se per far funzionare un testo scritto oralmente non bastava solo la voce, ma era necessario anche il movimento, con riguardo particolare ai movimenti gesticolatori, la questione diventava capire in che maniera il corpo fosse coinvolto nel linguaggio, strutturalmente, tipologicamente e funzionalmente (quali gesti, quali spinte dinamiche, quale rapporto con la gravità e in che maniera voce e gesto fossero in relazione). In quanto linguista invece, l'interesse era rivolto allo studio del linguaggio e alla crescente attenzione che la letteratura attuale sta dando alla costituzione di corpora multimodali annotati di linguaggio e alla formazione di modelli di annotazione del gesto. Da una parte è ormai assodato che all'interno del variegato e complesso excursus di manifestazioni corporali accomunate dalla definizione di comportamento non-verbale (Ekman & Friesen 1969), o di linguaggio del corpo (Maricchiuolo 2017), è possibile individuare una tipologia specifica di gestualità considerata costitutiva e complementare al linguaggio parlato (McNeill 1985), che occorre co-verbalmente, alternativamente o in sostituzione ad esso[2], e che come tale non può più essere ignorata nello studio di quest'ultimo. D'altro canto ci si trova in un momento storicamente favorevole all'osservazione della gestualità, contrariamente a quanto avvenuto per la gran parte del secondo scorso, grazie alla diffusione di strumentazione audiovideo di alto livello, facile utilizzo e co-

[2] Si veda la classificazione "Kendon's continuum" in McNeill 1992.

sti accessibili. In questo quadro di domanda individuale e opportunità generale è da collocarsi dunque il lavoro qui esposto.

Saranno presentati la tipologia di dato acquisito e i criteri operativi adottati per la sua acquisizione (2.); i modelli teorici presi a riferimento (3.); la metodologia di annotazione, nello specifico la strutturazione di quali e quanti livelli predisposti e le modalità di segmentazione del gesto e della prosodia utilizzate (4.); aspetti relativi alla validazione (5.); evidenze osservate e produttività del modello di annotazione approntato (6.); infine saranno riportate in breve alcune conclusioni (7.).[3]

2. *Tipologia di dato acquisito e criteri operativi di acquisizione*

Per ottenere materiale comparabile di due diverse varietà linguistiche si è scelto di far recitare a tre attori di prosa, uomini, italiani madrelingua, il medesimo pezzo teatrale prima e poi di sottoporli ad un'intervista strutturata. Il pezzo era tratto da "Il giuoco delle parti" di Luigi Pirandello[4]. Due degli attori sono stati registrati in uno studio di posa e hanno recitato il brano espressamente per la ricerca; il terzo è stato registrato durante lo spettacolo dal vivo dell'intera commedia di Pirandello in scena proprio nel periodo in cui è stata condotto il presente lavoro; e alle tre interpretazioni ottenute si è aggiunta quella di un quarto attore dalle Teche Rai. Si sono collezionati così circa 8 minuti di varietà recitata, in cui il personaggio principale, pur trattandosi di un dialogo, prende la parola e la mantiene quasi in esclusiva. I tre attori coinvolti sono stati poi sottoposti all'intervista, che ha riguardato il loro lavoro e il loro approccio alle battute di un copione, in particolare alla scrittura di Pirandello. Anche in questo caso, nonostante la natura dialogante e il contesto conversazionale, in linea con gli studi precedenti (p.es., Loehr 2004), dai circa 30 minuti complessivi registrati ne sono stati estrapolati 10 in totale in cui per ciascun parlante fossero compresi: una sequenza di flusso verbale di

[3] Per un approfondimento di processi, metodi e risultati si rimanda a Cantalini 2018; Cantalini et al. 2020; Cantalini & Moneglia 2020; Cantalini & Moneglia 2022.

[4] Luigi Pirandello: "Il giuoco delle parti" (1918. Mondadori: 1993 – Atto I, Scena Terza).): si tratta della scena in cui il protagonista, Leone Gala, discorre dei casi della vita utilizzando la metafora dell'uovo e del suo guscio: in breve il secondo (il guscio svuotato) è l'auspicabile per quanto cinicamente indifferente atteggiamento da assumere nella vita di contro alla forza distruttiva delle passioni (l'uovo dentro il guscio).

almeno 30 secondi consecutivi (di fatto ogni turno registrato è stato mediamente più lungo); la porzione analizzata non avesse tagli al suo interno ed esponesse almeno una risposta completa ad una delle domande dell'intervista strutturata (abbiamo estrapolato in realtà una media di ben due turni per parlante); ci fosse evidente coinvolgimento gestuale nelle selezioni estrapolate; e infine il materiale prodotto fosse prevalentemente di tipo monologico. In questo modo la comparabilità delle registrazioni è stata data dall'identità testuale nel recitato, da quella individuale tra spontaneo e recitato, dalla tipologia di contesto comunicativo per lo spontaneo; e per entrambe le varietà si sono selezionate porzioni testuali monologiche interne a dialoghi. Ai parlanti è stato detto che sarebbero stati osservati da un punto di vista 'linguistico' e 'comunicativo', con ciò intendendo che non avremmo in nessun modo affrontato o valutato aspetti di natura artistica, bensì aspetti definibili come 'grammaticali'. In questo modo si è evitato di menzionare la gestualità per evitare che ciò influenzasse il loro comportamento e al contempo li si è rassicurati sul fatto che in nessun modo la nostra ricerca avrebbe riguardato la loro qualità di attori, così da condizionarne il meno possibile l'interpretazione.

L'elicitazione di dati linguistici multimediali che offrano un flusso conversazionale il più naturale possibile[5] è questione metodologica delicata che richiede particolari accorgimenti, come già testimoniato in letteratura. Molti e diversi sono stati gli espedienti messi in atto per ottenerlo: la mediazione di un secondo interlocutore che non fosse l'intervistatore nel caso di interviste a singoli[6]; predisposizione di set in cui i parlanti fossero composti da gruppi di amici[7]; riprese effettuate in luoghi pubblici e familiari per gli intervistati[8]. Anche gli argomenti scelti di cui parlare possono favorire un flusso non trattenuto: parlare del proprio lavoro di attori come nel nostro caso è stato sicuramente stimolante per gli intervistati; Loehr (2004) ha lasciato che la conversazione tra più parlanti coinvolti, tra di loro amici, prendesse piede da sola, estrapolandone i momenti di maggior coinvolgimento gestuale. Tutto ciò perché, è innegabile, la sola telecamera, così come la relazione con un interlocutore sconosciuto e investito del ruolo di

[5] "conversation flow as natural as possible" in Loehr 2004: 72.
[6] McNeill 2005.
[7] Loehr 2004.
[8] Kendon 1972.

'osservatore esterno', inibisce e condiziona il comportamento di un parlante, influenzando e bloccandone soprattutto la gestualità[9]. E' perciò importante dare tempo alla conversazione di svilupparsi, così da permettere ai parlanti di dimenticarsi della presenza della telecamera, e di intervistarli assecondandone i momenti in cui appaiono particolarmente coinvolti in quanto stanno dicendo. In queste specifiche condizioni di acquisita familiarità rispetto al contesto estraneo e di investimento nel proprio pronunciamento, la gesticolazione non è mai davvero trattenuta ed è possibile osservarla in quanto naturalmente utilizzata. Inoltre è bene far sedere gli intervistati su sedie con spalliere, ma senza braccioli: in assenza di appoggi cui aggrapparsi per 'tenere sotto controllo la gesticolazione' le braccia inevitabilmente inizieranno ad accompagnare in maniera spontanea l'elocuzione[10]. Ed è bene che non siano fatti sedere su poltrone perché la loro schiena resti libera di muoversi, non affondando all'indietro. L'excursus gestuale, nella nostra osservazione, arriva alle mani partendo dal torso ed espandendosi lungo le braccia. Il modo di far star seduto il parlante, con i succitati accorgimenti, può favorire pertanto a nostro avviso sia l'ampiezza del gesticolare (e con l'ampiezza una più compiuta formalizzazione del gesto favorendone l'interpretazione), che il completo rilasciamento del movimento gestuale.

3. *Modelli teorici*

Per annotare il dato raccolto ci si è mossi all'interno di due quadri teorici, per la prosodia e per il gesto, teoricamente e proceduralmente comparabili. In realtà la scelta del modello prosodico da utilizzare in relazione all'analisi gestuale è questione cruciale e variamente affrontata in letteratura. Se il modello di struttura dell'excursus gestuale primariamente individuata da Kendon[11] è stata di fatto confermata negli studi successivi e non ha subito sostanziali modifiche, i quadri teorici prosodici invece non sono unitari né nell'approccio né nella descrizio-

[9] È tuttora culturalmente diffuso il concetto che gesticolare quando si parli sia sbagliato e diastraticamente di basso livello.

[10] Purtroppo con il parlante "U", attore di chiara fama intervistato nel suo camerino, è stato impossibile chiedergli di tenere completamente staccate le braccia dagli appoggi intorno a lui.

[11] Kendon 1972.

ne. D'altro canto gli studi sul gesto sono stati per lungo tempo ostacolati dall'assenza di strumentazione adeguata, mentre quelli prosodico fonetici sicuramente si sono avvalsi molto presto dei più diffusi strumenti di registrazione solo sonora. Così da una parte il gesto ha fornito da subito un modello standard di analisi, ma non le opportunità per la sua indagine; dall'altra il parlato è stato oggetto di approfondite indagini, ma in assenza di uno standard fonetico di annotazione condiviso e universale. Tutto ciò fa sì che ad oggi gli studi a disposizione sulla sincronizzazione gesto-parlato non siano realmente sovrapponibili per quanto riguarda le unità linguistiche prese a riferimento[12].

Stante l'eterogeneità dei lavori precedenti, la nostra scelta è stata quella di mettere insieme il modello gestuale di Kita et al.[13], che sistematizza le fasi del movimento gestuale, già identificate in unità configurazionali, secondo un sistema di regole sintagmatiche all'interno di una gerarchia a inclusione, e affiancare ad esso un modello teorico prosodico analogo per approccio descrittivo e procedurale: il modello L-AcT[14], anch'esso configurazionale, che postula unità linguistiche percettivamente riconoscibili e altrettanto gerarchicamente organizzate a inclusione. In particolare, in entrambi i modelli si osserva un segmento nucleare necessario e sufficiente costituito da un picco d'intensità (acustica o cinetica) intorno al quale si dispongono facoltativamente ulteriori unità strutturalmente differenti, nell'insieme delle quali, unità facoltative più picco, sono articolate le unità superiori.

Nel modello gestuale il nucleo realizza la fase semioticamente attiva del gesto ed è denominato *stroke*. Prima e dopo lo stroke è possibile trovare fasi di tenuta o *hold* (per esteso, *pre-stroke hold* se prima dello stroke; e *post-stroke hold* se dopo[15]) che ne estendono la portata espressiva, realizzando un segmento composito invece che singolo denominato *expressive phase*, Prima e dopo il solo stroke o la più composita expressive-phase, possono occorrere le fasi di preparazione (*preparation*) e ritrazione (*retraction*), le cui denominazioni indicano i segmenti di preparazione della mano a realizzare il picco o di successivo e progressivo rilassamento dopo averlo realizzato. Con *fase* si

[12] Kendon 1972, 1980; McClave 1981; Loehr 2004; Turk 2020.
[13] Kita et al. 1998.
[14] Language into Act Theory: Cresti 2000; Cresti & Moneglia 2018.
[15] L'*independent hold* è invece alternativo allo stroke (Kita et al. 1998: 28).

indica quindi per il gesto l'unità discreta più piccola, potenzialmente separabile, della gerarchia[16].

L'intera escursione del gesto intorno al picco d'intensità, sia se realizzato da un singolo stroke che articolato in più o tutte le possibili estensioni date dalle fasi menzionate, costituisce il livello superiore della gerarchia o sintagma gestuale (*gesture phrase*). Il sintagma a sua volta può essere contenuto in un'unità superiore variamente articolata, l'unità gestuale (*gesture unit*), che si realizza ogni qualvolta la mano lascia una superficie di appoggio (*home position*) per poi ritornarvi. Una volta appoggiata, ma non ancora completamente ferma, la mano è nella fase di riposo (*rest position*), considerata nel presente lavoro unità attiva e come tale inserita nella gerarchia. A livello percettivo quindi il gesture phrase è riconoscibile grazie al suo centro (il culmine del movimento), mentre la gesture unit lo è grazie alla chiara individuazione delle sue estremità (insorgenza del movimento e sua chiusura). I tre livelli possono articolarsi differenziati o coincidere (per esempio: uno stroke che parta dalla superficie d'appoggio e vi ritorni subito dopo)[17].

A partire dal modello l'annotazione del gesto procede allora predisponendo sul foglio informatico del software Elan[18] (Wittenburg *et al*. 2006) organizzato in griglia a partire dalle indicazioni date da LASG[19] (Bressem *et al*. 2013), tre righe sovrapposte e concomitanti, in ciascuna delle quali segnare inizio e fine delle unità sovradescritte: al livello superiore iniziando dalle gesture unit, per poi dettagliare al secondo i gesture phrase, e infine riportando al terzo le gesture phase.

L'annotazione di inizio e fine unità (*break*) è alla base anche del modello linguistico-prosodico L-AcT. Postulando la corrispondenza tra unità prosodiche e unità linguistiche, L-AcT procede su due livelli individuando e annotando i *break prosodici terminali*, e successivamente al loro interno i *break prosodici* percepiti come non *terminali.*

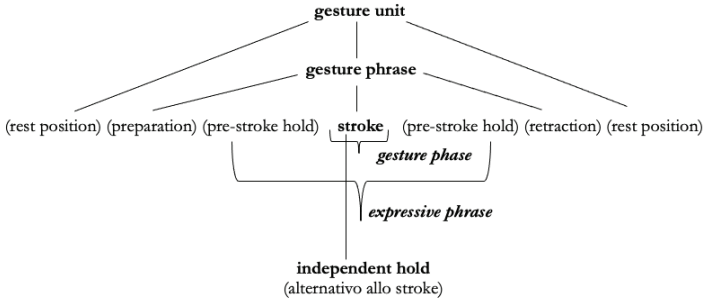
[16] Bressem et al. 2013.

[17] Figura 1.

[18] Si tratta di uno strumento professionale di annotazione e trascrizione di registrazioni audio o video manuali o semiautomatiche,

[19] Language Annotation System for Gestures.

Figura 1 - *La struttura del gesto*[20]



da Kita, van Gijn & van der Hulst 1998

L'annotazione di inizio e fine unità (*break*) è alla base del modello linguistico-prosodico L-AcT. Postulando la corrispondenza tra unità prosodiche e unità linguistiche, L-AcT procede su due livelli individuando e annotando i *break prosodici terminali*, e successivamente al loro interno i *break prosodici* percepiti come non *terminali.* Una sequenza conclusa da un break terminale (*pattern prosodico*) corrisponde a una unità di riferimento autonoma dal punto di vista pragmatico (*enunciato*); le unità delimitate da break non terminali (*unità prosodiche*) individuano le *unità informative*, in cui è segmentato il flusso del pensiero[21]. Il parlato consiste quindi di due livelli gerarchici prosodicamente riconoscibili, di cui il *comment* costituisce l'unità fondante. Analogamente allo stroke, il comment è unità necessaria e sufficiente alla realizzazione di un atto linguistico e può occorrere non accompagnato da ulteriori unità informative, nel qual caso abbiamo allora una neutralizzazione dei due livelli e la loro coincidenza in un enunciato non modellizzato (*unpatterned)*. Diversamente le unità informative intorno al comment possono essere variamente articolate realizzando enunciati modellizzati (*patterned*). Nel foglio di annotazione i due livelli saranno disposti su due righe predisposte che si aggiungeranno alle tre del gesto già menzionate.

[20] Il grafico è nostra raffigurazione del modello, le parentesi indicano opzionalità dell'elemento contenuto.

[21] *"flow of thought"* in Chafe 1998.

4. *Metodologia di annotazione*

L'annotazione multimodale è stata condotta separatamente, audio (parlato) e video (gesto), da due annotatori differenti. Il parlato è stato osservato senza accesso alle immagini, il gesto ascoltando il parlato ma non accedendo alla sua trascrizione. In letteratura la metodologia relativamente alla separatezza o meno delle annotazioni non è univoca. McNeill trascrive prima il parlato e al suo interno successivamente inserisce gli indici gestuali allo scopo di segnarne con esattezza l'occorrenza in relazione a singole sillabe o vocali. L'autore ribadisce quindi l'assoluta necessità di trascrivere le unità gestuali non sono ascoltando il parlato ma anche guardandone la trascrizione[22]. D'altro canto Cornelia Müller (2018: comunicazione personale) consiglia di osservare i gesti senza alcun accesso all'audio. Nel nostro caso volendo verificare l'eventuale sincronia tra modalità, abbiamo optato per l'approccio di McNeill come però attuato da Loehr (2004): osservare il gesto ascoltandone il parlato co-occorrente senza tuttavia avere accesso alla sua trascrizione, quindi senza conoscerne preventivamente i break prosodici.

La trascrizione gestuale viene fatta sul software Elan, quella prosodica su WinPitch (Martin 2004) e su PRAAT (Boersma 2001) e da PRAAT importata successivamente in Elan e allineata ai dati gestuali.

In Figura 2 è riportato il totale delle unità annotate[23].

Figura 2 - *Unità annotate (dati aggregati)*

	Durata	Unità terminate (parlato)	Unità informative (parlato)	Unità comunicative (no parole)	Gesture Units (gesto)	Gesture Phrases (gesto)	Gesture Units no unità linguistiche	Gesture Phrases no unità linguistiche	Durata interruzione gestuale
Spontaneo	10:05,802	107	461	1	56	336	2	8	01:13,787
Recitato	07:49,345	113	271	7	53	177	0	8	02:22,136

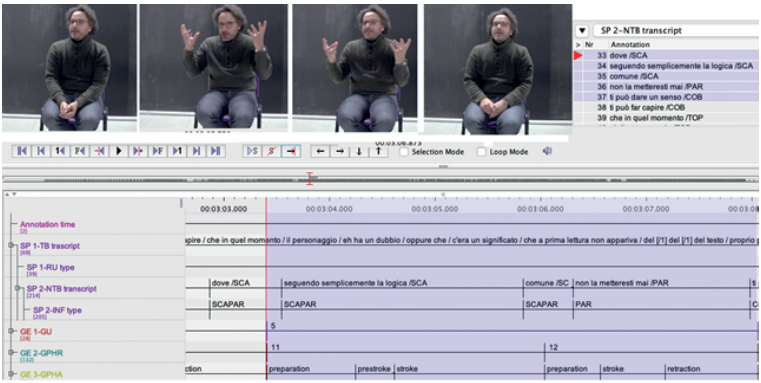
Una volta riconciliate le due annotazioni in Elan si è proceduto a verificare inizio e fine delle unità gestuali in relazione ai break prosodici

[22] In McNeill 2005: Appendix.

[23] I lavori precedenti presentano campioni osservati di lunghezza sensibilmente inferiore. L'annotazione gestuale richiede circa un'ora di lavoro per secondo di dato ed è quindi estremamente dispendiosa, pur con la strumentazione attuale. Loehr (2004) descrive 164 secondi e 147 gesti; Kendon (1972) 90 secondi: McClave (1991) 147. Il nostro lavoro cerca di offrire uno sguardo insolitamente ampio sul comportamento gestuale offrendo spezzoni di discorso per un totale di circa 17 minuti.

terminali e non-terminali, rilevando tendenze e modalità di sincronizzazione: inizio e fine delle gesture unit in relazione sia agli enunciati che alle unità informative[24]; ugualmente con i gesture phrase; di entrambi si è contato se e quante volte gli arrivi gestuali oltrepassassero i break sia terminali che non; infine rispetto agli stroke si sono condotte osservazioni tipologiche e qualitative che evidenziassero le corrispondenze e la frequenza di occorrenze rispetto alle differenti occorrenze di unità informative (queste ultime date da un set chiuso di circa 20 tipi, funzionalmente divisi tra testuali, dialogici, e privi di valore informativo). Figura 3 mostra i livelli di annotazione di una gesture unit, da posizione di riposo a posizione di riposo con due stroke, e come il gesto si allinea alla segmentazione del parlato in unità informative.

Figura 3 - *Sincronizzazione gesto prosodia in ELAN*



5. *Validazione*

Il processo di validazione prosodica è attestato in Moneglia et al. 2010.

Il processo di validazione gestuale sviluppato nella presente ricerca ha visto la replica del 20% dell'annotazione da parte di un secondo annotatore esperto, lo studio qualitativo dei disaccordi, e la riconduzione di questi a consenso producendo criteri condivisi di annotazione (Cantalini *et al*. 2020)[25].

[24] L'avvenuta corrispondenza è stata definita all'interno di un margine di 200 ms prima e dopo la fine dell'unità linguistica.

[25] Per maggiori dettagli sugli aspetti qualitativi discussi si veda il volume indicato.

Si è quindi proceduto alla validazione quantitativa. Il tasso di sovrapposizione delle unità e quello di categorizzazione sono stati calcolati attraverso le seguenti metriche standardizzate: "overlap /extent value", per gesture unit e gesture phrase, che ha mostrato un valore di 0.83 per le gesture unit e 0.61 per i gesture phrase; e "modified Cohen's Kappa", per le gesture phase, calcolata secondo le modalità descritte in Holle & Rein (2013), che negli stroke è risultata di circa 0.85. I risultati ottenuti sono paragonabili con gli unici altri dati pubblicati sulla validazione dell'annotazione gestuale (Dvoretska et al. 2013; per le metriche: Holle & Rein 2013 e Helmich & Lausberg 2014).

7. *Evidenze osservate e produttività dell'approccio adottato*

Il corpus raccolto e la metodologia attuata per osservarlo hanno rilevato le seguenti evidenze di sincronizzazione tra modalità e di variazione tra varietà, mostrando rilevante efficacia di analisi.

Per quanto riguarda la sincronizzazione tra gesto e unità linguistiche nel parlato spontaneo, sia le unità linguistiche terminate che quelle non terminate *segnano confini significativi* per l'allineamento delle unità gestuali: le gesture unit possono abbracciare più gesture phrase ed estendersi su più unità linguistiche; le gesture unit tendono a chiudere sui confini delle unità prosodiche terminali, mentre i gesture phrase su quelle non terminali; le gesture unit attraversano frequentemente i confini delle unità prosodiche terminali, mentre i gesture phrase quasi mai; lo stroke di conseguenza si posiziona sempre all'interno dell'unità di informazione. Tutto ciò oltre a mostrare corrispondenze sistematiche tra gesto e parlato evidenzia differenze di comportamento statisticamente significative tra i due livelli gestuali unit e phrase.

D'altra parte *la presenza di un dato normativo nello spontaneo consente di evidenziare le pecularità del recitato* sia per quanto riguarda il parlato che per il gesto. È riscontrabile infatti una diversa incidenza, nelle due varietà, di enunciati semplici, cioè di realizzazioni in cui i due livelli linguistici coincidono e il livello informativo è schiacciato in quello illocutivo: nel parlato recitato vi sono un maggior numero di enunciati semplici, sia scanditi che non (cioè accompagnati da unità prosodiche non terminate prive di valore informativo), mostrando una preferenza dei parlanti per l'interpretazione illocutiva rispetto a quella

informativa. Analogamente a livello del gesto le gesture unit tendono a coincidere con i gesture phrase e quindi a non abbracciare più porzioni di discorso bensì ad allinearsi con enunciati quando non addirittura con unità informative. Come con il parlato, anche per il gesto le strategie modellizzanti risultano sensibilmente ridotte.[26]

8. *Conclusioni*

Il modello di annotazione gesto – unità prosodiche (terminate e non terminate) approntato permette di studiare le modalità di sincronizzazione significative tra parlato e gesto. L'annotazione permette altresì di rilevare le differenze significative nelle modalità di sincronizzazione tra varietà (in questo caso tra spontaneo e recitato) e può restituire dati rilevanti di corrispondenze tra occorrenze gestuali e linguistiche. Consente inoltre avanzamenti nello studio della semantica del gesto.

Ci preme sottolineare come, con riguardo all'analisi linguistica, in molti casi di ambivalenza di interpretazione, sia proprio il gesto che può aiutare a disambiguare l'informazione linguistica e a contribuire ad una sua più accurata identificazione.

Riferimenti bibliografici

Boersma, Paul. 2001. Praat, a system for doing phonetics by computer. *Glot International* 5(9/10). 341–345.

Bressem, Jana & Ladwig, Silva H. & Müller, Cornelia. 2013. Linguistic Annotation System for Gestures (LASG). In Müller, Cornelia, Cienki, Alan, Fricke, Ellen, Ladewig, Silva H., McNeill, David & Teßendorf, Sedinha (a cura di), *Body – Language – Communication: An International Hand-book on Multimodality in Human Interaction*. Handbooks of Linguistics and Communication Science 38(1), 1098–1124. Berlin: De Gruyter Mouton.

Cantalini, Giorgina. 2018. *La gestualità co-verbale nel parlato spontaneo e nel recitato. Annotazione del gesto e correlati prosodici in campioni comparabili di attori italiani*. Roma: Università Roma Tre. (Tesi di dottorato.)

Cantalini, Giorgina & Moneglia, Massimo & Gagliardi, Gloria & Proietti, Morgana. 2020. La correlazione gesto/prosodia e la sua variabilità: il par-

[26] Per una disamina più approfondita dei risultati si veda Cantalini & Moneglia 2022.

lato spontaneo di contro alla performance attorale. In De Meo, Anna & Dovetto, Francesca M. (a cura di), *La Comunicazione Parlata / Spoken Communication. Napoli 2018*, 63–88. Roma: Aracne.

Cantalini, Giorgina & Moneglia, Massimo. 2020. The annotation of gesture and gesture / prosody synchronization in multimodal speech corpora. *Journal of Speech Sciences* (9): 07-30.

Cantalini, Giorgina & Moneglia, Massimo. 2022. Reduction of gesticulation and information patterning strategies in acted speech. In Fernandes, Carla & Evola, Vito & Ribeiro, Cláudia (a cura di), *Dance Data, Cognition, and Multimodal Communication*, 346 -362. London–New York: Routledge.

Chafe, Wallace. 1998. Language and the Flow of Thought. In Tomasello, Michael (a cura di), *The New Psychology of Language. Cognitive and Functional Approaches to Language Structure*, 87–104. New York: Routledge.

Cresti, Emanuela. 2000. *Corpus di italiano parlato*. Firenze: Accademia della Crusca, Firenze.

Cresti, Emanuela & Moneglia, Massimo. 2018. The illocutionary basis of Information Structure. Language into Act Theory (L-AcT). In Adamou, Evangelia & Haude, Katharina & Vanhove, Martine (a cura di), *Information structure in lesser-described languages*: *Studies in prosody and syntax*, 359-401. Amsterdam: Benjamins.

Dvoretska, Daniela & Petermann, Kerstin & Skomroch, Harald. 2013. Calculating Temporal Interrater Agreement for Binary Movement Categories. In Lausberg, Hedda (a cura di), *Understanding Body Movement. A Guide to Empirical Research on Nonverbal Behaviour (With an Introduction to the NEUROGES Coding System)*, 253-260. Frankfurt am Main, Peter Lang Verlag.

Ekman, Paul, & Friesen, Wallace V. 1969. The repertoire of nonverbal behavior. *Semiotica* 1. 49–98.

Helmich, Ingo & Lausberg, Hedda. 2014. Hand Movements with a Phase Structure and Gestures that Depict Action Stem from a Left Hemispheric System of Conceptualization. *Experimental Brain Research* 232(10). 3159–3173.

Holle, Henning & Rein, Robert. 2013.. The Modified Cohen's Kappa: Calculating Interrater Agreement for Segmentation and Annotation. In Lausberg, Hedda (a cura di), *Understanding Body Movement. A Guide to Empirical Research on Nonverbal Behaviour (With an Introduction to*

the NEUROGES Coding System), 261–275. Frankfurt am Main: Peter Lang Verlag.

Kendon, Adam. 1972. Some relation between body motion and speech: An analysis of an example. In Siegman, Aron W. & Pope, Benjamin (a cura di), *Studies in Dyadic Communication,* 177–210. New York: Elsevier.

Kendon, Adam. 1980. Gesticulation and speech: Two aspects of the process of utterance. In Ritchie Key, Marie (a cura di), *The Relationship of Verbal and Nonverbal Communication and Language*, 207–227. The Hague: Mouton.

Kendon, Adam. 2004. *Gesture: Visible Action as Utterance*. Cambridge: Cambridge University Press.

Kita, Sotaro & van Gijn, Ingeborg & van der Hulst, Harry. 1998. Movement phases in signs and co-speech gestures, and their transcription by human coders. In Wachsmuth, Ipke & Fröhlich, Martin (a cura di), *Gesture and Sign Language in Human-Computer Interaction,* 23–35. Berlin: Springer-Verlag.

Loehr, Dan. 2004. *Gesture and Intonation*. Ph.D dissertation. Washington, DC: Georgetown University.

Maricchiuolo, Fridanna. 2017. La comunicazione non verbale. Caratteristiche e funzioni. *The Inquisitive Mind* 12.

Martin, Philippe. 2004. WinPitch Corpus, a text to speech alignment tool for multimodal corpora. In Lino, Maria Teresa & Xavier, Maria Francisca & Ferreira, Fatima & Costa, Rute & Silva, Raquel (a cura di), *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04),* 537–540. ELRA.

McClave, Evelyn Z. 1991. *Intonation and Gesture.* Ph.D. dissertation. Washington, DC: Georgetown University.

McNeill, David. 1985. So you think gestures are nonverbal? *Psychological Review* 92(3). 350–371.

McNeill, David. 1992. *Hand and Mind: What Gestures Reveal about Thought*. Chicago: University of Chicago Press.

McNeill, David. 2005. *Gesture and Thought*. Chicago: University of Chicago Press.

Moneglia, Massimo & Raso, Tommaso & Malvessi-Mittmann, Maryualê & Mello, Heliana. 2010. Challenging the perceptual relevance of prosodic breaks in multilingual spontaneous speech corpora: C-ORAL-BRASIL / C-ORAL-ROM. In *5th International Conference on Speech Prosody 2010.* International Speech Communications Association.

Müller, Cornelia. 2018. Personal communication. In *8th Conference of the International Society for Gesture Studies: Gesture and Diversity. Cape Town.*

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

Pirandello, Luigi. 1993[1918]. *Il giuoco delle parti*. Milano, Mondadori.

Turk, Olcay. 2020. *Gesture, Prosody and Information Structure Synchronisation in Turkish*. Victoria University of Wellington. (Tesi di dottorato.)

Wittenburg, Peter & Brugman, Hennie & Russel, Albert & Klassman, Alex & Sloetjes, Han. 2006. ELAN: A Professional Framework for Multimodality Research. In Calzolari, Nicoletta & Choukri, Khalid & Gangemi, Aldo & Meegaard, Bente & Mariani, Joseph & Odijk, Jan & Tapias, Daniel (a cura di), *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06),* 1556-1559. ELRA.

Web Sites

ELAN https://www.mpi.nl/corpus/html/elan/index.html

NEUROGES https://neuroges.neuroges-bast.info/

PRAAT http://www.fon.hum.uva.nl/praat/

WinPitch https://www.winpitch.com/

DOI: 10.17469/O2106SLI000009

FEDERICA COMINETTI, LORENZO GREGORI, EDOARDO LOMBARDI VALLAURI, ALESSANDRO PANUNZI

IMPAQTS: un corpus di discorsi politici italiani annotato per gli impliciti linguistici

Il contributo presenta il corpus multimodale di discorsi politici italiani IMPAQTS e lo schema di annotazione pragmatica ad esso applicato. Il corpus è costruito per essere rappresentativo del linguaggio politico della Repubblica italiana: nel contributo se ne descrivono i criteri di bilanciamento e i metadati. Il corpus è inoltre annotato per contenuti discutibili veicolati implicitamente. Dello schema di annotazione pragmatica si illustrano i presupposti teorici e i correlati applicativi. Infine, si forniscono alcune indicazioni qualitative e quantitative emerse dall'analisi di una prima tranche di 477 discorsi annotati.

Parole chiave: discorso politico, corpus multimodale, annotazione pragmatica, impliciti linguistici.

1. *Introduzione*[1]

In questo contributo presentiamo il corpus IMPAQTS e lo schema di annotazione pragmatica ad esso applicato. La realizzazione e annotazione del corpus sono uno degli obiettivi del PRIN IMPAQTS (*Implicit Manipulation in Politics – Quantitatively Assessing the Tendentiousness of Speeches*) finanziato dal MIUR per il triennio 2020-2022, che coinvolge l'Università Roma Tre (PI: Edoardo Lombardi Vallauri) e l'Università di Firenze. Il contributo è organizzato come segue: in §2 si dà una descrizione del corpus, completa di criteri di bilanciamento e metadati; in §3 si descrive lo schema dell'annotazione pragmatica; in §4 è presentata l'interfaccia per le ricerche online; in §5 si propongono alcuni esempi e considerazioni derivati dall'analisi di una prima tranche annotata del corpus.

[1] L'articolo è frutto della collaborazione di tutti gli autori. Federica Cominetti ha scritto le sezioni 1, 2, 3 e 5. Lorenzo Gregori ha scritto la sezione 4.

2. *Descrizione del corpus*

L'idea alla base della progettazione e realizzazione del corpus IMPAQTS è quella di raccogliere una vasta collezione di discorsi che siano rappresentativi del linguaggio politico della storia della Repubblica Italiana. Mentre scriviamo, l'opera di raccolta del corpus è ancora in corso, ma siamo in grado di darne una accurata descrizione sulla base del *corpus-design* e dei significativi progressi portati avanti nei primi due anni del progetto, che permettono di prevedere che le tempistiche previste (metà 2023) saranno rispettate.

IMPAQTS conterrà 1500 discorsi pronunciati da politici italiani tra il 1946 e il 2022. I discorsi hanno una lunghezza media di circa 10mila caratteri, con una certa variabilità[2]. Quando completo, il corpus avrà dunque una consistenza di circa 12,5 milioni di caratteri, corrispondenti a circa 3,5 milioni di parole. Questa notevole consistenza renderà IMPAQTS non solo il più grande corpus di discorso politico italiano, ma anche il più grande corpus di italiano parlato controllato e uno dei più grandi corpora di italiano parlato *tout court*.

IMPAQTS è pensato come corpus multimodale. La trascrizione di ogni discorso è infatti allineata enunciato per enunciato al file video corrispondente, con l'eccezione di alcuni discorsi degli anni Sessanta e Settanta, per cui è disponibile solo il file audio, e di alcuni discorsi degli anni Quaranta e Cinquanta, per cui non è disponibile neppure l'audio. Nonostante si tratti di un corpus di parlato, per facilitarne la fruizione si è scelto di adottare nelle trascrizioni uno stile di punteggiatura ortografico, pur con alcune semplificazioni.

La punteggiatura segue tuttavia un criterio prosodico, più che sintattico, e il punto fermo tende a coincidere con intonazione conclusiva di enunciato. Oltre alla trascrizione con punteggiatura convenzionale, è peraltro disponibile una versione del corpus annotata prosodicamente per break terminali e non terminali secondo le convezioni Lablita/C-ORAL-ROM (Cresti & Moneglia 2005). Il corpus sarà automaticamente lemmatizzato e annotato per parti del discorso (*POS-tagging*).

Ogni discorso del corpus è classificato secondo due tipi di metadati: informazioni sociolinguistiche relative al parlante e informazioni relative al discorso.

[2] Il discorso più breve è di 921 caratteri, il più lungo di 37874, la media di 9654.

Le informazioni sociolinguistiche relative al parlante includono i seguenti parametri[3]: *nome e cognome; sesso; età; ruolo politico; forza politica; orientamento politico.*

Il metadato relativo al ruolo politico è espresso nel modo più dettagliato possibile, tenendo conto della eventuale sovrapposizione di più cariche. Ad esempio, quasi tutti i discorsi di Enrico Berlinguer presenti nel corpus recano alla voce "ruolo politico" la doppia dicitura *Deputato* e *Segretario del Partito*.

Il parametro "orientamento politico" nasce invece per semplificare le ricerche, dal momento che sotto la voce "forza politica" sono annoverate addirittura 72 voci diverse, che rendono conto della frammentazione partitica che ha interessato la storia della Repubblica. Si è dunque deciso di aggiungere al dato dettagliato relativo al partito quello più ampio dell'orientamento politico, una classe chiusa con sei possibili realizzazioni: sinistra, centro-sinistra, centro, centro-destra, destra, indipendente. L'estrema volatilità delle appartenenze partitiche è anche la ragione per cui la forza politica viene annotata per ogni discorso, e non viene "automaticamente" assegnata a ogni singolo politico[4].

I metadati relativi al discorso – seguiti dalle loro possibili realizzazioni – sono i seguenti:

- *tipo*: assemblea, comizio, riunione di partito, dichiarazione trasmessa, dichiarazione in presenza, dichiarazione nuovi media (cfr. sotto per dettagli)*;*
- *canale*: in presenza, trasmesso, nuovi media;
- *pubblico:* pubblico generale, istituzioni, sostenitori, colleghi di partito;
- *struttura:* assemblea ufficiale, comizio, riunione di partito, dichiarazione;
- *numero di caratteri:* della trascrizione, inclusi gli spazi.

Tra quelli appena descritti, il parametro più importante per la classificazione dei discorsi è il *tipo*, che deriva dalla constatazione che i tre parametri di *canale*, *pubblico* e *struttura* non sono indipendenti tra loro,

[3] Tutti i dati sociolinguistici sono da intendersi con riferimento al momento dell'enunciazione del discorso.

[4] Un caso paradigmatico è quello di Francesco Rutelli, che nella sua storia politica ha fatto parte di 9 partiti diversi.

ma si presentano in combinazioni tipiche. Le possibili realizzazioni del tipo di discorso sono qui dettagliate:

- *assemblea* (discorso pronunciato in presenza davanti alle istituzioni, tipicamente in una seduta parlamentare o di una istituzione locale – Consiglio Regionale, Consiglio Comunale, ecc. –);
- *comizio* (discorso pronunciato in presenza davanti a un pubblico prevalentemente di sostenitori, tipicamente in campagna elettorale o in occasione di un evento pubblico);
- *riunione di partito* (discorso pronunciato in presenza davanti a un pubblico di colleghi di partito, tipicamente in occasione di un congresso o assemblea di partito);
- *dichiarazione trasmessa* (discorso pronunciato non in presenza di un pubblico di ascoltatori ma per la trasmissione video/audio, destinato al pubblico generale; es. messaggio di fine anno del Presidente della Repubblica, videomessaggio autofinanziato o trasmesso dal servizio pubblico in occasione di campagna elettorale o referendaria);
- *dichiarazione in presenza* (discorso pronunciato davanti a un pubblico generico o di giornalisti; es. dichiarazioni rilasciate in conferenza stampa);
- *dichiarazione nuovi media* (discorso registrato o trasmesso tramite nuovi media, destinato a un pubblico prevalentemente di *followers*; es. dirette Facebook).

Dalla descrizione delle categorie riconducibili al "tipo" di discorso, si vede dunque come alcuni degli altri parametri risultino in parte ridondanti: selezionando discorsi del tipo "assemblea", ad esempio, si avranno solo discorsi con *canale* "in presenza", *pubblico* "istituzioni", *struttura* "assemblea ufficiale".

Ogni discorso è univocamente individuato da un codice alfanumerico di 9 caratteri, come ad esempio GALM79-A1. Le prime quattro lettere sono l'iniziale del nome del parlante seguita dalle prime tre iniziali del cognome. Nel caso di cognomi composti da più di una parola, la seconda lettera del codice è la prima lettera della prima parola del cognome, mentre la terza e la quarta lettera del codice sono la prima e la seconda lettera della seconda parola del cognome (ad es. GALM = Giorgio Almirante; EDNI = Enrico De Nicola; RRIE = Rosa Russo Iervolino). Seguono due numeri che rappresentano le ultime due cifre dell'anno di enunciazione del discorso (46 per

1946, 20 per 2020). Segue il segno "-" (meno). Segue una lettera che indica il tipo di discorso secondo la tabella seguente:

Tabella 1 - *Codici attribuiti ai diversi tipi di discorso*

Assemblea	A
Comizio	C
Riunione di partito	P
Dichiarazione trasmessa	T
Dichiarazione in presenza	D
Dichiarazione nuovi media	N

Il codice GALM79-A1 corrisponde pertanto a un discorso pronunciato da Giorgio Almirante nel 1979 durante un'assemblea ufficiale. Il numero finale disambigua eventuali sovrapposizioni nel codice (dovute alla presenza nel corpus di più di un discorso di un certo tipo pronunciato dallo stesso parlante nello stesso anno). Ad esempio, MLUP12-C1 e MLUP12-C2 indicano due discorsi pronunciati da Maurizio Lupi in occasione di comizi tenutisi entrambi nel 2012.

Come anticipato, il corpus è rappresentativo del linguaggio politico della Repubblica. I criteri di bilanciamento – in ordine gerarchico – sono i seguenti: numero di discorsi per politico; scansione temporale; tipo di discorsi.

Il primo criterio di bilanciamento è stato individuato nel *numero di discorsi per politico*, fissato a 10, cifra che permette un congruo spazio di variazione personale sincronica e in molti casi diacronica. Di conseguenza, i 1500 discorsi del corpus corrispondono a 150 personalità politiche.

Il secondo criterio di bilanciamento riguarda la *scansione temporale*. Da questo punto di vista, il corpus è divisibile in tre sotto-corpora temporali:

– *Sotto-corpus A* (discorsi dal 1946 al 1972);
– *Sotto-corpus B* (discorsi dal 1973 al 1993);
– *Sotto-corpus C* (discorsi dal 1994 al 2022).

Il punto di rottura 1972-1973 è determinato dal cambiamento dei temi affrontati in Parlamento, e conseguentemente del tono comunicativo, con le prime discussioni sul divorzio e l'aborto. Il punto di rottura 1993-1994 è invece individuato nel referendum per la riforma elettorale e nel conseguente avvio della cosiddetta "Seconda Repubblica".

I tre sotto-corpora non hanno uguale consistenza. Il sotto-corpus A contiene circa 170 discorsi, il sotto-corpus B circa 330 e il sotto-corpus C circa 1000. Pertanto, dal punto di vista diacronico il corpus segue le proporzioni di bilanciamento 1: 2: 6. Tale squilibrio rende conto del maggiore interesse dei discorsi più recenti per uno degli obiettivi del PRIN IMPAQTS, ovvero la diffusione nella società civile di uno spirito critico nei confronti degli intenti manipolatori dei politici.

La selezione dei politici inclusi nel corpus risponde in primo luogo alla scansione temporale suddetta: circa 100 dei 150 politici afferiscono alla fascia temporale più recente, 33 hanno operato prevalentemente nel periodo tra il 1973 e il 1992, e 17 nel periodo dal 1946 al 1973. Si è poi considerata l'esigenza di rispecchiare il più possibile la composizione del Parlamento (e in misura minore degli organi politici locali). Ad esempio, il sotto-corpus B comprende una maggioranza di rappresentanti della DC, una minoranza di esponenti del PCI e alcuni rappresentanti dei partiti più piccoli, mentre il sotto-corpus C rispecchia il sostanziale bipolarismo del ventennio 1994-2013 e la configurazione politica più sfaccettata dell'ultimo decennio. Sulla base di questi criteri, sono state selezionate le figure politiche più significative e note.

Per la selezione dei dieci discorsi di ogni politico si è adottato un criterio di bilanciamento basato sul tipo di discorso, presentato nella tabella seguente:

Tabella 2 - *Schema di bilanciamento per ogni politico*

Tipo di discorso	*Numero di discorsi*
A	4
T, D, N	3
C	2
P	1

Come si vede nella tabella, per ogni politico vengono inclusi nel corpus quattro discorsi pronunciati durante assemblee ufficiali, tre fra dichiarazioni trasmesse, dichiarazioni in presenza e dichiarazioni rilasciate tramite nuovi media, due discorsi pronunciati in occasione di comizi, e un discorso pronunciato in una riunione di partito. Naturalmente, i discorsi del tipo N (nuovi media) sono disponibili

solo per i politici appartenenti al sotto-corpus C che fanno uso dei social media, disponendo di un profilo Facebook o YouTube ufficiali, e rendono conto solo degli anni più recenti (in particolare dell'ultimo decennio). I politici che hanno operato in anni meno recenti e quelli che non usano i social media avranno dunque tre discorsi totali dei tipi D (dichiarazione in presenza) o T (dichiarazione trasmessa).

3. *Annotazione pragmatica*

Il corpus IMPAQTS costituisce la base di un progetto di annotazione pragmatica per contenuti impliciti tendenziosi. Tale progetto ha il duplice obiettivo di fornire indicazioni qualitative e quantitative rispetto all'uso manipolatorio degli impliciti da parte dei politici italiani, e di produrre un vasto catalogo di esempi autentici di impliciti dal contenuto discutibile.

Lo schema di annotazione è basato sul modello proposto da Lombardi Vallauri & Masia (2014), ampliato in Lombardi Vallauri (2019), e comprende quattro macrocategorie di implicito linguistico: implicature, vaghezza, presupposizioni, topicalizzazioni. È importante precisare che gli impliciti vengono annotati solo quando veicolano contenuti non *bona fide* veri, cioè informazioni che sarebbero più difficilmente credute vere se venissero asserite in modo esplicito. L'applicazione di questo parametro risponde alla necessità di distinguere l'uso manipolatorio degli impliciti dall'uso comunicativamente "onesto" che essi hanno ai fini di un'economia testuale: contenuti già introdotti nell'universo di discorso o già noti al ricevente non necessitano infatti di essere asseriti ex-novo, come nell'esempio seguente:

(1) *L'immigrazione clandestina è diventata un business e di fatto una nuova forma di finanziamento pubblico ai partiti.* [ADBA17-A1]

In (1) l'aggettivo *nuova* presuppone che esistano altre forme di finanziamento pubblico ai partiti, ma si tratta di un contenuto che può essere legittimamente dato per presupposto, e pertanto non viene annotato.

Per la categoria della vaghezza, una proprietà del linguaggio in qualche misura fisiologica, vengono annotate solo le espressioni vaghe che sono sfruttate per sottrarre all'attenzione dei destinatari det-

tagli che li porterebbero a riconoscere il messaggio come poco credibile o manipolatorio.

Le implicature presenti nel corpus sono annotate seguendo la diffusa classificazione in convenzionali, conversazionali generalizzate e conversazionali particolarizzate. A queste si aggiungono due etichette che indicano sottotipi di implicature conversazionali particolarizzate: le implicature da lista e quelle da metafora. Per la vaghezza, sono individuate tre categorie: vaghezza semantica, sintattica e da metafora. L'annotazione delle presupposizioni prevede la precisazione del tipo di attivatore, secondo uno schema dettagliato che comprende tutti i principali attivatori noti in letteratura, a cui se ne aggiungono altri che sono stati identificati nel corso del progetto[5]. La categoria del topic è annotata distinguendo attivatori prosodici e attivatori sintattici.

Ogni implicito tendenzioso è annotato tramite l'etichetta corrispondente ed è accompagnato dall'esplicitazione il più possibile dettagliata del contenuto implicito[6], come nell'esempio seguente:

(2) *[Signor presidente del Consiglio, è la seconda volta in 28 mesi che lei interviene alla Camera. La prima volta fu per l'insediamento, concluse allora il suo discorso con uno squillante "Viva il Parlamento"]IMPL_CVRS*
[PBER10-A1]
Implica che il presidente del Consiglio sia andato in Parlamento troppo di rado, e che ciò dimostri la sua scarsa considerazione dell'istituzione, contraddicendo le sue affermazioni.

Oltre all'indicazione della categoria e all'esplicitazione, ogni contenuto implicito è accompagnato dalla descrizione della *funzione comunicativa* svolta, secondo il modello proposto da Garassino, Brocca e Masia (2019, 2022). Ogni contenuto implicito è associato a una delle seguenti funzioni: attacco (TT), difesa (DI), autoelogio (AU), elogio (EL), opinione personale (OP). Ad esempio, la funzione comunicativa dell'implicatura contenuta in (2) è di attacco, pertanto il contenuto sarà accompagnato dall'etichetta TT.

L'annotazione pragmatica viene prodotta su ogni discorso del corpus da tre annotatori indipendenti attraverso la piattaforma multi-an-

[5] Lombardi Vallauri *et al.* 2021; Cominetti & Giunta 2022; Cimmino & Cominetti (*accettato*).
[6] Seguendo il suggerimento di Sbisà (2021).

notatore WebAnnoMM. Le tre versioni vengono poi confrontate da un curatore, che licenzia l'annotazione definitiva.

4. *Le ricerche online*

La consultazione dei dati del corpus IMPAQTS sarà possibile attraverso un'applicazione web dedicata (EMMACorp)[7], ad oggi in fase di sviluppo, che consentirà di effettuare ricerche linguistiche avanzate sul corpus.

EMMACorp è un software specifico per l'interrogazione di corpora linguistici multimodali, che affianca alle funzionalità avanzate di ricerca linguistica su corpora annotati la gestione di contenuti audio e video allineati. In particolare, per ogni ricerca effettuata sul corpus si ottengono, oltre al testo trascritto, ai metadati e alle annotazioni, anche i segmenti audio/video allineati. L'utente può quindi avere accesso diretto al parlato originale a fianco della sua trascrizione.

EMMACorp è basato su NoSketchEngine[8], la versione open-source del popolare strumento per l'interrogazione di corpora SketchEngine[9]. Il frontend dell'applicazione è realizzato appositamente per IMPAQTS con Angular[10] ed è progettato per fornire un'interfaccia moderna che sia in grado di gestire i contenuti multimodali. EMMACorp permette i seguenti tipi di ricerche:

– Ricerche sulla trascrizione: è possibile ricercare una parola o una sequenza all'interno di tutto il materiale trascritto;
– Ricerche sull'annotazione testuale: il corpus è suddiviso in token e annotato automaticamente per parti del discorso e lemmi; sono quindi possibili ricerche per questi due livelli, sia su singolo token, sia su sequenze;
– Ricerca di pattern linguistici: i pattern sono delle sequenze variabili di token specificate attraverso espressioni regolari in linguaggio CQL (Corpus Query Language[11]); EMMACorp consente di fare ricerche direttamente in CQL, oppure di definire pattern linguistici attraverso uno strumento grafico;

[7] *Engine for MultiModal Aligned Corpora*.
[8] https://nlp.fi.muni.cz/trac/noske
[9] https://www.sketchengine.eu/
[10] https://angular.io/
[11] https://www.sketchengine.eu/documentation/corpus-querying/

– Ricerche sull'annotazione pragmatica: è possibile fare query che coinvolgano tutte le categorie di implicito; gli impliciti possono essere ricercati direttamente, oppure come filtro per circoscrivere una ricerca testuale;
– Ricerche su metadati: tutte le ricerche possono essere filtrate utilizzando la serie di informazioni collegate a ogni discorso.

È importante notare che i vari livelli di annotazione possono essere combinati in un'unica query, consentendo, ad esempio, di ricercare una sequenza testuale o un pattern linguistico all'interno di un certo tipo di implicito in un gruppo di discorsi filtrati per metadato (ad esempio, pronunciati in un certo periodo storico, o appartenenti a una certa area politica).

I risultati delle ricerche vengono visualizzati in formato KWIC (Key Word In Context, Manning & Schütze 1999: 35) oppure per enunciato, di cui è disponibile il relativo segmento audio/video del discorso originale, direttamente riproducibile online.

EMMACorp, inoltre, mette a disposizione una serie di funzionalità per l'analisi statistica dei dati: ordinamento, campionamento, collocazioni, liste di frequenza.

5. *Prime analisi ed esempi*

Nel giugno 2021 è stata analizzata statisticamente una prima tranche di 477 discorsi annotati, per una consistenza di 1.664.514 parole e di 65.614 enunciati. In questa sezione sono state individuate 9160 implicature a contenuto discutibile (di cui 8664 conversazionali, con una frequenza di 550 ogni 100mila parole), 8014 presupposizioni (481 ogni 100mila parole), 1723 topicalizzazioni (103 ogni 100mila parole) e 4876 istanze di vaghezza (292 ogni 100mila parole).

Dal momento che la sezione del corpus analizzata non è bilanciata, non si possono fare considerazioni significative rispetto ai metadati, anche se alcune tendenze, che andranno verificate nel corpus bilanciato, sembrano emergere: nello specifico, la tendenza più evidente riguarda la variabile di genere, con le donne che paiono essere meno implicite degli uomini rispetto a tutte le categorie linguistiche prese in considerazione.

Un'altra tendenza che potrebbe trovare conferma statistica riguarda la frequenza dei diversi tipi di impliciti in relazione al tipo di discorso: i discorsi del tipo A (assemblee ufficiali) risultano essere in media i più impliciti rispetto a tutte le variabili, con l'eccezione della vaghezza, che è ancora più frequente nei comizi (tipo C). Questa tendenza si può spiegare con il fatto che i comizi spesso hanno luogo durante la campagna elettorale, un momento in cui i politici sono particolarmente propensi a fare promesse, spesso molto vaghe, per assicurarsi il consenso dell'elettorato, e ad attaccare gli avversari, anche in questo caso avvalendosi di strategie di vaghezza per proteggersi dalle conseguenze di attacchi troppo diretti. Si consideri il seguente esempio:

(3) *I prossimi cinque anni non saranno solo un confronto tra programmi diversi, ma sarà una battaglia politica, culturale, etica, programmatica, tra [chi non avendo idee per il futuro, vuole riportare indietro l'Europa]VAG_SEM, e chi invece vuole e scommette su [una nuova Europa del lavoro, del benessere, della scienza, della storia]VAG_SEM.*
[NZIN19-C1]

In (3) si osservano due passaggi etichettati come vaghezza semantica[12]: nel primo, il parlante attacca senza nominarli avversari che non hanno idee per il futuro e vogliono dunque riportare indietro l'Europa. L'accusa è evidentemente molto grave e probabilmente esagerata: se il parlante avesse detto: "Il partito X o il politico Y, non avendo idee per il futuro, vuole riportare indietro l'Europa su questi precisi aspetti dell'agenda politica", gli ascoltatori avrebbero più facilmente collegato questa affermazione alla realtà e avrebbero notato che l'accusa era esagerata o inesatta. Inoltre, le persone attaccate avrebbero probabilmente reagito, accusando l'oratore di calunnia. La vaghezza permette invece al parlante di instillare negli ascoltatori il timore che i suoi avversari non siano alternative affidabili, proteggendolo però da reazioni critiche. La seconda istanza di vaghezza invece lascia sottospecificato che cosa intenda il parlante per "nuova Europa del lavoro, del benessere, della scienza, della storia": in questo caso la vaghezza non serve a salvare la faccia del parlante ma a permettergli di prospettare uno scenario che risulti accattivante per un pubblico il più possibile

[12] In (3) vi sono per la verità altri contenuti impliciti, che non riportiamo per facilitare la lettura.

ampio. Ad esempio "Europa del lavoro" potrebbe significare un'Europa con meno disoccupazione, ma anche un'Europa con maggiori tutele e garanzie sul lavoro, oppure un'Europa che facilità la mobilità lavorativa, e così via.

Un'ulteriore considerazione qualitativa che emerge dall'analisi della prima tranche riguarda, soprattutto nel caso dei discorsi ufficiali (A), un possibile ruolo esercitato sulla frequenza degli impliciti manipolatori dall'appartenenza dell'oratore alla maggioranza o all'opposizione: sembra infatti delinearsi una tendenza dei politici all'opposizione a essere più impliciti di quelli di maggioranza[13]. Tale tendenza deriva forse dalla maggiore propensione dell'opposizione alla critica e all'attacco, operazioni comunicative socialmente "costose", che spesso si avvalgono del ricorso alle strategie di implicitezza. Si veda l'esempio seguente:

(4) *Presidente e colleghi. Sarebbe fin troppo facile per chi, come noi del Movimento sociale italiano, non le votò la fiducia nello scorso mese di aprile, mettere in evidenza che poi non avevamo tanto torto quando sostenevamo che da [un governo che nasceva unicamente [perché [la paura di andare anticipatamente alle urne] PPP_DDEF era più forte [del desiderio di farlo]PPP_DDEF] PPP_SUB] TOP_PROS+PPP_DINDEF, non c'era molto da aspettarsi.*
[GFIN92-A1]

L'esempio (4) contiene l'inizio di una replica di Gianfranco Fini all'intervento alla Camera del Presidente del Consiglio Giulio Andreotti. Fin dal primo enunciato, il testo è molto ricco di impliciti, anche "sovrapposti": le descrizioni definite "la paura di andare anticipatamente alle urne" e "il desiderio di farlo" presuppongono rispettivamente che la DC avesse paura e desiderio delle elezioni anticipate; la subordinata introdotta da "perché" presuppone che tale paura fosse più forte del desiderio; il costituente introdotto da "un governo" presuppone, tramite una descrizione indefinita anaforica[14], che il Governo Andreotti sia nato solo perché tale paura era più forte del desiderio; inoltre, lo stesso costituente indefinito risulta topicalizzato. Discorsi particolarmente ricchi di impliciti da parte di membri dell'opposizione, come quello da cui è tratto (4), compaiono spesso in situazioni di crisi di

[13] Tendenza emersa anche nell'analisi di un piccolo corpus multilingue di discorso politico, cfr. Cominetti *et al*. (*in valutazione*).
[14] Cfr. Lombardi Vallauri *et al.* (2021).

Governo o di fine legislatura, in cui la tendenza all'attacco sembra particolarmente marcata. Non mancano del resto nel corpus discorsi molto ricchi di impliciti tenuti da membri del Governo o da parlamentari di maggioranza, soprattutto in circostanze in cui è in discussione un provvedimento controverso, o qualche membro del Governo o della maggioranza viene accusato di qualcosa. L'esempio seguente è tratto dal discorso alla Camera di Aldo Moro in occasione della discussione relativa allo scandalo Lockheed:

(5) [*L'affidarsi a frammentarie notizie della lunga vicenda]PPP_DDEF+VAG_SINT, [il pensare che tutto sia stato già udito e compreso]PPP_DDEF+VAG_SINT, [[immaginarci in una sorta di situazione obbligata, in una posizione di partito, in una ragione di disciplina]VAG_SINT, [l'essere in una esigente corrente di opinione di partito] PPP_DDEF+VAG_SINT] IMPL_CVRS: tutto questo è in contraddizione, tutto questo è incompatibile con la funzione del giudicare che il nostro ordinamento, con una scelta che può essere discussa ma non disattesa, ci attribuisce.*
[AMOR77-A1]

L'esempio (5) contiene una interessante sequenza di infiniti nominali preceduti da articoli determinativi, che ne fanno delle descrizioni definite presuppositive. L'infinito nominale è inoltre una strategia di vaghezza sintattica, dal momento che permette la rimozione dell'agente: l'espressione "l'affidarsi a frammentarie notizie della lunga vicenda" non solo presuppone che qualcuno si affidi a notizie frammentarie, ma lascia vago chi lo faccia. L'enunciato contiene inoltre una implicatura conversazionale, dove Moro suggerisce che sulla vicenda Lockheed alcuni parlamentari non ragionino secondo giustizia ma seguendo direttive di partito.

Contesti parlamentari che sembrano al contrario collegati a uno scarso ricorso a strategie di implicitezza sono i casi in cui Ministri riferiscono al Parlamento su questioni di loro competenza e altre circostanze in cui l'esposizione è incentrata su fatti accaduti e vicende oggettive.

Gli esempi mostrati in questa sezione suggeriscono inoltre un altro dato che verrà verificato nel corpus bilanciato: la presenza, anche massiccia, di impliciti linguistici nei discorsi politici sembra essere un *fil rouge* che percorre l'intera storia della Repubblica.

Riferimenti bibliografici

Cimmino, Doriana & Cominetti, Federica. Italian *davvero* as a trigger of implicit contents in persuasive discourse. (accettato), in Edoardo Lombardi Vallauri, Laura Baranzini, Doriana Cimmino (eds.) The dynamic contribution of implicit meaning to the context: variability in real usage. *Special Issue of Journal of Pragmatics*.

Cominetti, Federica & Giunta, Giulia. 2022. Change of State and Factive Nominals and Nominalizations as Presupposition Triggers. *Italian Journal of Linguistics*, 34.1 (2022). 59-102.

Cominetti, Federica & Cimmino, Doriana & Coppola Claudia & Mannaioli, Giorgia & Masia, Viviana. Manipulative Effects of Implicit Communication: A Comparative Analysis of French, Italian and German Political Speeches. (in valutazione)

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

Garassino, Davide & Masia, Viviana & Brocca, Nicola. 2019. Tweet as you speak. The role of implicit strategies and pragmatic functions in political communication: Data from adiamesic comparison. *Rassegna Italiana di Linguistica Applicata (RILA)*, 2-3. 187-208.

Garassino, Davide & Brocca, Nicola & Masia, Viviana. 2022. Implicit categories and their pragmatic functions in a dimesic comparison. In Lombardi Vallauri, Edoardo & Cominetti, Federica & Masia, Viviana (eds.), *The persuasive and manipulative power of implicit communication*, *Journal of Pragmatics* 183 Special Issue.

Lombardi Vallauri, Edoardo & Masia, Viviana. 2014. Implicitness impact: Measuring texts. *Journal of Pragmatics 61*. 161-184.

Lombardi Vallauri, Edoardo. 2019. *La Lingua Disonesta*. Bologna: Il Mulino.

Lombardi Vallauri, Edoardo & Cominetti, Federica & Baranzini, Laura. 2021. Presupposing indefinite descriptions. In Lombardi Vallauri, Edoardo & Cominetti, Federica & Masia, Viviana (eds.), *The persuasive and manipulative power of implicit communication*, *Journal of Pragmatics* 183 Special Issue.

Manning, Chris & Schütze, Hinrich. 1999. *Foundations of Statistical Natural Language Processing*. Cambridge (MA), MIT Press.

Sbisà, Marina. 2021. Implicit meaning: varieties and functions. In Lombardi Vallauri, Edoardo & Federica Cominetti & Viviana Masia (eds.), *The persuasive and manipulative power of implicit communication*, *Journal of Pragmatics* 183 Special Issue.

DOI: 10.17469/O2106SLI000010

FRANCESCA M. DOVETTO, ALESSIA GUIDA, ANNA CHIARA PAGLIARO, RAFFAELE GUARASCI, LUCIA RAGGIO, ASSUNTA SORRENTINO, SIMONA TRILLOCCO

Corpora di Italiano Parlato Patologico dell'età adulta e senile*

Corpora di Italiano Parlato Patologico dell'età adulta e senile è un progetto che attualmente comprende tre diversi corpora di parlato non normofasico, ciascuno relativo a una patologia che non coinvolge l'età evolutiva e che quindi riguarda l'età adulta e/o quella senile.

Parole chiave: linguistica dei corpora, parlato, schizofrenia, malattia di Alzheimer, Mild Cognitive Impairment.

1. *Introduzione*

Il primo corpus del progetto raccoglie il parlato prodotto da pazienti ai quali è stata diagnosticata la malattia della schizofrenia, un disturbo mentale dell'età riproduttiva che si manifesta perlopiù nell'arco di un periodo che va dai 18 ai 35 anni, più frequente negli uomini rispetto alle donne (APA 2013). La schizofrenia è una malattia ubiquitaria, presente in tutte le parti del mondo e in tutti i contesti eco-ambientali e sociali nella medesima proporzione, è invariabile nel tempo e ha un tasso di incidenza superiore all'1% della popolazione mondiale.

Il secondo e terzo corpus riguardano invece forme di demenza tipiche dell'età senile: la demenza è una condizione di disfunzione cronica e progressiva delle funzioni cerebrali che porta a un declino delle facoltà cognitive della persona e che interessa dall'1 al 5% della popo-

* Al progetto "Corpora di Italiano Parlato Patologico dell'età adulta e senile", ideato e coordinato da Francesca M. Dovetto (Dovetto in press), hanno collaborato, in modi e momenti diversi, Amalia C. Bruni, Raffaele Guarasci, Alessia Guida, Valentina Laganà, Anna Chiara Pagliaro, Lucia Raggio, Simona Schiattarella, Assunta Sorrentino, Simona Trillocco. I co-autori sopra indicati hanno strettamente collaborato alla costruzione della versione DEMO dei tre corpora qui presentati.

lazione al di sopra dei 65 anni di età, con un'incidenza che raddoppia ogni quattro anni, arrivando a colpire il 30% della popolazione all'età di 80 anni (Alzheimer's Association 2016; Alzheimer's Disease International 2021). I due corpora sinora raccolti riguardano uno la malattia di Alzheimer, una sindrome dementigena prodotta da una patologia degenerativa diffusa del sistema nervoso centrale che comporta un progressivo decadimento delle funzioni cognitive, e l'altro riguarda invece una "condizione di rischio" definita come Disturbo Cognitivo Lieve (*Mild Cognitive Impairment*, MCI), ossia uno stadio di transizione tra quelle che sono le normali funzioni cognitive nell'invecchiamento fisiologico e la probabile insorgenza di una demenza di tipo Alzheimer.

Più in particolare, la malattia di Alzheimer rientra nella macro-categoria delle demenze a decorso degenerativo progressivo, pertanto l'esordio non è dovuto a un evento acuto – come nel caso delle demenze vascolari – bensì prevede una progressione lenta e una fase preclinica che può precedere anche di anni la comparsa di sintomi della sfera cognitiva. Il criterio principale per poter attribuire la diagnosi di fase clinica di demenza si fonda sulla perdita di autonomia nella gestione degli eventi della vita quotidiana. Le aree cognitive coinvolte sono: attenzione complessa, funzione esecutiva, apprendimento e memoria, linguaggio, percezione e cognizione sociale. La manifestazione della sintomatologia e la progressione della malattia presentano tuttavia un certo grado di variabilità tra i pazienti, dato sia da eventuali comorbilità, sia da fattori ambientali che possono fungere tanto da fattori di protezione quanto di rischio (Marini 2018; WHO 1993).

L'MCI (*Mild Cognitive Impairement*) rappresenta la fase prodromica delle malattie neurodegenerative progressive, nella quale il soggetto presenta, in una condizione di autonomia preservata, un declino delle funzioni cognitive leggermente maggiore di quello dovuto al naturale processo di invecchiamento – di cui si accorge lo stesso soggetto o qualche familiare – ma non ancora ascrivibile a un quadro patologico (Petersen 2004; Petersen *et al.* 2014). La sfera cognitiva solitamente più compromessa è quella della memoria, alla quale però si possono associare sintomi di declino relativi ad altri domini cognitivi, tra cui attenzione, funzioni esecutive e linguaggio. Sulla base del coinvolgimento della memoria si può distinguere tra MCI amnestico e non amnestico e sulla base del coinvolgimento di uno o più domini

cognitivi si può distinguere tra MCI a dominio singolo o multiplo. Tra le possibili manifestazioni di questa condizione, la tipologia MCI amnestico ha il più alto tasso di conversione in malattia di Alzheimer (60-70%); al contrario le forme non amnestiche possono evolvere in altre patologie neurodegenerative o – anche se in minima parte – regredire verso uno status non patologico (Sanford 2017).

I tre corpora hanno struttura aperta e sono in via di implementazione grazie a progetti in corso.

I corpora sono archiviati presso il Laboratorio scientifico LiSa, *Lingua e salute*, nell'Area di Ricerca "Processi e pratiche linguistiche" del Centro LUPT della Federico II; direzione di Francesca M. Dovetto.

2. *CIPPS*

Il primo corpus è denominato CIPPS: *Corpus di Italiano Parlato Patologico Schizofrenico*. Conta un totale di 17 ore di sonoro prodotto da 4 pazienti. 10 ore di registrazione sono state trascritte ortograficamente. Nel 2012, nel volume *Il parlar matto* a cura di Dovetto e Gemelli, sono state pubblicate le trascrizioni ortografiche insieme a una prima analisi multidisciplinare del corpus (dal punto di vista medico-clinico, filosofico, linguistico e psicolinguistico); nel 2013, una nuova edizione de *Il parlar matto* ha messo a disposizione la revisione delle trascrizioni insieme all'audio delle 10 sedute di psicoterapia. Il progetto di nuove acquisizioni, attualmente in corso, riguarda il parlato di pazienti schizofrenici farmacoresistenti.

Il corpus è stato raccolto, previo consenso informato dei pazienti e dei loro tutori legali, dal 2005 al 2007 e trascritto dal 2007 al 2012, grazie alla collaborazione tra la Scuola Sperimentale per la Formazione alla Psicoterapia e alla Ricerca nel Campo delle Scienze Umane Applicate della ASL Napoli 1, diretta dal dott. Carlo Pastore, e il Centro Interdipartimentale di Ricerca per l'Analisi e la Sintesi dei Segnali dell'Università Federico II di Napoli, allora diretto dal prof. Federico Albano Leoni.

In totale, il corpus comprende il parlato di quattro pazienti, indicati con le prime quattro lettere dell'alfabeto latino (A-B-C-D). Tutti i parlanti sono adulti (età 35-45), di sesso maschile e provenienti dall'area del napoletano. Parlano un italiano regionale, tendenzialmente dialettale, soprattutto il paziente D.

Ogni registrazione dura circa 60 minuti, pari a una intera seduta di psicoterapia. Per i pazienti A, C e D, la seduta comprende una prima parte di parlato letto, in cui il soggetto schizofrenico legge un testo da lui scritto in precedenza, e una seconda parte, più ampia, di parlato spontaneo in cui vengono affrontati diversi argomenti, come le abitudini quotidiane, soprattutto alimentari, e racconti di favole.

Il campione è costituito da pazienti in esordio (A), pazienti con patologia farmacoresistente (B) e pazienti cronici che non seguono terapia farmacologica (C) o che la assumono ma in bassi dosaggi.

Il totale dei token dell'intero corpus è di ca 59.000 tokens, precisamente 58.613, di cui 46.369 prodotti dai 4 pazienti: ossia il 79% del totale dei token corrisponde alla produzione verbale dei pazienti e solo il 20% ca agli interventi del terapeuta.

L'ampia variabilità nel numero dei turni, che contano le prese di parola di ciascun paziente e del terapeuta nel corso della seduta di psicoterapia, riflettono le diverse manifestazioni della patologia e l'estrema variabilità individuale della sindrome clinica.

2.1 Criteri di trascrizione ortografica

Le registrazioni acquisite sono state trascritte ortograficamente seguendo le specifiche del metodo CLIPS (*Corpora e Lessici dell'Italiano Scritto e Parlato*) relativamente alla modalità di turnazione, per la quale è stata ritenuta comunque fondamentale indicazione di unità la coerenza dal punto di vista semantico-pragmatico interna alla produzione di uno stesso locutore. In particolare, in analogia con CLIPS, è stato considerato *turno* ogni presa di parola da parte di uno dei due interlocutori, sia che interrompesse il turno dell'altro sia che si sovrapponesse a quest'ultimo (Savy 2007). Analogamente a CLIPS sono stati considerati e annotati gli *elementi linguistici lessicali* come le forme con aferesi o elisione, le parole con interruzione interna e le non parole da lapsus o errori e *gli elementi non lessicali* come le false partenze o le parole troncate; i *fenomeni verbali non lessicali* (come le esclamazioni e segnalazioni di assenso da parte del locutore, le pause piene con vocalizzazione o nasalizzazione) *e vocali non verbali* (come il colpo di tosse, lo sbadiglio, la risata ecc.), *i fenomeni non vocali* (come il "comunicativo generico" NOISE). Particolare attenzione è stata dedicata all'annotazione dei fenomeni di disfluenza.

I file di trascrizione hanno un formato ASCII (estensione txt), come per CLIPS, per favorirne la fruibilità. La conversione della codifica secondo le norme TEI (*Text Encoding Initiative*), per uniformarli alla più diffusa e riconosciuta forma di standardizzazione con linguaggio XML, non è stata ultimata.

In vista della ricodifica TEI, alcune etichette presenti nelle specifiche di trascrizione di CLIPS come "commenti del trascrittore" sono state modificate aprendo una parentesi graffa e segnalando l'inizio del fenomeno tra parentesi uncinate e segnalando poi la fine del fenomeno con parentesi uncinata, slash, ripetizione del fenomeno e chiusura di parentesi uncinata e di parentesi graffa. Diversamente da CLIPS anche le sovrapposizioni sono riportate in stringhe comprese tra parentesi graffe.

Il commento del trascrittore è riportato dopo il turno come <note> ... </note>. Un *beep* nel sonoro, e tre asterischi nella trascrizione, coprono i nomi propri per motivi di *privacy*.

Attualmente sono in corso l'annotazione pragmatica, un approfondimento dei fenomeni di disfluenza e di riparazione presenti nel corpus all'interno del concetto di 'catena' ("a cluster of several disfluent phenomena occurring in a sequence that constitutes a coherent unit within a turn" – Dovetto *et al.* in press) e una ricognizione sistematica delle modalità di occorrenza dei termini di natura cronologica e topografica che costituiscono una caratteristica significativa della *Lebenswelt* psicotica, caratterizzata da un rallentamento o arresto del tempo vissuto, evidente soprattutto nell'incidenza degli elementi deittici statico-spaziali.

È in corso uno studio sull'annotazione semiautomatica del corpus con l'obiettivo di realizzare uno strumento efficace per l'etichettatura multilivello di CIPPS ma adattabile anche ad altri tipi di corpora non-standard e in grado di automatizzare il più possibile le funzioni di annotazione (Dovetto, Panunzi, Gregori 2017; Panunzi *et al.* forthcoming).

3. *CIPP-ma*

Il secondo corpus è denominato CIPP-ma, *Corpus di Italiano Parlato Patologico (della) malattia di Alzheimer*. Attualmente raccoglie circa 5 ore di parlato registrato, di cui 2 h e 21min ca sono prodotte da

pazienti affetti da malattia di Alzheimer e 2 h e 40min ca da controlli; il totale dei pazienti sinora registrati è 20, 18 sono invece i controlli. La dimensione finale prevista, interrotta a causa dell'emergenza epidemiologica, era – ed è – di 40 pazienti e 40 controlli per un totale di ca 10 ore audio.

Il corpus è attualmente costituito dal parlato di 20 pazienti, di cui 11 donne e 9 uomini, di età media pari a 74:4 anni e media di anni di scolarizzazione pari a 8:3 anni, e 18 soggetti di controllo, bilanciati per età, scolarità, genere e status economico. L'età media dei soggetti di controllo è di 73 anni, mentre i loro anni di istruzione sono in media 8:1. Per due pazienti e due controlli si possiede anche una acquisizione longitudinale, ossia seconda acquisizione effettuata a circa 6 mesi di distanza dalla prima.

L'eloquio dei pazienti è stato registrato presso la Seconda Clinica di Neurologia dell'Università della Campania Luigi Vanvitelli; l'eloquio dei controlli, invece, è stato registrato presso le loro abitazioni. In entrambi i casi, i soggetti sono stati registrati previo consenso informato.

Sono state trascritte le registrazioni dell'intero gruppo dei pazienti, pari a 2 ore e 21 minuti di sonoro corrispondenti a circa 17mila tokens e 2.300 types e dell'intero gruppo di controllo (18 soggetti), pari a 2 ore e 40 minuti, corrispondenti a circa 25mila tokens e 3.300 types. Si prevede di implementare il corpus con ulteriori acquisizioni.

3.1 Task per l'acquisizione del parlato MA

I colloqui sono stati guidati da un'intervistatrice del gruppo di ricerca, a volte alla presenza del *caregiver* e di un'assistente che ha raccolto i dati sociolinguistici e anamnestici.

L'acquisizione del parlato, sia dei pazienti sia dei controlli, si basa su due diversi campioni di linguaggio. Il primo compito prevede la descrizione di una figura complessa (Capasso & Miceli 2001: 4) in cui sono raffigurati in primo piano i componenti di una famiglia dalle cui espressioni si percepisce una disposizione serena e felice; sullo sfondo della vignetta sono invece rappresentati due ladri nell'atto di rubare. Le espressioni e le azioni compiute dai personaggi della figura sono destinati anche a suscitare emozioni positive e negative in chi descrive. Il secondo compito prevede l'acquisizione di parlato (semi-)spontaneo, attraverso un'intervista semistrutturata: nella prima parte l'intervista ha come tema la televisione, oggetto presente nella vignetta e

che pertanto costituisce uno snodo di natura associativa tra un task e l'altro; nell'ultima parte dell'intervista il tema riguarda la famiglia e le abitudini giornaliere del soggetto.

4. *CIPP-mci*

Il terzo corpus, denominato CIPP-mci, *Corpus di Italiano Parlato Patologico (della condizione di rischio) Disturbo Cognitivo Lieve*, comprende attualmente circa 4 ore e 45min di parlato, diviso tra 6 soggetti a cui è stato diagnosticato il *Mild Cognitive Impairment*, cioè il Disturbo Cognitivo Lieve e 6 controlli. L'obiettivo del progetto è l'acquisizione del parlato di 20 soggetti mci e di 20 controlli.

Il corpus CIPP-mci è stato raccolto presso il Centro Regionale di Neurogenetica di Lamezia Terme nel 2019. Allo stato attuale comprende 6 pazienti (2 donne e 4 uomini) e 6 soggetti di controllo rappresentati dai rispettivi *caregivers* (moglie o marito del paziente; più raramente un figlio o altro familiare) acquisiti previo consenso informato. L'età media dei pazienti è di 70:6 anni, il loro livello di istruzione è di 12:17 anni in media e il punteggio medio ottenuto al MMSE è di 24,5/30.

I dati anagrafici dei controlli sono equiparabili a quelli dei pazienti per provenienza, età e scolarità. Il gruppo di controllo costituito dai *caregivers* dei pazienti permette un bilanciamento ottimale tra i due gruppi grazie alla condivisione dello stesso ambiente e alla similarità delle *enciclopedie* dei parlanti.

Le 12 registrazioni corrispondono a un totale di circa 4h e 45min e sono state interamente trascritte per un totale di 15.023 tokens e 2.169 types del gruppo dei pazienti e 16.610 tokens e 2.250 types del gruppo di controllo. È prevista l'acquisizione di altri soggetti per un totale di 20 pazienti e 20 controlli, corrispondenti previsionalmente a circa 16 ore di sonoro e circa 105.000 tokens.

L'acquisizione del parlato, sia dei pazienti sia dei controlli, si basa su campioni di linguaggio che prevedono, come in CLIPS e CIPPS, la compresenza di un Giver (*Instruction Giver*), ossia l'intervistatore che guida lo scambio conversazionale, e di un Follower (*Instruction Follower*), il soggetto paziente/controllo intervistato. Attraverso l'utilizzo del software Praat il parlato di G e F è segmentato in turni dialogici secondo la stessa definizione già seguita in CLIPS e in CIPPS.

Il turno rappresenta quindi la presa di parola da parte di uno dei due interlocutori: esso può sia interrompere il turno dell'altro locutore sia sovrapporsi a quest'ultimo senza costituire necessariamente interruzione. Analogamente sui *textgrid* sono annotati la fine e l'inizio del turno, segnalando anche la presenza di eventuali sovrapposizioni, e vengono segmentate anche le pause silenti, sia fra turni di locutori diversi sia interne al turno di uno stesso locutore, distinguendole in base alla lunghezza tra *sp*, *mp* e *lp*.

Per la lunghezza delle pause, in questo caso annotate con maggior dettaglio rispetto a CIPPS, abbiamo seguito Giannini (2008) considerando la pausa breve inferiore a 0,25s, la pausa media tra 0,25s e 1s e le pause lunghe superiori a 1s.

4.1 Task per l'acquisizione del parlato MCI

L'acquisizione del parlato si basa su due diversi task in cui i pazienti e i controlli sono intervistati dal personale medico. Il primo task consiste in un'intervista semi-strutturata coincidente con il colloquio neuropsicologico su temi a carattere quotidiano e familiare nonché sulla percezione delle condizioni fisiche e mentali proprie e dei propri familiari. Questa intervista produce un parlato dialogico semi-spontaneo.

Il secondo task è costituito dalla somministrazione della batteria SAND (*Screening for Aphasia NeuroDegeneration* – Catricalà *et al.* 2017) che consente l'accertamento di disturbi acquisiti del linguaggio causati da malattie neurodegenerative. La batteria si compone di 9 sub-test: 1. denominazione di figure, 2. comprensione di parole, 3. comprensione di frasi, 4. ripetizione di parole, 5. ripetizione di frasi, 6. lettura, 7. scrittura, 8. associazione semantica, 9. descrizione di figura.

5. *Prodotti della ricerca*

Di seguito si elencano i prodotti della ricerca del Gruppo di Lavoro, sottoarticolati con riferimento ai tre diversi *corpora*.

5.1 CIPPS

– Basile, Grazia & Dovetto, Francesca M. Forthcoming. *Metalinguistic ability, Pragmatic and Schizophrenia*, *Conference Pragmasophia 3, 13-15 July 2021*, Noto.

- Cresti, Emanuela & Moneglia, Massimo. 2017. Prosodic Monotony and Schizophrenia. In Dovetto, Francesca M. (a cura di), *Lingua e patologia. Le frontiere interdisciplinari del linguaggio*, Collana "Linguistica delle differenze" n. 2, Roma: Aracne, 147-197.
- Cresti, Emuanuela & Dovetto, Francesca M. & Rocha, Bruno. 2015. Schizophrenia and Prosody. First Investigations. In Manfredi, C. (ed.), *IX International Workshop Model and Analysis of Vocal Emissions for Biomedical Applications (MAVEBA), September 2-4 2015, Proceedings*, 139-142. Firenze: Firenze University Press.
- Dovetto, Francesca M. & Gemelli, Monica. 2009. Marcatori discorsivi nel parlato schizofrenico. In Gili Fivela, Barbara & Bazzanella, Carla (a cura di), *Fenomeni di intensità nell'italiano parlato*, 181-193. Firenze: Franco Cesati Editore.
- 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.] (Contributi di: Federico Albano Leoni; Federico Leoni; Carlo Pastore; Monica Gemelli; Francesca M. Dovetto; Isabella Chiari; Annamaria Cacchione; Cristina Bartolomeo, Elvira Improta & Manuela Senza Peluso).
- Dovetto, Francesca M. 2014. Schizofrenia e deissi. *Studi e Saggi Linguistici* LII, 101-132.
- Dovetto, Francesca M. 2015. Uso delle parole nella schizofrenia. In Mariottini, Laura (a cura di*), Identità e discorsi. Studi offerti a Franca Orletti*, 161-174. Roma: Roma TrE-Press.
- Dovetto, Francesca M. 2017. Usi della prima persona plurale nel testo schizofrenico. In Sorianello, Patrizia (a cura di), *Il parlato disturbato. Modelli, strumenti e dati empirici*, 49-66. Roma: Aracne.
- Dovetto, Francesca M. Forthcoming. Time and space in schizophrenia: the deixis of movement verbs. In Cardella, Valentina (a cura di), *Reti Saperi Linguaggi*, Special Issue *The challenge of mental disorder: psychopathology and the cognitive sciences*.
- Dovetto, Francesca M. In press. *Speech in Schizophrenia. A Corpus Analysis*, trad. di Laura Tagliaferro, Roma, tab edizioni, 2022.
- Dovetto, Francesca M. & Cresti, Emanuela & Rocha, Bruno. 2015. Schizofrenia tra prosodia e lessico. Prime analisi. In Orletti, Franca & Cardinaletti, Anna & Dovetto, Francesca M. (a cura di),

Studi Italiani di Linguistica Teorica e Applicata (SILTA), Numero tematico: Tra linguistica medica e linguistica clinica. Il ruolo del linguista XLIV/3, Nuova Serie. 486-507.
– Dovetto, Francesca M. & Panunzi, Alessandro & Gregori, Lorenzo. 2017. Sull'annotazione di un corpus orale mistilingue non standard (patologico schizofrenico). In De Meo, Anna & Dovetto, Francesca M. (a cura di), *La Comunicazione parlata / Spoken Communication*, *Napoli 2016, Collana "La comunicazione parlata"*, 345-361. Roma: Aracne.
– Panunzi, Alessandro & Gregori, Lorenzo & Dovetto, Francesca M. & Trillocco, Simona & Sorrentino, Assunta. Forthcoming. L'annotazione di corpora speciali. Ancora sull'annotazione e sul corpus CIPPS. In Dovetto, Francesca M. (a cura di), *Lingua e patologia. Parole dentro, parole fuori*, Roma: Aracne.

5.2 CIPP-ma

– Dovetto, Francesca M. & Guida, Alessia & Guarasci, Raffaele & Pagliaro, Anna Chiara. Forthcoming. Espressione e tematizzazione delle emozioni nelle malattie neurodegenerative: studio pilota su una paziente affetta da malattia di Alzheimer. In Castagneto, Marina & Ravetto, Maria (a cura di), *La comunicazione parlata / Spoken Communication*, Vercelli 5-7 Maggio 2021.
– Dovetto, Francesca M. & Guida, Alessia & Pagliaro, Anna Chiara & Guarasci, Raffaele. Forthcoming. Silences and disfluencies in a corpus of patients with Alzheimer's Disease (CIPP-ma). In Rose, Ralph & Eklund, Robert (eds), *Proceedings of DISS – Special Day on Disfluency in Speech and Language Disorders, Parigi 27 Agosto 2021.*
– Dovetto, Francesca M. & Schiattarella, Simona & Guida, Alessia & Coppola, Cinzia & Melone, Marina & Guarasci, Raffaele & Pagliaro, Anna Chiara & Raggio, Lucia & Sorrentino, Assunta & Trillocco, Simona. 2020. Relationships between cognition, emotion and language in Dementia Syndrome, from interdisciplinary to transdisciplinary research: A case study. *34th Virtual International Conference of Alzheimer's Disease International (Singapore, 10-12 dicembre 2020)* (poster).
– Melone, Marina & Dovetto, Francesca M. & Schiattarella, Simona & Guida, Alessia & Coppola, Cinzia. 2020. Parola, linguaggio ed emozioni nelle malattie neurodegenerative. Dalla fisiopatologia

agli studi clinici con uno studio pilota sulla tematizzazione delle emozioni. In Dovetto, Francesca M. (a cura di), *Lingua e patologia. I sistemi instabili, Collana "Linguistica delle differenze" n. 5*, 123-177. Roma: Aracne.

- Trotta, Daniela & Albanese, Teresa & Stingo, Michele & Guarasci, Raffaele & Elia, Annibale. 2018. Multi-Word Expressions in spoken language. *PoliSdict*, *Quarta Conferenza Italiana di Linguistica Computazionale – CLiC-it*.
- tesi di dottorato: Guida, Alessia, *Costruzione e annotazione di un corpus dell'età senile in pazienti affetti da malattia di Alzheimer*, Dottorato in Filologia moderna (35° ciclo), Università Federico II di Napoli (tutor Francesca M. Dovetto).
- post-doc: Pagliaro, Anna Chiara, *Linguistica dei corpora e corpora non-standard*, assegno di ricerca in Glottologia e Linguistica (L-LIN/01), a.a. 2020/2021, Dipartimento di Studi Umanistici, Università Federico II di Napoli (resp. scientifico Francesca M. Dovetto).
- tesi di laurea magistrale: Raggio, Lucia, *Analisi e trascrizioni su un corpus di pazienti affetti da malattia di Alzheimer*, Tesi di Laurea Magistrale in Glottologia e Linguistica (L-LIN/01), Dipartimento di Studi Umanistici, Università Federico II di Napoli (relatrice Francesca M. Dovetto).

5.3 CIPP-mci

- Bruni, Amalia C. & Dovetto, Francesca M. & Guida, Alessia & Laganà, Valentina & Pagliaro, Anna Chiara & Guarasci, Raffaele & Schiattarella, Simona & Sciutto, Paola. Forthcoming. Test neuropsicologici per l'analisi dei deficit linguistici. Spunti di riflessione per l'analisi linguistica su un caso di studio, in Dovetto, Francesca M. & Raso, Tommaso & Sorianello, Patrizia (a cura di), *Chimera. Romance Corpora and Linguistic Studies*, Num. monografico *Le patologie del linguaggio: studi e risorse tra crossdisciplinarità e interdisciplinarità*.
- Laganà, Valentina & Guida, Alessia & Pagliaro, Anna Chiara & Sciutto, Paola & Dovetto, Francesca M. & Puccio, Gianfranco & Bruni, Amalia. 2021. Integrating linguistic analysis with neuropsychological assessment: an opportunity to identify early signs of AD. *XVI Convegno Nazionale Sindem – SOCIETÀ ITALIANA*

DI NEUROLOGIA "Associazione Autonoma Aderente alla SIN per le Demenze", 25-27 novembre 2021, Firenze (poster).

Riferimenti bibliografici

Alzheimer's Association. 2016. Alzheimer's disease facts and figures. *Alzheimers Dement. 12(4),* 459-509.

Alzheimer's Disease International. 2021. *World Alzheimer's Report 2021.* https://www.alzint.org/u/World-Alzheimer-Report-2021.pdf

APA. American Pyschiatric Association. 2013. *Diagnostic and statistics Manual of mental disorders. Fifth edition*. CBS Publishers & Distributors.

Capasso, Rita & Miceli, Gabriele. 2001. *Esame Neuropsicologico per l'Afasia: ENPA* (Vol. 4). Berlino: Springer Science & Business Media.

Catricalà, Eleonora & Gobbi, Elena & Battista, Petronilla & Miozzo, Antonio & Polito, Cristina & Boschi, Veronica & Esposito, Valentina & Cuoco, Sofia & Barone, Paolo & Sorbi, Sandro & Cappa, Stefano F. & Garrard, Peter. 2017. SAND: a Screening for Aphasia in NeuroDegeneration. Development and normative data. *Neurological Sciences* 38. 1469-1483.

Dovetto, Francesca M. & Guida, Alessia & Pagliaro, Anna Chiara & Guarasci, Raffaele. Forthcoming. Silences and disfluencies in a corpus of patients with Alzheimer's Disease (CIPP-ma). *DISS – Special Day on Disfluency in Speech and Language Disorders, Parigi 27 Agosto 2021.* In Rose, Ralph & Eklund, Robert (eds), *Proceedings*.

Dovetto, Francesca M. & Panunzi, Alessandro & Gregori, Lorenzo. 2017. Sull'annotazione di un corpus orale mistilingue non standard (patologico schizofrenico). In De Meo, Anna & Dovetto, Francesca M. (a cura di), *La Comunicazione parlata / Spoken Communication, Napoli 2016*, Roma: Aracne, 345-361.

Giannini, Antonella. 2008. I silenzi del telegiornale. In Pettorino, Massimo & Giannini, Antonella & Vallone, Marianna & Savy, Renata (a cura di) *La comunicazione parlata, Atti del Congresso Internazionale, Napoli 23-25 febbraio 2006, Tomo I*, 97-108. Napoli: Liguori.

Marini, Andrea. 2018. *Manuale di Neurolinguistica. Fondamenti teorici, tecniche di indagine, applicazioni.* Nuova edizione. Roma: Carocci.

Panunzi, Alessandro & Gregori, Lorenzo & Dovetto, Francesca M., Trillocco, Simona & Sorrentino, Assunta. Forthcoming. L'annotazione di corpora speciali. Ancora sull'annotazione e sul corpus CIPPS. In Dovetto, Francesca M. (a cura di), *Lingua e patologia. Parole dentro, parole fuori*, Roma: Aracne.

Petersen, Ronald C.. 2004. Mild cognitive impairment as a diagnostic entity. *J Intern Med. 256(3)*, 183-194.

Petersen, Ronald C. & Caracciolo, Barbara & Brayne, Carol & Gauthier, Serge & Jelic, Vesna & Fratiglioni, Laura. 2014. Mild cognitive impairment: a concept in evolution. *J Intern Med. 275(3),* 214-228.

Sanford, Angela M.. 2017. Mild Cognitive Impairment. *Clinics in Geriatric Medicine 33(3),* 325–337.

Savy, Renata. 2007. Specifiche per la trascrizione ortografica annotata dei testi raccolti. *Progetto CLIPS*. http://www.clips.unina.it/it/

WHO. World Health Organization. 1993. *The ICD-10 classification of mental and behavioural disorders.* World Health Organization.

Sitografia

CLIPS. Corpora e Lessici di Italiano Parlato e Scritto.
http://www.clips.unina.it/it/

LISA. Lingua e Salute.
https://www.lupt.it/attivita/lisa.html

TEI. Text Encoding Initiative.
https://www.tei-c.org/release/doc/tei-p5-doc/it/html/TS.html

DOI: 10.17469/O2106SLI000011

HELIANA MELLO, TOMMASO RASO, MIGUEL OLIVEIRA, TONY BERBER SARDINHA, CLÁUDIA FREITAS, SANDRA MARIA ALUÍSIO, THIAGO PARDO, MAGALI DURAN, SIDNEY LEAL, MARK DAVIES, CHARLOTTE GALVES-CHAMBELLAND

Brazilian Portuguese: Spoken, Written and Diachronic Corpora

In this paper, a team of corpus linguists who work with Portuguese corpora have joined to report on the major corpora projects dealing with Brazilian Portuguese. All of the corpora are accessible digitally. The projects cover spoken (C-ORAL-BRASIL, Nurc Digital), written (Corpus Brasileiro, Linguateca corpora, NILC corpora) and diachronic corpora (Tycho Brahe). The corpora major characteristics, the projects sites as well as a list of publications pertaining to them are provided.

Keywords: Brazilian Portuguese, spoken corpora, written corpora, diachronic corpora.

1. *Introduction*

In this paper we present sequentially, in the same order that can be seen in the SLI Demo video (https://underline.io/lecture/33029-brazilian-portuguese-spoken,-written-and-diachronic-corpora), the Brazilian Portuguese corpora discussed by presenters during the 2021 SLI conference. The sequence will be as follows: C-ORAL-BRASIL (Mello & Raso), Nurc Digital (Oliveira Jr.), Corpus Brasileiro (Berber Sardinha), Linguateca Corpora (Freitas), NILC Corpora (Aluísio *et al*.), Corpus do Português (Davies) and Tycho Brahe (Galves-Chamberlland).

2. *C-ORAL-BRASIL*

The C-ORAL-BRASIL corpora have been compiled at the LEEL Lab, at the Federal University of Minas Gerais (UFMG), Brazil, by Tommaso Raso and Heliana Mello and their research team.

The C-ORAL-BRASIL corpora portray spontaneous Brazilian Portuguese speech in natural context, media and telephonic interactions. The corpora comprise sound, sound to text alignment, transcription, PoS and parsing files.

The corpora are useful for several types of studies focusing on spontaneous speech, however they were designed so as to be especially adaptable for informational structuring, prosodically based pragmatic studies.

For each corpus, a sample is manually informationally tagged, following the Language into Act Theory (Cresti 2000). The complete collection of corpora encompasses the following:

- C-ORAL-BRASIL I: informal spontaneous speech (Raso & Mello 2012);
- C-ORAL-BRASIL II: formal in natural context speech, media and telephone (forthcoming);
- C-ORAL-ESQ: schizophrenic patients and doctor interactions (forthcoming);
- C-ORAL-ANGOLA: Angolan Portuguese (forthcoming);
- COBAI: Brazilian learners of English speech (2012);
- COLPI: Indigenous Brazilian Portuguese (2016);
- Informationally tagged minicorpora of Brazilian Portuguese, Italian and American English.

The C-ORAL-BRASIL I and II portray an ample range of diaphasic variation, covering private, public, informal and formal monologues, dialogues and conversations, in addition to an assortment of radio and tv programs, besides telephonic conversations. Diastratic variation is well represented, while diatopic variation for the C-ORAL-BRASIL I and the formal and telephonic parts of C-ORAL-BRASIL II are representative of the Belo Horizonte metropolitan area, while the C-ORAL-BRASIL II media part represents the variation that is carried through Brazil's major TV and radio networks.

The C-ORAL-BRASIL materials and products are accessible through the following links:

- Download area and publications: www.c-oral-brasil.org;
- Corpus queries: www.c-oral-brasil.org/db-com;
- C-ORAL-BRASIL I book download: https://www.dropbox.com/sh/s2n9w30dycnkauc/AAB1W_nIQtSVFdyPYeaGS2Fqa?dl=0

3. *NURC Digital*

Project NURC was a major corpus compilation initiative in Brazil, which started in 1969 with the aim of documenting and studying the spoken linguistic variety of five Brazilian state capital cities: Recife, São Paulo, Rio de Janeiro, Porto Alegre and Salvador.

The materials collected by project NURC have been used in a large number of academic papers, theses, dissertations and works of great importance, such as The Spoken Portuguese Grammar.

Project NURC Digital, which focused on the NURC data collected in Recife, had the following as its major goals: to digitize the entire data collection of the NURC Recife Project; to catalog and store all metadata in digital format; to propose a multilevel annotation system for the NURC Project data; to digitize the transcription data referring to the shared corpus of the NURC Recife Project; to archive all digitized data in international language databases and to make all NURC Recife material available on a dedicated website.

In order to achieve its goals, the NURC Digital Project followed the recommendations proposed by the Technical Committee of the International Association of Sound and Audiovisual Archives (IASA) and by the Open Archival Information Systems (OAIS). These recommendations were observed in all phases of the Project's development.

All the treated material were archived locally at the Federal University of Alagoas (UFAL) and at The Language Archive (TLA), based at the Max Planck Institute.

The NURC Digital audio files were transcribed in PRAAT (Boersma & Weenink 2015), then imported into ELAN along with the transcriptions in order for the sound-transcript allignment to be performed. The transcripts were parsed using the PALAVRAS parser (Bick 2000).

All materials from the NURC Recife Project are available in high quality digital format, annotated and revised. They encompass 346 recordings reaching 300 hs, 417 speakers, divided among 208 files. All multilevel annotations in TextGrid and eaf format are available for download.

The NURC Digital materials can be accessed and downloaded from the NURC Digital Portal at https://fale.ufal.br/projeto/nurcdigital/ as well as from The Language Archive at https://hdl.handle.net/1839/4C5B6AAD-97D8-4C53-846F-AB39FAD85F55.

4. *Corpus Brasileiro*

Corpus Brasileiro (The Brazilian Corpus) is the first mega corpus of Brazilian Portuguese. It was developed between May 2008 and April 2010, by Tony Berber Sardinha along with José Lopes Moreira Filho and Elaine Albert.

The corpus has had three versions. The first one, lauched in 2020, was hosted on a local server at PUC-SP, was tagged with TreeTagger and had a relational database format. Version 2 came out in 2012, had the same design as the previous version and was hosted in Linguateca and SketchEngine. The third version was released in 2015, with the same design, but tagged with PALAVRAS. The third version is a 10.5 GB download, with tagged and untagged versions, through gzipped files.

The Corpus Brasileiro design by mode portrays the following characterization: 8% spoken (83,055,313 words) and 92% written (1,005,163,599 words). The design by domain shows the following distribution: 53% academic, 24% news, 8% education, 8% politics, 4% encyclopedia, 1% technical, 1% literary fiction, 1% legal texts.

The corpus can be downloaded for free after a license form is filled out, through the link: https://form.jotform.com/51214092562953

More information about Corpus Brasileiro can be obtained from Berber Sardinha & Ferreira (2014) or through contacting the corpus manager, Telma São Bento Ferreira at telma.ferreira@corpuslg.org

5. *Linguateca Corpora*

Linguateca is an infrastructure project for Portuguese corpora and computational resources that is over twenty years old. It portrays both Brazilian and European Portuguese corpus resources, along with other Portuguese varieties.

One of its subprojects, AC/DC – "access and availability of corpora"contains materials developed by the Linguateca team as well as other corpus linguistics groups. All corpora in AC/DC are syntactically annotated and may also contain semantic annotation.

Some examples of Brazilian Portuguese corpora in AC/DC are: ReLi (Freitas *et al*. 2014), OBras (Santos *et al*. 2018) and DHBB (Higuchi *et al*. 2019).

ReLi is a book review corpus, containing 1,600 reviews and 133,000 words. It is a user-generated content corpus, comprised of book reviews

posted on the internet. Its annotation portrays opinions about books and their polarities (positive or negative).

OBras is a Brazilian literature in the public domain corpus. It contains 272 texts, 43 authors and 7.2 million words. Obras was originally created to be the Brazilian counterpart of Vercial, a corpus of public domain literary works from Portugal. Its annotation encompasses places, people and human traits, literary genres (novel, short story, romance, etc.) and literary manifestation (realism, moderninsm, naturalism, etc.).

DHBB is comprised of Brazilian historical and biographical dictionaries and contains 7,685 entries and 9.8 million words. It belongs to the encyclopedic corpus genre and was designed to support Brazilian historical research and information extraction in the History domain. Its annotation portrays places, people, family relations, political parties and organizations.

The AC/DC corpora can be queried through the interface available at: https://www.linguateca.pt/acesso/info_acesso_English.php

6. *NILC Corpora*

The team at the Interinstitutional Center for Computational Linguistics (NILC) at USP-São Carlos have developed many Brazilian Portuguese corpora and computational resources. Three corpora will be presented below.

The first one is the PorSimplesSent corpus. This is a corpus comprised of aligned sentence pairs for the task of sentence readability assessment in Portuguese, as described in Leal *et al.* (2018). It is the first resource of this kind for the Portuguese language. The authors of the project made available four baselines for the corpus as well as an approach based on pairwise ranking to compare two versions of a sentence. Their model uses 17 lexical, syntactic and psycholinguistic features and identifies the readability level of sentence pairs with an accuracy of 74.2%. The corpus is publicly available at http://www.nilc.icmc.usp.br/nilc/index.php/tools-and-resources.

The second corpus is the PropBank.Br (Duran & Aluísio 2011). This corpus project aimed at adding a layer of semantic role labels (SRL) to a treebank of Brazilian Portuguese. The first phase of such annotation provided a training corpus that is currently being used to develop SRL classifiers. The SRL annotation was added to the syntactic trees gener-

ated by the parser PALAVRAS (Bick 2000) in the Brazilian portion of Bosque, a manually revised subcorpus of Floresta Sinta(c)tica (Afonso *et al.* 2002). PropBank.Br 1.0 and 2.0 are available for download at: http://143.107.183.175:21380/portlex/index.php/en/downloadsingl

The third corpus is the CSTNews corpus (Cardoso *et al.* 2011), which is a collection of texts annotated according the CST (Cross-document Structure Theory) model. The CSTNews corpus contains 50 Brazilian Portuguese text collections. Each collection has approximately 3 documents on the same subject but from different sources. This corpus portrays a complex multilevel annotation system that encompasses: PoS and syntactic automatic annotation, wh-question/aspect annotation, text-summary sentence alignment, subtopics, noun and verb Wordnet senses, multidocument discourse annotation, single document discourse automatic summaries, and single and multidocument manual summaries. The corpus is available at https://sites.icmc.usp.br/taspardo/sucinto/cstnews.html

7. *Corpus do Português*

The Corpus do Português (Portuguese Corpus) encompasses a collection of corpora covering the following: Genre/Historical (45 million words), Web/Dialects (1 billion words), NOW (1.1 billion words). WordsAndPhrase (top 40,000 words).

The Genre/Historical corpus contains 45 million words of data from the 1200s-1900s, and it can be used to look at the history of Portuguese. For the 1900s, it is equally divided between spoken, fiction, newspaper, and academic texts, which means that it can be used to compare genres of Portuguese.

The Web/Dialects corpus contains about one billion words of data in web pages from four different Portuguese-speaking countries (Brazil, Portugal, Angola, Mozambique). This corpus allows a look at very recent Portuguese (the texts were collected 2013-14), and comparison among the different dialects.

In 2022, many new features were added to this corpus: 1) browsing and searching the top 40,000 lemmas in the corpus 2) detailed "word pages" with information on each of these 40,000 words, including definitions, synonyms, links to images and videos, frequency information (by genre and country), collocates, related topics, and concordance lines), 3) the ability to input and analyze entire texts,

find keywords in these texts, and then see detailed information (cf. 2) for each word, as well as the ability to highlight phrases in a text and find related phrases in the corpus, and 4) extensive links to external resources in the frequency and conordance displays.

The NOW corpus is the newest addition to the Corpus do Português. It contains more than 1.1 billion words from four different Portuguese-speaking countries.

Finally, the WordandPhrase corpus allows searching and browsing through the top 40,000 words in Portuguese (based on frequency in the corpus). For each word, detailed information can be seen (all on one page) – definition, synonyms, frequency by genre, frequency by country, collocates (nearby words, which give great insight into meaning and usage), topics (co-occurring words on the same web pages), and 200 sample concordance lines (to see the patterns in which it occurs) – all with useful links from one word to another.

The Corpus do Português has a query interface that allows for several different types of searches. It is available at https://www.corpusdoportugues.org/x.asp

8. *Tycho Brahe*

The Tycho Brahe Parsed Corpus of Historical Portuguese (Galves 2019) is an electronic corpus of texts written in Portuguese by authors born between 1380 and 1978, encompassing both Portuguese and Brazilian authors.

At present, 88 texts (3,544,628 words) are available for research, with a linguistic annotation system in two stages: part-of-specch tagging (58 texts, a total of 2,280,819 words); and syntactic annotation (27 texts, a total of 1,234,323 words). The complete catalog of texts, along with author, title, number of words, as well as annotation status is available at http://www.tycho.iel.unicamp.br/~tycho/corpus/en/catalogo.html

The text preparation manual, along with the morphological and syntactic annotation manuals are available at http://www.tycho.iel.unicamp.br/~tycho/corpus/manual/index.html

The corpus page, through which all its specifications can be found, along with the download links is available at http://www.tycho.iel.unicamp.br/~tycho/corpus/en/index.html

Acknowledgements

Mello and Raso thank FAPEMIG for the research grants that have enabled the development of the C-ORAL-BRASIL project. Berber Sardinha thanks FAPESP for the research grant that has enabled the development of Corpus Brasileiro.

Bibliography

Afonso, Susana & Bick, Eckhard & Haber, Renato & Santos, Diana. 2002. Floresta sintá(c)tica: a treebank for Portuguese. In *Proceedings of LREC-2002*. Available at: https://www.linguateca.pt/documentos/AfonsoetalLREC2002.pdf

Bick, Eckhard. 2000. *The parsing system PALAVRAS: automatic grammatical analysis of Portuguese in a constraint grammar framework*. Aarhus: Aarhus University Press.

Boersma, P. & Weenink, D. 2015. Praat: doing phonetics by computer. Available at: http://www.praat.org

Cardoso, Paula C.F. & Maziero, Erick G. & Castro Jorge, Maria Lucía R. & Seno, Eloize M.R. & Di Felippo, Ariani & Rino, Lucia H.M. & Nunes, Maria das Graças V. & Pardo, Thiago A.S.. 2011. CSTNews – A Discourse-Annotated Corpus for Single and Multi-Document Summarization of News Texts in Brazilian Portuguese. In *Proceedings of the 3rd RST Brazilian Meeting*, 88-105. Available at: https://sites.icmc.usp.br/taspardo/RST2011-CardosoEtAl1.pdf

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

Duran, Magali Sanches & Aluísio, Sandra Maria. 2011. Propbank-Br: a Brazilian Portuguese corpus annotated with semantic role labels. In *Proceedings of the 8th Symposium in Information and Human Language Technology*, October 24-26, 2011, Cuiabá/MT, Brazil. Available at: https://aclanthology.org/W11-4519.pdf

Freitas, Cláudia & Motta, Eduardo & Milidiú, Ruy Luiz & César, Juliana. 2014. "Sparkling Vampire... lol! Annotating Opinions in a Book Review Corpus". In Aluísio, Sandra & Tagnin, Stella E.O. (eds.), *New Language Technologies and Linguistic Research: A Two-Way Road*, 128-146, Cambridge Scholars Publishing.

Galves, Charlotte. 2019. O corpus Tycho Brahe: um corpus sintaticamente anotado do português histórico. *Revista Binacional Brasil Argentina: Diálogo entre às Ciências*, 181-204.

Higuchi, Suemi & Santos, Diana & Freitas, Cláudia & Rademaker, Alexandre. 2019. Distant reading Brazilian history". In Navarreta, Costanza & Agirrezabal, Manex & Maegard, Bente (eds.), *Proceedings of the Digital Humanities in the Nordic Countries 4th Conference (Copenhagen, Denmark, March 5-8, 2019)*, 190-200.

Janssen, Maarten & Freitas, Tiago. 2010. Spock – a spoken corpus client. In Oliveira Miguel Jr. (ed.), *Estudos de Corpora: da Teoria à Prática*, 11-126. Lisboa: Edições Colibri.

Leal, Sidney Evaldo & Duran, Magali Sanches & Aluísio, Sandra Maria. 2018. A nontrivial sentence corpus for the task of sentence readability assessment in Portuguese. In *Proceedings of the 27th International Conference on Computational Linguistics*, 401–413. Available at: https://aclanthology.org/C18-1034/

Plichta, Bartek. 2002. Best practices in the acquisition, processing, and analysis of acoustic signals. *University of Pennsylvania Working Papers in Linguistics* 8(3), Article 16. Available at: https://repository.upenn.edu/pwpl/vol8/iss3/16

Raso, Tommaso & Mello, Heliana (eds.). 2012. *C-ORAL-BRASIL I. Corpus de referência do português brasileiro falado informal*. Belo Horizonte: Editora UFMG.

Santos, Diana & Freitas, Cláudia & Bick, Eckhard. 2018. OBras: a fully annotated and partially human-revised corpus of Brazilian literary works. In the public domain, *OpenCor*, Canela, RGS, Brasil, 24 de setembro de 2018.

Wittemburg, Peter & Brugman, Hennie & Russel, Albert & Klassmann, Alex & Sloetjes, Han. 2006. ELAN: a Professional Framework for Multimodality Research. In *Proceedings LREC 2006*. Available at: http://www.lrec-conf.org/proceedings/lrec2006/pdf/153_pdf.pdf

DOI: 10.17469/O2106SLI000012

FABIO TAMBURINI

I corpora del FICLIT, Università di Bologna: CORIS/CODIS, BoLC e DiaCORIS

Presenteremo i tre grandi corpora sviluppati dal nostro gruppo di ricerca e liberamente consultabili sul Web.
Il progetto più rilevante ha riguardato la creazione di CORIS, un corpus generale rappresentativo dell'italiano contemporaneo; è disponibile online dal 2001 e costantemente aggiornato, ogni tre anni, per cogliere le variazioni nel lessico e nei rapporti di frequenza. CODIS, la sua controparte dinamica, consente allo studioso di combinare liberamente insiemi predefiniti di testi per costruire il corpus ideale per studi specifici o per studi interlinguistici.
BoLC è un corpus bilingue (italiano/inglese) di linguaggio giuridico contenente una sezione parallela e una sezione comparabile. Consente lo studio e il confronto della terminologia adottata nei due sistemi giuridici ed è disponibile online dal 2011.
DiaCORIS, online dal 2006, riproduce la struttura in macro-varietà testuali di CORIS in una prospettiva diacronica. Contiene testi dall'Unità d'Italia agli anni 2000 adeguatamente annotati con i principali metadati.

Parole chiave: corpora, rappresentatività, struttura dinamica.

1. *Introduzione*

Questo breve contributo intende presentare i grandi corpora che sono stati sviluppati al Dipartimento di Filologia Classica e Italianistica (FICLIT) dell'Università di Bologna.

Negli ultimi 25 anni abbiamo sviluppato tre grandi corpora: il primo, e più importante progetto, riguarda CORIS/CODIS, il corpus di riferimento per l'italiano scritto contemporaneo; il secondo corpus che abbiamo sviluppato è un corpus terminologico, bilingue, contenente linguaggio giuridico e composto da una sezione in lingua inglese e una sezione in italiano; il terzo Corpus, DiaCORIS, è un corpus diacronico di riferimento per l'italiano dall'Unità d'Italia al 2001.

2. *CORIS/CODIS*

CORIS è il progetto decisamente più complesso che abbiamo affrontato (Rossini Favretti *et al*. 2002). È configurato come un COrpus di Riferimento per l'Italiano Scritto contemporaneo, è stato sviluppato a partire dalla fine degli anni '90 e attualmente, con l'ultimo aggiornamento inserito nell'estate 2021, contiene circa 165 milioni di parole, o meglio di *token*. CORIS viene aggiornato ogni tre anni mediante un monitor corpus in modo da mantenere il lessico e i rapporti di frequenza tra le varie aree semantiche aggiornati e di cogliere tutte le evoluzioni storiche avvenute in questi ultimi anni.

Come possiamo vedere nella Tabella 1, CORIS è costituito da una prima porzione di testi dagli anni '80 al 2000 contenente circa 100 milioni di parole ed è stato regolarmente aggiornato ogni tre anni inserendo un monitor corpus di 10 milioni di parole al fine di catturare i grandi eventi degli ultimi vent'anni.

Tabella 1 - *Struttura delle sezioni temporali di CORIS/CODIS e alcuni eventi storici rilevanti considerati nei vari monitor corpora*

Sezione temporale	*Dim.*	*Descrizione/Eventi*
CORIS 1980-2000:	100Mw	◃ Sezione iniziale di CORIS
Monitor 2001-04:	10Mw	◃ 11 settembre 2001, Euro, Guerre in Medio Oriente
Monitor 2005-07:	"	◃ Italia Camp. Mondo di Calcio
Monitor 2008-10:	"	◃ Crisi economica globale
Monitor 2011-13:	"	◃ Incidente di Fukushima
Monitor 2014-16:	"	◃ Fond. Islamico, Terrorismo
Monitor 2017-20:	14Mw	◃ Primo anno di pandemia

CORIS è disponibile online dal 2001 per tutti gli studiosi e gli studenti: si configura quindi come un corpus di riferimento generale sincronico, contiene unicamente lingua scritta, testi autentici e integrali ed è annotato automaticamente rispetto alle categorie grammaticali (PoS-tag) e ai lemmi.

Per definire la struttura gerarchica del corpus (Biber 1993), ci siamo attenuti rigorosamente a criteri esterni (Atkins *et al*. 1992); per il bilanciamento dei vari subcorpora abbiamo considerato in prima istanza parametri quantitativi relativi alla circolazione dei quotidiani e alla distribuzione dei volumi, modulati, in una seconda fase, da parametri qualitativi, per riproporzionare al meglio i rapporti tra diverse

tipologie testuali, come ad esempio il livello di attenzione cognitiva e il tipo di uso dei testi che compongono il corpus.

CORIS e tutti gli aggiornamenti successivi rispettano quindi le proporzioni indicate nella Tabella 2 tra le sei principali macro-varietà testuali considerate.

Tabella 2 - *Tipologia e proporzioni tra le principali macro-varietà testuali (subcorpora) di CORIS/CODIS*

Subcorpus	*Proporzione*
Stampa	38%
Narrativa	25%
Prosa Accademica	12%
Prosa Giuridico-Amministrativa	10%
Miscellanea	10%
Ephemera	5%

Accanto a CORIS, ovvero composto dagli stessi documenti ma strutturato in modo completamente diverso, abbiamo introdotto CODIS, il COrpus Dinamico dell'Italiano Scritto (Tamburini 2002). Due sono state le ragioni che ci hanno spinto a sviluppare una versione dinamica di CORIS: innanzitutto, il nostro studio di rappresentatività, per quanto molto accurato, potrebbe non essere condiviso da altri studiosi e, in seconda istanza, per consentire studi interlinguistici basati su corpora. Come si può osservare dalla Tabella 3 (un estratto da Tamburini 2002) la struttura dei vari corpora nel panorama internazionale mostra una rilevante variazione nelle tipologie testuali e nelle proporzioni.

Tabella 3 - *Struttura e bilanciamento di alcuni corpora nel panorama internazionale (valori riferiti al 2002)*

Corpus	*Composizione*	
BNC	Volumi	52.5 Mw – 58.6%
90Mw – Inglese	Stampa	27.8 Mw – 31%
(solo lingua scritta)	Miscellanea	7.4 Mw – 8.3%
LSWE	Fiction	5 Mw – 17.8%
28Mw – English	News	10.6 Mw – 37.7%
(solo lingua scritta)	Prosa Accademica	5.3 Mw – 19%
	Prosa generica	6.9 Mw – 24.6%

Corpus	*Composizione*	
The Oslo Corpus	Fiction	3.8 Mw – 17%
22.3 Mw – Norvegese	Quotidiani/Riviste	10.6 Mw – 47.5%
	Prosa	7.8 Mw – 35%
Corpus de Referência do	Quotidiani	55 Mw – 60.8%
Português Contemporâneo	Volumi	20.5 Mw – 22.6%
(CRPC) – 92 Mw	Periodici	7 Mw – 7.7%
Portoghese	Decisioni della	1.8 Mw – 2%
(solo lingua scritta)	Suprema Corte di Giustizia	
	Miscellanea	3.9 Mw – 4.3%
	Opuscoli	0.3 Mw – 0.3%
	Lettere	0.1 Mw – 0.1%

La struttura dinamica di CODIS consente allo studioso di selezionare quali porzioni dei materiali di CORIS considerare nelle ricerche. I testi sono esattamente gli stessi di CORIS e sono stati semplicemente raggruppati in "sezioni" di svariati milioni di parole (si veda la Tabella 4): ogni subcorpus è suddiviso in quattro sezioni di materiali testuali con proporzioni differenti e lo studioso/studente può comporre il corpus da utilizzare per le indagini rispetto alle proprie sensibilità o le proprie necessità. Combinando queste sezioni liberamente, CODIS consente studi interlinguistici basati su corpora o consente di lavorare con un corpus strutturato diversamente o pensato con differenti criteri di rappresentatività rispetto alla progettazione iniziale.

Tabella 4 - *Dimensione delle sezioni di materiali selezionabili in CODIS rispetto ai vari subcorpora*

Subcorpus	*Dimensioni (in %)*			
Stampa	20	10	5	3
Narrativa	13	7	3	2
Prosa Accademica	5	4	2	1
Prosa Giuridico-Amministrativa	4	3	2	1
Miscellanea	4	3	2	1
Ephemera	2	1	1	1

3. *BoLC*

Il secondo grande progetto riguarda il *Bononia Legal Corpus* (Rossini Favretti *et al*. 2007). BoLC è un corpus di linguaggio giuridico bilin-

gue, contiene testi in lingua inglese britannica e in lingua italiana ed è stato strutturato in due sezioni distinte: una prima sezione contiene documenti paralleli, cioè in rapporto di traduzione, ed è basata prevalentemente su documenti dell'Unione Europea (direttive e sentenze); la seconda sezione contiene un corpus comparabile contenente testi estratti dalla giurisprudenza e dalla legislazione dei due paesi. La Tabella 5 mostra le tipologie di testi introdotte nelle due sezioni.

In questa duplice prospettiva, si è prevista la possibilità di giungere a una comparazione di testi giuridici, le cui forme sono espressione, da un lato, di un ordinamento giuridico comune, quale quello comunitario, e, dall'altro, di ordinamenti e culture giuridiche diverse, quali quelle sviluppate nei singoli stati. Si è inteso, in tal modo, tenere conto sia la progressiva formazione di un diritto uniforme a livello europeo sia la pluralità di sistemi normativi nazionali nell'ambito dell'Unione Europea.

Tabella 5 - *La struttura del corpus BoLC*

	Sezione 1	*Sezione 2*
Inglese	Directives, Judgments	Acts of Parliament, Chancery Division, Court of Appeal, Family Division, House of Lords, Privy Council, Queen's Bench Division, Statutory Instruments
Italiano	Direttive, Sentenze	Costituzione, Codice Civile, Codice Penale, Codice di Procedura Civile, Codice di Procedura Penale, Decreti Legislativi, Leggi Costituzionali, Leggi Ordinarie, Sentenze Penali Corte di Cassazione, Sentenze Civili Corte di Cassazione, Sentenze e Ordinanze della Consulta

4. *DiaCORIS*

L'ultimo grande progetto che presenteremo riguarda DiaCORIS, il corpus diacronico di italiano scritto sviluppato in collaborazione con l'Accademia della Crusca e L'Università di Modena e Reggio Emilia (Onelli *et al*. 2006). L'obiettivo di DiaCORIS è quello di integrare CORIS in una prospettiva diacronica: contiene testi a partire dall'Unità d'Italia nel 1861 fino al 2001, esattamente in corrispondenza dell'inizio temporale dei monitor corpora di CORIS, costituendo quindi un continuum temporale che copre l'evoluzione storica della lingua italiana dall'Unità fino a oggi. L'intervallo temporale è stato suddiviso in cinque sezioni composte da 5 milioni di parole ciascuna

per un totale di 25 milioni di parole. Ogni sezione ricalca la struttura delle macro-varietà testuali di CORIS (eccetto l'Ephemera, non presente in diacronia); le proporzioni tra le varie sezioni cambiano nel tempo in funzione dell'importanza della macro-varietà testuale nel periodo storico considerato (si veda la Tabella 6).

Tutti i testi di DiaCORIS contengono i metadati fondamentali che li caratterizzano (si veda la Tabella 7) in modo da consentire studi storici accurati e completi.

Tabella 6 - *Suddivisione temporale e proporzione tra le varie sezioni del corpus DiaCORIS*

Sezione	*Subcorpus*	*Prop.*
1861-1900 Dopo l'unificazione	STAMPA	15%
	NARRATIVA	30%
	SAGGISTICA	30%
	PROSA GIUR.-AMM.	10%
	MISCELLANEA	15%
1901-1922 Il Periodo Liberale	STAMPA	25%
	NARRATIVA	25%
	SAGGISTICA	25%
	PROSA GIUR.-AMM.	10%
	MISCELLANEA	15%
1923-1945 Periodo Fascista	STAMPA	30%
	NARRATIVA	25%
	SAGGISTICA	25%
	PROSA GIUR.-AMM.	10%
	MISCELLANEA	10%
1946-1967 Dopo la 2a Guerra Mondiale	STAMPA	35%
	NARRATIVA	20%
	SAGGISTICA	25%
	PROSA GIUR.-AMM.	10%
	MISCELLANEA	10%
1968-2001 Dopo la Rivoluzione del '68	STAMPA	40%
	NARRATIVA	20%
	1968 SAGGISTICA	20%
	PROSA GIUR.-AMM.	10%
	MISCELLANEA	10%

Tabella 7 - *Struttura dei metadati inseriti nei testi che compongono DiaCORIS in formato XML*

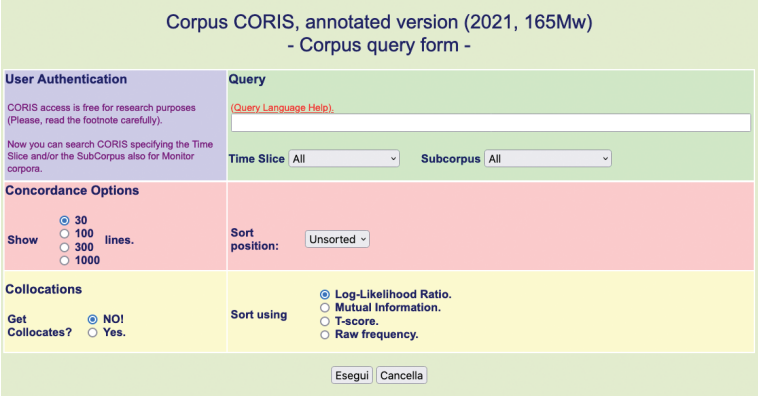
```
<text id="filename.xml" lingua="Italiano"
      titolo="titolo del testo"
      autore="nome autore" ed_test="editore/testata"
      anno="anno di pubblicazione"
      sez="sezione" subc="subcorpus">
word1
word2
...
</text>
```

5. *Interfaccia di consultazione dei corpora*

La Figura 1 mostra l'interfaccia di consultazione di tutti i corpora. Consente di effettuare tutte le più comuni operazioni per l'estrazione di informazioni da corpora elettronici (Sinclair 1991; 2004): l'estrazione delle concordanze di un termine, eventualmente incrociando le informazioni lessicali con le annotazioni, l'estrazione delle collocazioni di un nodo, avvalendosi di vari indici statistici di associazione, e l'estrazione di informazioni sulla frequenza dei termini.

L'accesso ai corpora è completamente libero all'indirizzo https://corpora.ficlit.unibo.it.

Figura 1 - *L'interfaccia di consultazione di CORIS, comune a tutti i corpora del FICLIT*



Riferimenti bibliografici

Atkins, Sue & Clear, Jeremy & Ostler, Nicholas. 1992. Corpus Design Criteria, *Literary and Linguistic Computing*, 7(1), 1-16.

Biber, Douglas. 1993. Representativeness in corpus design. *Journal of Literary and Linguistic Computing*, 8(4), 243-257.

Onelli, Corinna & Proietti, Domenico & Seidenari, Corrado & Tamburini, Fabio. 2006. The DiaCORIS project: a diachronic corpus of written Italian. In *Proc. 5th International Conference on Language Resources and Evaluation – LREC 2006*, 1212-1215. Genova.

Rossini Favretti, Rema & Tamburini, Fabio & De Santis, Cristiana. 2002. CORIS/CODIS: A corpus of written Italian based on a defined and a dynamic model. In Wilson, Andrew & Rayson, Paul & McEnery, Tony (a cura di), *A Rainbow of Corpora: Corpus Linguistics and the Languages of the World*, 27-38. Lincom-Europa, Munich.

Rossini Favretti Rema, Tamburini Fabio & Martelli Edoardo. 2007. Words from Bononia Legal Corpus. In W. Teubert (a cura di), *Text Corpora and Multilingual Lexicography*, John Benjamins Publishing Company, 11-30.

Sinclair, John. 1991. *Corpus Concordance Collocation*. Oxford: Oxford University Press.

Sinclair, John. 2004. Carter, Ronald (a cura di), *Trust the Text: Language, Corpus and Discourse*. Routledge.

Tamburini, Fabio. 2000. Annotazione grammaticale e lemmatizzazione di corpora in italiano. In Rossini Favretti, Rema (a cura di), *Linguistica e*

informatica: multimedialità, corpora e percorsi di apprendimento, 57-73. Bulzoni, Roma.

Tamburini, Fabio. 2002. A dynamic model for reference corpora structure definition. In *Proc. Third International Conference on Language Resources and Evaluation – LREC2002*, 1847-1850. Las Palmas, Canary Islands, Spain.

Tamburini, Fabio. 2007. CORISTagger: a high-performance PoS tagger for Italian. *Intelligenza Artificiale*, IV(2). 14-15.

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.

DOI: 10.17469/O2106SLI000014

PAOLO D'ACHILLE, CLAUDIO IACOBINI

Il corpus MIDIA: concezione, realizzazione, impieghi

MIDIA è un corpus diacronico bilanciato della lingua italiana liberamente consultabile in rete che comprende testi che vanno dall'inizio del XIII alla prima metà del XX secolo per un totale di circa otto milioni di occorrenze. Si caratterizza rispetto ad altri corpora per la scansione temporale, la tipologia di testi, la lemmatizzazione. In questo intervento descriviamo la concezione e la realizzazione di MIDIA, i suoi possibili sviluppi, e illustriamo due esempi di impiego di MIDIA per ricerche diacroniche di tipo morfologico e sintattico che considerano anche la prospettiva dei generi testuali.

Parole chiave: corpus, lingua italiana, diacronia, morfologia, generi testuali.

1. *Introduzione*

Il nostro contributo è diviso in due parti: nella prima parte (§§ 2-4) presentiamo il corpus MIDIA illustrandone brevemente la concezione, la struttura attuale e le prospettive di sviluppo; nella seconda parte (§ 5) sono proposti due esempi di utilizzo dei dati ricavabili da MIDIA tratti da nostre ricerche in corso.

2. *Caratteristiche di MIDIA*

MIDIA (acronimo di Morfologia dell'Italiano in DIAcronia) è un corpus liberamente consultabile in rete a partire dal settembre 2014 all'indirizzo http://www.corpusmidia.unito.it in quanto ospitato nei server dell'Università di Torino, dapprima in un server fisico, poi, nel corso del 2021, in un servizio cloud che ha richiesto alcuni aggiornamenti al software. Abbiamo approfittato di questo passaggio per apportare anche alcune piccole modifiche, che consistono nella correzione di alcuni errori materiali, nella integrazione al corpus di alcuni testi al fine di garantire un ancora migliore bilanciamento, e qualche

integrazione alla documentazione a corredo del sito insieme a una interfaccia in lingua inglese.

In estrema sintesi, MIDIA è un corpus diacronico bilanciato della lingua italiana che comprende testi che vanno dall'inizio del XIII alla prima metà del XX secolo, per un totale di quasi otto milioni di occorrenze. È stato realizzato grazie a un finanziamento di un progetto PRIN 2009, coordinato da Paolo D'Achille, che ha visto la proficua collaborazione tra linguisti generali e storici della lingua italiana. Il corpus MIDIA presenta due peculiarità: la prima è quella di non suddividere i testi raccolti nel corpus secondo una estrinseca suddivisione in secoli (modalità che è di gran lunga quella privilegiata negli studi di italianistica, e adottata anche nel corpus LIZ poi BIZ), ma di adottare una suddivisione temporale in cinque periodi cronologici significativi per la storia della lingua italiana; la seconda, quella di operare una distinzione in generi testuali, prendendo in esame anche scritture estranee alla sfera letteraria.

I cinque periodi cronologici (1. Dall'inizio del Duecento al 1375; 2. dal 1376 al 1532; 3. dal 1533 al 1691; 4. dal 1692 al 1840; 5. dal 1841 al 1947) sono scanditi da fatti di storia linguistica, letteraria e culturale che possono essere considerati come punti di svolta nella storia della lingua italiana.

Il primo periodo (dall'inizio del Duecento al 1375) parte dallo sviluppo della letteratura (e in genere della scrittura in volgare) in area toscana fino all'anno della morte di Boccaccio e dell'inizio dell'attività cancelleresca da parte di Coluccio Salutati; la data finale è la stessa che delimita il corpus testuale dell'OVI – TLIO.

Il secondo periodo (dal 1375 al 1532) abbraccia l'esperienza dell'Umanesimo e del Rinascimento, e in particolare, per quanto riguarda la lingua, accoglie testi che si collocano tra lo sviluppo del fiorentino "argenteo" e la scelta in direzione classicista del fiorentino "aureo" teorizzata nelle *Prose della volgar lingua* di Pietro Bembo (1525). La data finale coincide con quella della terza edizione dell'*Orlando Furioso*, attuazione in poesia delle teorie bembiane.

Il terzo periodo (dal 1533 al 1691) comprende il tardo Rinascimento, il Manierismo e il Barocco. La data di chiusura coincide con la terza edizione del *Vocabolario degli Accademici della Crusca* (1691), all'indomani della fondazione dell'Arcadia (1690).

Il quarto periodo (dal 1692 al 1840) coincide con l'età dell'Arcadia, dell'Illuminismo e del Romanticismo: è questa, sostanzialmente, l'epoca in cui alcuni studiosi (Durante 1981; Tesi 2005), poco propensi a riconoscere una continuità tra italiano antico e italiano moderno, hanno collocato la nascita dell'italiano moderno. Il periodo termina con l'edizione definitiva dei *Promessi sposi*, basata, come è noto, sul fiorentino dell'uso vivo, e per tanti aspetti modello linguistico dell'italiano postunitario.

Il quinto periodo (dal 1841 al 1947) è quello in cui a partire dal Risorgimento, passando per l'Italia unita e le due guerre mondiali, si arriva alla nascita della Repubblica e alla promulgazione della Costituzione. Sull'importanza linguistica dell'unificazione hanno giustamente insistito vari studiosi, primo fra tutti Tullio De Mauro (1963), che più di recente ha valorizzato anche, sul piano linguistico, gli anni della Repubblica (De Mauro 2014). L'approdo quasi alla metà del Novecento – prima, dunque, di quest'ultimo periodo – fa sì che i testi selezionati, pur se non troppo lontani del tempo, possano fornire alcuni elementi di differenziazione in diacronia rispetto all'italiano contemporaneo, oggetto degli studi più recenti sulla formazione delle parole.

La distinzione in generi testuali è un'altra peculiarità di MIDIA. I testi del corpus sono suddivisi in sette generi: i. testi espositivi; ii. testi giuridico-amministrativi; iii. testi personali; iv. poesia; v. prosa letteraria; vi. testi scientifici; vii. teatro, oratoria, mimesi dialogica. Rileviamo anzitutto che la distinzione all'interno dei testi letterari di due generi, poesia e prosa letteraria, riprende una distinzione tradizionale (percepita come tale già nel Cinquecento ed effettivamente molto importante nella storia della lingua italiana, in cui l'istituto della poesia – come ha mostrato Serianni 2009 – ha garantito la sopravvivenza, a tutti i livelli di analisi linguistica, di tratti usciti da tempo dall'uso), mentre la letteratura teatrale (su cui negli ultimi decenni si sono intensificati gli studi, aperti da un magistrale saggio di Nencioni 1976), è stata inserita nella sezione teatro, oratoria, mimesi dialogica, che accoglie testi scritti in vista di una fruizione orale o derivati da essa (tra cui prediche, discorsi, registrazioni di verbali di processi) e altre simulazioni di dialogo (quali i manuali di conversazione), al fine di cogliere, per quanto possibile, fenomeni rappresentativi della modalità parlata. Le altre quattro sezioni rappresentano le maggiori novità del

corpus. Una è costituita da testi personali (lettere, autobiografie, diari, memorie, libri di conti) in genere non destinati alla pubblicazione e che, dato il loro carattere privato, possono aprire finestre su aspetti della lingua d'uso specialmente in ambito familiare che molte persone colte utilizzavano nei secoli passati accanto al dialetto. La sezione dei testi espositivi comprende trattati, saggi, descrizioni, biografie e altre opere non rientranti nella categoria della prosa d'arte e disponibili ad accogliere tecnicismi e voci di matrice locale. Ancora maggiore, ovviamente, è la quantità di tecnicismi che si possono trovare nella sezione dei testi scientifici, che comprende soprattutto opere che hanno per oggetto le cosiddette scienze dure: la matematica, la fisica, la biologia, la chimica, la medicina. Vale la pena di precisare che in questo caso per i periodi più recenti abbiamo raccolto anche testi di discipline quali la statistica e la psicologia, mentre specialmente per i primi due periodi temporali il corpus accoglie opere di alchimia, bestiari, volgarizzamenti di trattati scientifici classici e altre opere di simile contenuto. Infine, la sezione dei testi giuridico-amministrativi raccoglie leggi, statuti, regolamenti, atti amministrativi, che rappresentano, accanto ai trattati scientifici, i testi definiti "molto vincolanti" nella tipologia testuale di Francesco Sabatini (1990).

3. *L'annotazione del corpus*

Un corpus così differenziato per generi testuali ed esteso nel tempo ha richiesto l'impiego di un programma di annotazione automatica che fosse adatto a trattare un corpus con tali caratteristiche. I criteri di annotazione e di lemmatizzazione, così come il programma di interrogazione e l'interfaccia web, sono stati concepiti e realizzati dall'unità di ricerca dell'Università di Salerno coordinata da Claudio Iacobini, che si è avvalsa del prezioso contributo di diversi collaboratori, tra i quali in particolare Giovanna Schirato per la parte linguistica e Aurelio De Rosa, attualmente Software Engineering Manager presso Facebook a Londra, per la parte informatica.

L'interfaccia di interrogazione è articolata e flessibile: permette infatti la combinazione di diversi criteri, tra cui la parte del discorso, la forma (o parti di essa), il lemma, il contesto, il periodo temporale, il genere testuale, l'autore del testo. I risultati, oltre che essere visualiz-

zati, possono essere scaricati in formato .csv; è possibile fare ricerche anche tramite espressioni regolari.

Il ricco polimorfismo e la stratificazione lessicale della lingua italiana nelle sue diverse fasi cronologiche hanno rappresentato uno dei principali problemi da affrontare al momento di scegliere quale programma di annotazione automatica utilizzare e come adattarlo al meglio alle nostre esigenze. Infatti, i programmi di annotazione automatica disponibili mirano nella quasi totalità all'analisi dell'attuale stato sincronico di lingua (o al più a un determinato stato sincronico), oppure a una determinata modalità (ad esempio lingua parlata, invece che scritta) oppure a uno specifico genere testuale. Una delle caratteristiche distintive di MIDIA, cioè quella di rendere disponibili e confrontabili in un corpus bilanciato testi di diverse epoche e generi, è stata dunque anche la causa di maggiore difficoltà nella etichettatura e lemmatizzazione del corpus.

Per poter fornire un'analisi automatica quanto più possibile adeguata alle variazioni dipendenti dai diversi generi linguistici e dall'estensione temporale dei testi presenti nel corpus, la soluzione più efficiente si è dimostrata essere l'utilizzo del programma TreeTagger combinato con un formario che modifica e integra quello realizzato da Marco Baroni (consultabile all'indirizzo https://docs.sslmit.unibo.it/doku.php?id=resources:morph-it) per la lingua italiana contemporanea[1]. Il nostro formario, che attualmente conta circa 550.000 forme associate ciascuna a un lemma o a una o più parti del discorso, è stato arricchito di forme tratte da testi appartenenti ai diversi generi compresi nel corpus in tutta la sua estensione temporale. L'arricchimento del formario si è rivelato una strategia particolarmente efficace per migliorare significativamente gli errori di assegnazione di parte del discorso o di lemmatizzazione inevitabili per un programma tarato su testi dell'italiano contemporaneo.

Il formario da noi raccolto si può considerare un prodotto secondario del lavoro di realizzazione di MIDIA e, oltre che indispensabile per l'analisi automatica del nostro corpus, può essere anche riutilizzabile per altri lavori di etichettatura di corpora che riguardino periodi dell'italiano precedenti a quello contemporaneo. Speriamo di riuscire

[1] Ringraziamo ancora una volta Marco Baroni per il prezioso e amichevole aiuto fornitoci nelle fasi iniziali del nostro lavoro.

ad accrescerlo ulteriormente e di raffinarlo grazie alla partecipazione a nuovi progetti ispirati o collegati a MIDIA.

L'impegno profuso nel miglioramento dello strumento di annotazione automatica e gli interventi semi-automatici successivi miranti a correggere errori sistematici risultati dal processo di annotazione automatica precedenti alla messa on-line di MIDIA non hanno potuto ovviamente impedire che il corpus oggi consultabile presenti errori di annotazione, alcuni dei quali sono stati del resto da noi stessi individuati nel corso delle nostre ricerche. Al momento non ci è stato possibile correggere i singoli errori di annotazione individuati, perché ciò comporterebbe onerosi interventi manuali che sono al di là delle nostre possibilità. Chi utilizza MIDIA deve dunque in ogni caso effettuare un controllo dei dati ottenuti.

4. *Finalità, limiti e possibili sviluppi*

Il fine per cui è stato realizzato MIDIA è quello di costituire la base per un progetto che speriamo prima o poi di portare a compimento: la pubblicazione di un testo di riferimento per lo studio della formazione delle parole dell'italiano in prospettiva diacronica che completi e si integri con quello curato da Grossmann & Rainer (2004) per la formazione delle parole dell'italiano in prospettiva sincronica.

Gli studi che finora hanno utilizzato come base di dati MIDIA, in effetti, hanno riguardato principalmente la morfologia derivazionale (tra i contributi al riguardo segnaliamo D'Achille & Grossmann 2016 su *-(t)ore* e *-trice*), ma diversi hanno affrontato anche altri ambiti, quali lo studio della morfologia flessiva, del lessico e delle costruzioni. Del resto, già in occasione del convegno di chiusura del PRIN (2014), come documentano gli atti editi a cura di D'Achille & Grossmann (2017), alcune delle diverse possibilità di uso del corpus erano state indicate; da allora, ovviamente, il quadro si è molto arricchito e non è irrilevante il numero dei lavori, apparsi sia in Italia sia soprattutto all'estero, di cui abbiamo avuto notizia, che a MIDIA hanno attinto.

Siamo consapevoli del limite quantitativo rappresentato dalle dimensioni del corpus MIDIA e proprio per questo, anche in vista dell'obiettivo iniziale di un ampio studio diacronico della formazione delle parole in italiano, avevamo presentato nel 2016 un nuovo progetto PRIN, che avrebbe consentito di allargare il corpus, cosa più

facile rispetto agli anni iniziali del progetto, dato che negli ultimi anni la disponibilità di testi in rete è diventata incomparabilmente maggiore di allora, anche per epoche e generi testuali che erano risultati allora abbastanza sguarniti e il cui riempimento ha costituito un notevole sforzo al fine della acquisizione in formato digitale di testi pubblicati solo a stampa. Il nuovo progetto prevedeva anche l'inserimento di testi dialettali, in modo da ampliare la riflessione sulla formazione delle parole all'intero dominio italo-romanzo, sia in sincronia sia in diacronia. Purtroppo, però, il nuovo progetto – pur essendo stato valutato positivamente – non riuscì a entrare nella rosa dei progetti finanziati.

Nonostante questa delusione, possiamo dichiararci piuttosto soddisfatti dell'accoglienza che l'impianto del progetto e i risultati ricavabili da MIDIA hanno avuto presso la comunità scientifica, in particolare all'estero. Possiamo anticipare che siamo stati coinvolti in due iniziative in corso che hanno MIDIA come riferimento importante, l'annotazione del corpus CODIT (*Corpus diacronico dell'italiano*) e il progetto CoRaLHis (*Comparison of Romance Languages through History*). Nella breve presentazione di questi due progetti sarà evidente la loro stretta correlazione con MIDIA e il ruolo che MIDIA ha svolto nella loro ideazione e realizzazione.

Il corpus CODIT[2], ideato e realizzato da Maria Silvia Micheli, è un corpus diacronico bilanciato di italiano scritto di circa 33 milioni di token (che evidentemente costituisce un significativo incremento quantitativo rispetto a MIDIA). La periodizzazione di CODIT riprende quella del corpus MIDIA e anche la tipologia dei testi è largamente coincidente. Il corpus CODIT è attualmente consultabile attraverso il portale dell'Istituto del corpus nazionale ceco http://www.korpus.cz collegato alla Università Carolina di Praga. Al momento si possono interrogare solo testi grezzi (non annotati) a partire dal seguente link: https://www.korpus.cz/kontext/query?corpname=codit. Il lavoro appena intrapreso di lemmatizzazione e annotazione in parti del discorso (promosso da Maria Silvia Micheli insieme a Jan Radimský) utilizza il formario raccolto per MIDIA. Si tratta evidentemente di una interazione fruttuosa, in primo luogo per l'etichettatura di CODIT, ma anche per l'ulteriore futuro arricchimento del formario MIDIA, che potrà incorporare le forme ricavate dal corpus CODIT.

[2] Su cui si possono avere maggiori informazioni dal sito https://wiki.korpus.cz/doku.php/en:cnk:codit.

CoRaLHis (acronimo di *Comparison of Romance Languages through History*) è un progetto già approvato e finanziato, ma non ancora partito, che ha come sede principale l'Università Sorbona di Parigi ed è coordinato da Anne Carlier e da Elisabeth Stark dell'Università di Zurigo. Il progetto ha lo scopo di realizzare una risorsa digitale multilingue ad accesso aperto che permetta ricerche empiriche, comparative e diacroniche basate su di un corpus di testi dal XIII al XVIII secolo comparabili per epoca e per generi testuali per le tre principali sotto-aree della Romania: area iberica, gallica e italica. Anche per questo progetto, le scelte adottate per MIDIA hanno fornito un riferimento utile, che va oltre lo studio della lingua italiana e l'ambito della formazione delle parole. L'integrazione del corpus MIDIA in CoRaLHis offrirà possibilità comparative tra le varietà romanze che arricchiranno notevolmente le possibilità di ricerca già offerte da MIDIA. Per rendere possibile la comparazione sarà necessaria una piccola integrazione ai metadati di MIDIA. Le indicazioni temporali dei testi dovranno infatti essere fornite non solo in riferimento alla periodizzazione basata su eventi rilevanti per la storia della lingua italiana, ma sarà necessario indicare il secolo o partizioni temporali più precise a cui ricondurre i testi. La modularità di MIDIA rende agevolmente possibile questa integrazione, che può essere facilmente ricavata dalle date di redazione o pubblicazione di ciascun testo che sono indicate nelle schede che lo accompagnano.

Crediamo quindi che MIDIA abbia dimostrato di poter contribuire con la propria architettura allo sviluppo di progetti importanti non solo per la storia della lingua italiana, ma anche per una visione storico-comparativa delle lingue romanze.

5. *Due esempi di ricerche basate su MIDIA*

Dopo aver delineato le caratteristiche essenziali di MIDIA, intendiamo qui di seguito contribuire a mettere in luce le possibilità offerte da MIDIA illustrando i risultati di due ricerche condotte specificamente per quest'occasione (oltre a quelle documentate nei contributi raccolti in D'Achille & Grossmann 2017 e in altri lavori posteriori): una di carattere propriamente morfologico e una di tipo sintattico, i cui dati quantitativi fanno riferimento alla versione di MIDIA non ancora aggiornata.

5.1 Il prefisso negativo *anti-*

Sia pur nei limiti della sua estensione quantitativa, il corpus MIDIA permette di ricavare utili indicazioni sulla formazione delle parole dell'italiano in prospettiva diacronica, quali ad esempio la diffusione nell'uso del prefisso negativo *anti-*. Da un impiego inizialmente ristretto a termini di ambito religioso nei soli testi di tipo espositivo (a parte una sporadica attestazione in un testo poetico), il prefisso ha avuto una lenta ma progressiva espansione nel numero di formazioni e ambiti semantici nei periodi successivi, fino ad arrivare nei testi dell'ultimo periodo a coprire tutti i generi testuali grazie a un numero di derivati significativamente maggiore rispetto a tutte le epoche precedenti.

Le parole derivate con *anti-* non sono presenti nei testi del primo periodo (dall'inizio del Duecento al 1375) con la sola eccezione di *anticristo*, parola di origine tardolatina, derivata a sua volta dal greco, attestata in un testo poetico, le *Rime* di Cecco Angiolieri. Nel secondo periodo (dal 1376 al 1532) le parole attestate sono ancora molto poche (*anticristiano*, *antipapa*) e solo in testi di tipo espositivo consistenti in cronache, storie e commenti di ambito religioso (la *Storia di fra' Michele Minorita*, le *Croniche* di Giovanni Sercambi, *Le Sposizioni di Vangeli* di Franco Sacchetti). Nel terzo periodo (dal 1533 al 1691), oltre a testi di tipo espositivo (si veda il termine di origine aristotelica *antiparistasi*), il prefisso è presente anche in parole usate in testi scientifici (*anticopernicano*). Nel quarto periodo (dal 1692 al 1840) i derivati con *anti-*, oltre che in testi scientifici ed espositivi, si trovano anche in testi giuridici (*antisociale*) e letterari (*anti-geometrico*). Solo nel periodo più recente (dal 1841 al 1947) si assiste a una decisa affermazione del prefisso, i cui derivati sono presenti in tutti i generi testuali del nostro corpus, e raggiungono un numero di formazioni decisamente più alto rispetto a quelle di tutti i secoli precedenti, oltre che una maggiore varietà di ambiti semantici (*anti-astensionista*, *anti-individualistico*, *anti-progressivo*, *antiborghese*, *anticonservatore*, *anticostituzionale*, *antiguerresco*, *antiintellettualismo*, *antimilitarista*, *antimodernismo*, *antipatriotta*, *antiré*, *antiscientifico*). Grazie ai dati ricavabili da MIDIA, viene confermato e documentato l'iter cronologico della diffusione dei prefissi nominali e aggettivali delineato in Iacobini (2019), ed è anche possibile fare un raffronto con i dati relativi all'impiego del prefisso *anti-* in francese e spagnolo pubblica-

ti da Martín García (1996), Fradin (1997), Montero Curiel (1998), Huertas Martínez (2015).

5.2 La perifrasi verbale *stare* + gerundio

Per quello che riguarda la sintassi, abbiamo preso invece in considerazione le occorrenze della perifrasi verbale *stare* + gerundio, che, come è noto, è diffusissima nell'italiano contemporaneo e che si è progressivamente estesa anche a verbi che già sul piano semantico indicano un'azione durativa. I dati di MIDIA, che, come si è detto, sono relativi a fasi precedenti agli sviluppi più recenti dell'italiano, sono riportati nella Tabella 1, che distribuisce le occorrenze in periodi (I-V) e generi testuali (così abbreviati: Po(esia), Pr(osa letteraria), T(eatro, ecc.), (Testi) E(spositivi), (Testi) S(cientifici), (Testi) G(iuridici), (Testi) Pe(rsonali):

Tabella 1 - *Le occorrenze di* stare + *gerundio in MIDIA*

	Po	*Pr*	*T*	*E*	*S*	*G*	*Pe*	TOT.
I	2	1	1	-	-	1	1	**5**
II	7	-	3	-	-	-	1	**11**
III	2	5	6	1	1	1	27	**43**
IV	11	12	5	2	2	-	29	**61**
V	5	18	31	9	7	1	17	**88**
TOT.	**27**	**36**	**46**	**12**	**10**	**3**	**74**	**208**

Dalla tabella risulta chiaramente come, sebbene la struttura sia documentata *ab antiquo*, le attestazioni nei primi periodi siano rare, specie in certi generi testuali, in cui mancano del tutto. La diffusione si ha a partire dal Periodo III, soprattutto nei testi personali, e poi più decisamente nel Periodo V, in cui la perifrasi è documentata in tutti i generi testuali, anche nella prosa letteraria e soprattutto nel teatro, più aperto all'oralità. Crediamo che i dati, pur nella forma bruta con cui sono stati presentati, possano apportare un contributo al tema, molto dibattuto soprattutto negli ultimi anni, della continuità tra italiano del passato e italiano contemporaneo: indicano infatti l'importanza della prospettiva di analisi quantitativa, che guarda alla frequenza di forme e strutture, accanto a quella qualitativa, finora privilegiata, che polarizza l'analisi in base al criterio della presenza/assenza.

A rafforzare l'ipotesi di un mutamento quantitativo avvenuto di recente in italiano per quanto riguarda la struttura in esame, aiuta

il confronto con altre perifrasi, anzitutto con *stare a* + infinito, che, come è noto, oggi o assume un carattere regionale, caratterizzando (con l'infinito apocopato) le varietà romana, laziale e abruzzese, oppure è standard, ma si usa solo in contesti in cui è impossibile il costrutto alternativo (in frase negativa, con verbi di percezione, ecc.; cfr. D'Achille & Giovanardi 1998, a cui si rimanda per una prima informazione bibliografica anche sulle altre perifrasi segnalate).

I dati forniti da MIDIA sono quelli riportati nella Tabella 2.

Tabella 2 - *Le occorrenze di* stare a + *infinito in MIDIA*

	Po	*Pr*	*T*	*E*	*S*	*G*	*Pe*	*TOT.*
I	7	3	-	-	5	8	-	**23**
II	8	9	7	11	2	6	2	**45**
III	4	10	11	-	3	2	10	**40**
IV	7	6	8	-	1	-	2	**24**
V	1	15	34	3	-	-	7	**60**
TOT.	**27**	**43**	**60**	**14**	**11**	**16**	**21**	**192**

Se il numero complessivo delle occorrenze è di poco inferiore a quello di *stare* + gerundio, la distribuzione sul piano cronologico e dei generi testuali è diversa: c'è un maggior numero di assenze, ma nel complesso si vede come la struttura, che prevaleva sulla concorrente nei primi due periodi, abbia ceduto il campo a partire dal Periodo III e più consistentemente nel Periodo IV e nel V, in cui tuttavia è ancora ben attestata, anche in questo caso soprattutto nella prosa letteraria e nel teatro, limitatamente ai contesti sintattici sopra richiamati.

Per completezza, a titolo di confronto, riportiamo nella Tabella 3 i dati relativi alla perifrasi "imminenziale" (che, normalmente, non è in concorrenza con le due precedenti) *stare per* + infinito.

Tabella 3 - *Le occorrenze di* stare per + *infinito in MIDIA*

	Po	*Pr*	*T*	*E*	*S*	*G*	*Pe*	*TOT.*
I	-	2	-	-	-	8	-	**10**
II	-	2	-	1	-	6	2	**11**
III	2	2	1	1	-	2	10	**18**
IV	2	7	-	2	-	-	2	**13**
V	1	15	8	5	4	-	7	**40**
TOT.	**5**	**28**	**9**	**9**	**4**	**16**	**21**	**92**

Anche questa perifrasi ha una continuità di presenze (se pure con un numero di assenze ancora maggiore, in particolare nei testi scientifici) fin dal Periodo I, ma il numero delle occorrenze risulta sempre inferiore a 10, tranne che nei testi personali del Periodo III e nella prosa letteraria del Periodo V, in cui dunque tutte e tre le perifrasi sono ben rappresentate, perché certamente funzionali al genere narrativo (largamente prevalente, in questo periodo, nei testi del corpus).

Tornando ai dati di *stare* + gerundio, è interessante anche il confronto con due perifrasi strutturalmente analoghe – anche se aspettualmente diverse (Squartini 1990; Brianti 1992; Dessì Schmid 2021) – che in un certo senso si possono considerare "alternative", *venire* + gerundio e *andare* + gerundio. Nel primo caso, in verità, come risulta dalla Tabella 4, i dati non risultano particolarmente significativi.

Tabella 4 - *Le occorrenze di* venire + *gerundio in MIDIA*

	Po	*Pr*	*T*	*E*	*S*	*G*	*Pe*	TOT.
I	10	8	-	3	6	1	-	**28**
II	10	9	5	1	4	-	2	**31**
III	7	2	6	2	7	1	4	**29**
IV	7	12	4	7	1	-	2	**33**
V	11	2	1	25	13	1	2	**55**
TOT.	**45**	**33**	**16**	**38**	**31**	**3**	**10**	**176**

Il numero delle occorrenze è più o meno equivalente a quello di *stare* + gerundio, ma risulta distribuito più equamente nei vari periodi e generi testuali, con scarse presenze tra i testi personali e con un incremento, nel Periodo V, più nei testi espositivi e scientifici che non nella prosa letteraria.

Quanto ad *andare* + gerundio, come risulta dalla Tabella 5, ha un numero di occorrenze significativamente più alto rispetto a tutte le altre perifrasi finora esaminate.

Tabella 5 - *Le occorrenze di* andare + *gerundio in MIDIA*

	Po	*Pr*	*T*	*E*	*S*	*G*	*Pe*	*TOT.*
I	58	33	6	22	5	15	-	**139**
II	78	40	24	57	23	8	23	**253**
III	40	58	61	60	152	19	81	**471**
IV	31	24	32	34	23	2	56	**202**
V	19	28	21	17	25	-	24	**134**
TOT.	**226**	**183**	**144**	**190**	**228**	**44**	**184**	**1199**

La struttura risulta ben diffusa anche nel Periodo I (dove manca solo nei testi personali) e ha una particolare espansione nel Periodo III, quello di costituzione della norma, con un numero esorbitante di presenze (che merita un esame specifico) nella prosa scientifica. Anche nel Periodo V il numero delle occorrenze, pur se un po' in calo, resta notevole e anzi, sia sul piano complessivo, sia nei singoli generi – a parte quelli giuridici, in cui, stranamente, manca, e il teatro, dove però le occorrenze non sono affatto trascurabili –, supera quello di *stare* + gerundio.

In questo caso, dunque, il "mutamento" quantitativo dell'italiano, che vede oggi *andare* + gerundio in declino probabilmente a causa dell'avanzata di *stare* + gerundio, è avvenuto oltre i limiti cronologici del nostro corpus, che tuttavia rivela già questa tendenza.

6. *Conclusioni*

In definitiva, pensiamo che MIDIA possa offrirsi come uno strumento utile, da usare soprattutto in combinazione con altri (corpora più ampi, che non offrono però le stesse possibilità di ricerche, e opere lessicografiche in rete), per indagini a diversi livelli di analisi linguistica, che possono fornire sia dati precisi su fatti specifici, sia motivi di riflessione su questioni di carattere più generale. Per questo abbiamo voluto che fosse presente in questo convegno dedicato ai corpora.

Riferimenti bibliografici

BIZ = *Biblioteca italiana Zanichelli*. Bologna: Zanichelli, 2010, DVD.

Brianti, Giovanna. 1992. *Périphrases aspectuelles de l'italien. Le cas de* andare*,* venire *et* stare + *gérondif*. Bern: Peter Lang.

D'Achille, Paolo & Giovanardi, Claudio. 1998. Conservazione e innovazione nella sintassi verbale dal romanesco del Belli al romanaccio contemporaneo. *Dal Belli ar Cipolla. Conservazione e innovazione nel romanesco contemporaneo*, 43–65. Roma: Carocci, 2001.

D'Achille, Paolo & Grossmann, Maria. 2016. I suffissati in *-(t)ore* e *-trice* nell'italiano del periodo 1841-1947. In Ruffino, Giovanni & Castiglione, Marina (a cura di), *La lingua variabile nei testi letterari, artistici e funzionali contemporanei. Analisi, interpretazione, traduzione. Atti del XIII Congresso della SILFI (Palermo, 22-24 settembre 2014)*, 787–805. Firenze: Franco Cesati.

D'Achille, Paolo & Grossmann, Maria (a cura di). 2017. *Per la storia della formazione delle parole in italiano: un nuovo corpus in rete (MIDIA) e nuove prospettive di studio*. Firenze: Franco Cesati.

De Mauro, Tullio. 1963. *Storia linguistica dell'Italia unita*. Bari: Laterza.

De Mauro, Tullio. 2014. *Storia linguistica dell'Italia repubblicana dal 1946 ai nostri giorni*. Roma-Bari: Laterza.

Dessì Schmid, Sarah. 2021. Zur Beziehung von progressiven Verbalperiphrasen und *states*. Ein erster Bericht aus Studien zu romanischen Sprachen. *Romanistisches Jahrbuch* 72(1). 31–62.

Durante, Marcello. 1981. *Dal latino all'italiano moderno. Saggio di storia linguistica e culturale*. Bologna: Zanichelli.

Fradin, Bernard. 1997. Une préfixation complexe: le cas de *anti-*. *Neuphilologische Mitteilungen* 98(4). 333–349.

Grossmann, Maria & Rainer, Franz (a cura di). 2004. *La formazione delle parole in italiano*. Tübingen: Niemeyer.

Huertas Martínez, Sheila. 2015. Aspectos de la formación de palabras en *anti-* en el español del siglo XIX. *Études Romanes de Brno* 36(1). 41–60.

Iacobini, Claudio. 2019. "Rapiéçages faits avec sa propre étoffe": Discontinuity and convergence in Romance prefixation. *Word Structure* 12(2). 176–207.

LIZ = LIZ 4.0. *Letteratura italiana Zanichelli*. CD-ROM dei testi della letteratura italiana. 4ª ed. per Windows. Bologna: Zanichelli, 2001.

Martín García, Josefa. 1996. Los valores semánticos y conceptuales de los prefijos *anti-* y *contra-* en español. *Cuadernos de Lingüística* 4. 133–150.

Montero Curiel, María Luisa. 1998. La evolución del prefijo *anti-*. In García Turza, Claudio & González Bachiller, Fabián & Mangado Martínez, José Javier (a cura di), *Actas del IV Congreso Internacional de Historia de la Lengua Española*, vol. 3, 321–328. La Rioja: Universidad de La Rioja.

Nencioni, Giovanni. 1976. Parlato-parlato, parlato-scritto, parlato-recitato. *Di scritto e di parlato. Discorsi linguistici*, 126–179. Bologna: Zanichelli, 1981.

OVI = Istituto Opera del Vocabolario Italiano, *Corpus OVI dell'Italiano antico*, http://gattoweb.ovi.cnr.it/.

Sabatini, Francesco. 1990. Analisi del linguaggio giuridico. Il testo normativo in una tipologia generale dei testi. *L'italiano nel mondo moderno. Saggi scelti dal 1968 al 2009*, vol. 2, 273–320. Napoli: Liguori, 2011.

Serianni, Luca. 2009. *La lingua poetica italiana. Grammatica e testi*. Roma: Carocci.

Squartini, Mario. 1990. Contributo per la caratterizzazione aspettuale delle perifrasi italiane *andare* + gerundio, *stare* + gerundio, *venire* + gerundio. *Studi e saggi linguistici* 30. 117–212.

Tesi, Riccardo. 2005. *Storia dell'italiano. La lingua moderna e contemporanea*. Bologna: Zanichelli.

TLIO = *Tesoro della Lingua Italiana delle Origini*, http://tlio.ovi.cnr.it/TLIO/.

DOI: 10.17469/O2106SLI000015

NAOMI NAGY, CHIARA CELATA

Un corpus per lo studio della variazione sociolinguistica dell'italiano in contesto migratorio

Presentiamo scopi, metodi e alcuni risultati rilevanti del progetto *Heritage Language Variation and Change in Toronto*, che dal 2009 raccoglie la produzione linguistica di 10 comunità alloglotte giunte nell'area in seguito a ondate migratorie di diversa storia e provenienza. Il progetto, ispirato alla metodologia variazionista nordamericana, ha portato alla costruzione di un corpus multilingue, stratificato socialmente e anagraficamente, con dati relativi alla lingua del patrimonio della prima generazione di immigrati così come di due generazioni di loro discendenti ed anche un campione di parlanti rimasti in ogni paese di origine. Vengono analizzate in particolare le caratteristiche del campione di italiano calabrese e vengono illustrati i principali risultati di due filoni di indagine, focalizzati rispettivamente su una variabile morfosintattica e una fonologica. Essi mettono in luce la complessità dei fattori linguistici ed extra-linguistici che influenzano il modo in cui le lingue del patrimonio vengono mantenute e trasmesse attraverso le generazioni.

Parole chiave: variazionismo, lingue del patrimonio, migrazione, socioindessicalità, italiano regionale calabrese, Canada.

1. *Introduzione*

Questo contributo illustra le caratteristiche del progetto *Heritage Language Variation and Change in Toronto* (Nagy 2011; 2015; 2018; http://ngn.artsci.utoronto.ca/HLVC/) e del corpus multilingue che ne scaturisce, con particolare riferimento al sotto-corpus di italiano come lingua del patrimonio della comunità di origine calabrese a Toronto, Canada. Oltre ad illustrare le caratteristiche generali del corpus, si ripercorreranno in sintesi alcuni dei principali risultati delle analisi linguistiche che sono state condotte negli anni recenti, allo scopo di mettere in luce le finalità, i metodi e il portato dell'analisi

variazionista applicata allo studio delle lingue del patrimonio e dei fenomeni di variazione (se non sempre mutamento) generazionale che le definiscono.

Il contributo è strutturato come segue: il §2 discute il concetto di lingua del patrimonio; il §3 presenta le caratteristiche complessive del progetto HLVC e del relativo corpus; il §4 ne illustra le potenzialità attraverso la rassegna di alcuni studi recenti di ambito fonologico e morfosintattico sul sotto-corpus di italiano; il §5 contiene, infine, alcune riflessioni conclusive.

2. *Le lingue del patrimonio*

L'espressione "lingue del patrimonio" è utilizzata in italiano come corrispettivo dell'inglese *heritage languages*. Nella maggior parte dei casi le lingue del patrimonio sono collegate a fenomeni migratori, individuali o collettivi.

È dunque necessario fare immediatamente una precisazione per quanto riguarda il concetto stesso di migrazione. Bisogna cioè distinguere tra che vive concretamente un'esperienza di migrazione, e chi discende da immigrati. Tradizionalmente si adotta una terminologia che qualifica i primi come immigrati "di prima generazione", e i secondi come immigrati "di seconda (terza, quarta etc.) generazione" (anche se, in realtà, questi ultimi non sono propriamente "immigrati"). Dal punto di vista linguistico, le lingue del patrimonio sono le lingue native degli immigrati, che poi vengono trasmesse, secondo modalità molto peculiari, alle generazioni successive. Ovviamente i livelli di competenza possono variare, e anche molto, tra le diverse generazioni. Inoltre, le lingue del patrimonio corrispondono solitamente alla lingua materna (L1) degli immigrati, ma tendenzialmente possono acquisire lo status di lingue non materne (L2) per le generazioni successive. La casistica, da questo punto di vista, è però molto varia: con lingue del patrimonio ci si riferisce a varietà che vengono parlate e acquisite come lingue ora "native", ora "non-native"; sempre, però, in contesto di lingua di minoranza.

Secondo alcune definizioni recenti (es. Chang 2021), i parlanti di lingue del patrimonio sono caratterizzati da una traiettoria di esposizione alla L1 che si è, per qualche motivo, interrotta; tale discontinuità si associa all'esposizione intensa ad una lingua non materna (L2).

Risulta evidente che in questa definizione possiamo far rientrare solo gli immigrati di prima generazione, e non anche i relativi discendenti. I motivi che hanno portato all'interruzione dell'esposizione alla L1 possono essere diversi (immigrazione volontaria, deportazione, adozione internazionale etc.).

Se, invece, volessimo descrivere la situazione delle seconde generazioni, potremo considerare come parlanti di lingue del patrimonio quei bilingui la cui L1 è stata appresa principalmente in casa come lingua di minoranza e la L2 principalmente fuori casa come lingua della maggioranza. Con questa descrizione si coglie, appunto, la condizione di quegli individui (tipicamente, i figli di immigrati, che nascono nel paese di accoglienza) la cui varietà linguistica di socializzazione primaria, almeno fino alla fine dell'età prescolare, è la lingua dei genitori ("patrimoniale", appunto), mentre la lingua maggioritaria nella comunità linguistica è appresa come una L2, nel momento dell'inserimento a scuola.

Neppure questa descrizione, però, esaurisce la casistica che interessa le lingue del patrimonio: in alcuni casi, infatti, la lingua del patrimonio è acquisita in contemporanea all'esposizione alla lingua della maggioranza, alla quale non è quindi possibile assegnare lo status di L2. Tendenzialmente, quello che si osserva è un progressivo cambiamento nello status reciproco delle diverse varietà, che può portare ad una vera e propria inversione delle condizioni iniziali, nel confronto inter-generazionale: se per gli immigrati di prima generazione la lingua del patrimonio è L1 e la lingua della maggioranza nel paese d'arrivo è acquisita in età adulta, per le terze o quarte generazioni la lingua della maggioranza è spesso L1 e la lingua del patrimonio è acquisita in modo frammentario e svincolato dall'uso quotidiano.

Ecco perché alcuni autori preferiscono allora caratterizzare le lingue del patrimonio non in base alla posizione che esse ricoprono nel repertorio linguistico dei parlanti o in base a rigidi schematismi che polarizzano le differenze in termini di "nativo" e "non-nativo", bensì in riferimento al valore socio-culturale che le lingue del patrimonio assumono per i parlanti che le usano. Esse, infatti, marcano un'appartenenza etnica, un'identità collettiva che si distingue da quella della maggioranza per la forte connessione ad un retroterra alloglotto, socialmente, storicamente e culturalmente diverso. Wiley (2001), ad esempio, si chiede se non sia più sensato, per definire chi siano i parlan-

ti di lingue del patrimonio, prendere in considerazione la loro "cultural connection to an ethnolinguistic group", rispetto alla loro "proficiency in the minority language". Similmente, e in rapporto specificamente al contesto canadese, Harrison (2000) propone una definizione di lingua del patrimonio come lingua parlata da "an individual with a cultural connection to a language other than English or French".

In certa narrazione pubblica, e a volte anche in quella scientifica, le lingue del patrimonio tendono ad essere descritte nei termini di un sistema linguistico lacunoso, deficitario, soggetto ad *attrition* (cioè ad "erosione"), in contrapposizione quindi all'idea di una competenza nativa pienamente sviluppata: i parlanti di queste varietà possiedono un vocabolario "limitato", una morfologia "incompleta", una sintassi "impoverita", una variazione di registro "non completamente sviluppata"; oppure vengono riportati consistenti fenomeni di "semplificazione" o "perdita". Una rappresentazione più utile è forse quella che sottolinea le specificità delle condizioni sociolinguistiche in cui si sviluppa il particolare tipo di bilinguismo che interessa i parlanti di lingue del patrimonio; come sottolinea ad esempio Montrul (2012), essi sono di fatto esposti ad una lingua minoritaria fin dall'infanzia e sono anche pienamente fluenti nella lingua della comunità linguistica più ampia, indipendentemente dal fatto di parlare una o entrambe le varietà con un "accento" non pienamente corrispondente alle attese del parlante "nativo". La competenza linguistica multipla dei parlanti di lingue del patrimonio tende a includere, infatti, elementi di reciproco influsso: della lingua della maggioranza sulla lingua del patrimonio e, viceversa, della lingua del patrimonio su quella della maggioranza (es. Mayr & Siddika 2016; de Leeuw & Celata 2019).

Le domande di ricerca che scaturiscono da questo complesso quadro interessano, dunque, uno spettro molto ampio di questioni, sia linguistiche che sociolinguistiche. In questo contributo ci concentriamo sulla variazione sociolinguistica e sul mutamento inter-generazionale come questioni che pongono sfide teoriche e metodologiche nello studio delle lingue del patrimonio. In prospettiva variazionista, non è tanto essenziale comprendere fino a che punto le lingue del patrimonio assomiglino o divergano dalla "norma" delle lingue native parlate nella terra d'origine degli immigrati di prima generazione, quanto piuttosto chiarire i meccanismi che stanno alla base della trasmissione delle varietà del patrimonio attraverso le generazioni, con particolare rife-

rimento a quelle che, nella varietà degli immigrati, funzionano come variabili sociolinguistiche. Ci si chiede quindi, in questa prospettiva, se il mutamento sia graduale attraverso le generazioni, oppure presenti degli elementi di discontinuità; e se la variazione sociolinguistica che caratterizza le varietà linguistiche parlate nei territori di emigrazione si mantenga, si modifichi o si perda completamente quando tali varietà diventano lingue di minoranza in un contesto alloglotto. Nei paragrafi che seguono si mostrerà come si possa tentare di rispondere a questi quesiti a partire da un corpus multilingue stratificato socialmente, costruito appositamente per rendere conto della variazione interna alle comunità minoritarie in uno specifico contesto nordamericano.

Ogni parlante di lingua del patrimonio nel corpus HLVC, descritto in §3, ha acquisito prima l'italiano ma abita in una città dove l'inglese è la (o una) lingua materna di un'ampia parte (47%) della popolazione.

3. *Il corpus HLVC*

Secondo un censimento del 2016, circa 3 milioni di persone residenti nella *Greater Toronto Area* hanno come lingua madre una lingua che non è né l'inglese né il francese (né alcuna delle lingue aborigene canadesi), cioè la metà della città (Statistics Canada 2017): la comunità più numerosa è quella dei parlanti di cinese (prevalentemente cantonese), seguono poi i parlanti di italiano (153,000 circa, perlopiù di provenienza calabrese o siciliana). Oltre a queste due lingue, ci sono parlanti di tagalog, portoghese, russo, polacco, coreano, ucraino e ungherese, e un piccolo gruppo di parlanti di francoprovenzale originari di Faeto e Celle San Vito, in Puglia; costituiscono il corpus HLVC, aggiungendo la dimensione del contrasto fra grandi e piccole comunità di immigrati. Le relative comunità si sono stabilite a Toronto in momenti storici diversi; le più anticamente attestate sono la comunità italiana (almeno dal 1906), quella russa e quella ucraina. Il progetto *Heritage Language Variation and Change in Toronto* (Nagy 2011; 2015; 2018; http://ngn.artsci.utoronto.ca/HLVC/) è incentrato su questo ricchissimo patrimonio linguistico e culturale.

Gli scopi del progetto sono sia descrittivi che teorici. In primo luogo vi è l'urgenza di documentare e descrivere le lingue del patrimonio parlate dagli immigrati stabilitisi nell'area di Toronto e da due generazioni di loro discendenti. Per far ciò è stato costruito un corpus

che permette la ricerca su una varietà di argomenti nelle 10 lingue del patrimonio collegate alle comunità citate sopra. Il progetto persegue poi lo scopo ambizioso si spingere la ricerca variazionista al di là dei limiti derivanti dalla sua tradizione, che è di fatto incentrata su comunità intese come monolingui, e su lingue non di minoranza. Inoltre vi sono anche scopi culturali che vanno al di là dell'ambito della ricerca scientifica, e che comprendono la promozione della vitalità delle lingue del patrimonio, inclusa la formazione e la mobilitazione di saperi trasversali che ruotano intorno a queste comunità.

Il corpus comprende la produzione linguistica di 3 generazioni di parlanti: gli immigrati di prima generazione (nati all'estero, arrivati a Toronto in età adulta almeno venti anni prima), i parlanti di seconda generazione (arrivati a Toronto prima del compimento del sesto anno di età oppure nati a Toronto, in entrambi i casi da genitori che si qualificano come di prima generazione) e quelli di terza generazione (nati a Toronto da genitori che si qualificano come di seconda generazione). L'assunzione di fondo è che l'insieme di questi parlanti mostri valori diversi di attaccamento alla lingua e cultura di origine, ma anche di vicinanza alla lingua e cultura canadese, in funzione della generazione di appartenenza; in particolare, si assume che l'attaccamento alla lingue e alla cultura d'origine e la competenza nella lingua del patrimonio siano maggiori nella prima generazione e diminuiscano progressivamente nelle generazioni successive, mentre l'attaccamento alla lingua inglese e alla cultura canadese e l'influsso dell'inglese sulla produzione in lingua del patrimonio siano maggiori nelle generazioni successive e minori nella prima generazione.

Non vi sono, però, solo le differenze generazionali a definire il quadro dei fenomeni che caratterizzano la trasmissione delle lingue del patrimonio; nel progetto HLVC si tiene conto anche di altre variabili sociali, tradizionalmente incluse nei protocolli di indagine variazionista. Essi sono il genere dei parlanti (il corpus è equamente diviso in maschi e femmine) e l'età, rispetto alla quale il corpus comprende quattro categorie di parlanti: minori di 21 anni, tra 22 e 39, tra 40 e 59, maggiori di 60. In totale, il corpus HLVC contiene registrazioni di 400 parlanti di una lingua del patrimonio.

Inoltre, il corpus comprende anche un nucleo, più ristretto ma in corso di ampliamento, di parlanti nativi non emigrati e residenti nelle regioni di emigrazione. Al momento sono stati raccolti i dati lingui-

stici di circa 120 individui. La loro produzione linguistica rappresenta una *baseline* rispetto a cui le varietà del patrimonio possono essere comparate, inclusa la variazione di età e di sesso.

Per ogni soggetto del corpus intervistato, i dati provengono da 3 diversi compiti linguistici, funzionali all'elicitazione di diversi stili di parlato e alla raccolta di informazioni extra-linguistiche. Il primo compito è l'intervista sociolinguistica (Labov 1984), che permette di elicitare un tipo di parlato conversazionale semi-guidato; gli intervistatori sono parlanti della stessa lingua del patrimonio degli intervistati. Gli argomenti tipicamente toccati durante il dialogo sono quelli che attengono al prima e al dopo rispetto all'esperienza migratoria, al confronto culturale tra la realtà di partenza e quella di arrivo, e alle vicende personali e familiari legate all'evento migratorio. Il secondo compito è un compito di descrizione di figura, in cui agli intervistati è richiesto di spiegare il contenuto di una serie di vignette in cui sono rappresentate scene di vita familiare. Infine, il terzo compito è un questionario orale volto ad accertare l'orientamento etnico dell'intervistato, rispetto ai due poli della cultura e lingua d'origine e della cultura canadese e lingua inglese. Il questionario comprende 37 domande, divise in diversi gruppi tematici: dall'autovalutazione etnica (es. "Ti consideri canadese, italiano, o italo-canadese?"), alle preferenze linguistiche (sia individuali che dei familiari) e culturali.

Il corpus HLVC è interamente trascritto ortograficamente in ELAN (la trascrizione ortografica è allineata temporalmente alla produzione orale). Alcune parti del corpus hanno poi anche ricevuto una codifica ulteriore, per aspetti linguistici specificamente indagati in studi differenti: ad esempio, il sotto-corpus italiano è stato annotato foneticamente per l'aspirazione delle occlusive sorde (cf. §4.2).

4. *L'italiano lingua del patrimonio a Toronto*

Lo studio dell'italiano regionale calabrese come lingua del patrimonio può contare, dentro al HLVC, su un campione di 40 parlanti residenti a Toronto, appartenenti alla prima, alla seconda e alla terza generazioni di immigrati e registrati tra il 2009 e il 2019, più un campione di 29 italiani tuttora residenti in Calabria, e registrati nel 2013.

Diverse variabili linguistiche sono state fatte oggetto d'analisi per l'italiano calabrese contenuto nel corpus HLVC: apocope, Baird *et*

al. 2021; (gl), Cervantes *et al.* in corso di stampa; (r), Cristiano 2022; DOM (marcatura differenziale dell'oggetto), Di Salvo & Nagy 2022; intonazione, Frascà 2015; pro-drop, Nagy 2015, 2017, 2018, Nagy *et al.* 2011; velocità d'eloquio, Nagy & Brook 2020; orientamento etnico, Nagy *et al*. 2014; covariazione tra variabili, Nagy & Gadanidis 2022*;* VOT, Nagy & Kochetov 2013; Nodari *et al.* 2019, Celata & Nagy 2022. La Tabella 1 presenta una visione d'insieme dei risultati ottenuti, con un'indicazione di massima rispetto alla presenza di mutamento intergenerazionale e alla differenza tra varietà del patrimonio e varietà parlata in patria. In questo paragrafo commenteremo in particolare i risultati ottenuti relativamente alla variazione intergenerazionale nel parametro del pro-drop (§4.1) e in quello dell'aspirazione delle occlusive sorde, che è stato analizzato separatamente nelle sillabe toniche e atone (§4.2).

Tabella 1 - *Sinossi dei risultati ottenuti nello studio dell'italiano calabrese come lingua del patrimonio nel corpus HLVC*

Variabile	*Stabile o in trasformazione*	*Differenze tra lingua del patrimonio e lingua dei residenti in patria*
Marcatura differenziale dell'oggetto	stabile	no
Pro-drop	stabile	no
(r) [ɾ/ɾ] ~ [ɹ]/[ɾ̥]	in trasformazione nei 2 luoghi	no
VOT in sillabe toniche	stabile	no
VOT in sillabe atone	in trasformazione nella lingua del patrimonio	sì
Apocope	in trasformazione	no
Velocità d'eloquio	in trasformazione nella lingua del patrimonio	sì
(gl) [ʎ] ~ [l] ~ [j]	in trasformazione nella lingua del patrimonio	sì

4.1 Pro-drop

L'oggetto pronominale nullo è uno degli aspetti di maggiore differenziazione tra l'inglese e l'italiano. L'analisi sul corpus HLVC ha riguardato 1147 token in frasi con verbo di modo finito. Oltre alle informazioni sulla generazione (prima, seconda o terza) e sul parlante, la codifica ha preso in considerazione 6 fattori linguistici indipen-

denti: la co-referenza del verbo con quello della frase precedente, il tipo di frase, persona e numero della forma verbale, tempo e aspetto della forma verbale, la presenza di negazione, e la presenza di un clitico preverbale con funzione di oggetto. L'ipotesi prevedeva che, nel confronto generazionale, il parametro del pro-drop mostrasse un avvicinamento progressivo ai valori dell'inglese (dunque, con l'obbligatorietà del soggetto pronominale preverbale espresso) e un allontanamento dall'ellissi pronominale dell'italiano.

I risultati (Nagy *et al.* 2011; Nagy 2015; 2018) hanno mostrato invece che il tasso di realizzazione e omissione dei pronomi soggetto non cambia significativamente nel confronto tra generazioni. Inoltre, i parlanti emigrati a Toronto mantengono lo stesso tasso di omissione del pronome che si rileva nei parlanti della varietà di origine. Un'analisi del fattore anagrafico ha anche mostrato che l'età dei parlanti non influisce sul modo di implementare il parametro del pro-drop. Similmente, non è stata trovata alcuna correlazione tra tasso di realizzazione del pronome soggetto e orientamento etnico dei parlanti, per come è stato rilevato dal questionario etnico-culturale (cf. §3). Dunque, non sembrano esserci fattori sociolinguistici rilevanti a modificare il comportamento dei soggetti rispetto a questa caratteristica della varietà del patrimonio. Infine, dei fattori linguistici presi in considerazione, i medesimi 3 (ossia, la continuità con il soggetto della frase precedente, il numero del pronome e il tempo del verbo) predicono nello stesso modo il pro-drop nel parlato di tutte e 3 le generazioni analizzate.

Questi risultati smentiscono l'opinione condivisa che le lingue del patrimonio siano uniformemente colpite da processi di "degenerazione" (*attrition*, o acquisizione incompleta) nel passaggio generazionale. Mostrano, al contrario, che vi sono aree della grammatica che rimangono pienamente funzionali e apparentemente non subiscono mutamento intergenerazionale, né in dipendenza di fattori sociali legati al parlante, né in rapporto a fattori linguistico-grammaticali.

4.2 VOT e aspirazione delle occlusive sorde

Il campione di dati che è stato analizzato in questo caso (Nagy & Kochetov 2013; Nodari *et al.* 2019; Celata & Nagy 2022) ammonta a 4973 parole contenenti [p], [t] o [k], prodotte da 23 parlanti di Toronto delle 3 diverse generazioni (circa 215 parole per ogni parlante).

Ogni occlusiva è stata codificata come target potenziale di una delle seguenti condizioni: (i) aspirazione di tipo calabrese, quindi in sillaba atona e con l'occlusiva preceduta da liquida o nasale (es. [ˈstaŋkʰo] *stanco*) oppure geminata (es. [ˈstak:ʰo] *stacco*); (ii) aspirazione di tipo inglese, quindi in sillaba tonica e con occlusiva scempia ad inizio di parola, preceduta da pausa o da parola terminante con vocale (es. [kʰaro] *caro*); (iii) nessuna aspirazione, in sillaba atona e con l'occlusiva a inizio di parola, preceduta da pausa o da parola terminante con vocale (es. [kʰaˈtena] *catena*). Assumendo il modello inizialmente illustrato (cf. §3), si ipotizza che l'aspirazione di tipo (i) sia più frequente nel parlato della prima generazione, come conseguenza di una maggiore competenza nella varietà di italiano calabrese, e che, per converso, l'aspirazione di tipo (ii) sia più frequente nel parlato delle generazioni successive alla prima; la condizione (iii) rappresenta il termine di controllo (non ci si aspetta, cioè, che il tasso di aspirazione si modifichi a livello intergenerazionale nei contesti di quel tipo, poiché essi non sono bersaglio di aspirazione né in italiano calabrese, né in inglese).

I contesti di tipo (i) e (ii) si differenziano anche e soprattutto per i tratti di socio-indessicalità che i parlanti eventualmente attribuiscono all'aspirazione. In Calabria, infatti, l'aspirazione delle occlusive sorde si correla in modo significativo con variabili sociali quali il sesso dei parlanti, il livello di scolarizzazione e l'orientamento verso la cultura locale (Nodari 2017; Nodari 2022; Falcone 1976). Lo stesso non può dirsi per l'aspirazione delle occlusive sorde nella varietà canadese di inglese, dove i fattori che influenzano la presenza dell'aspirazione sono tutti di tipo esclusivamente linguistico (come il luogo di articolazione dell'occlusiva o l'altezza della vocale che segue; fattori, peraltro, che svolgono un ruolo anche in italiano calabrese, Nodari *et al.* 2019). Pertanto, analizzando l'aspirazione nell'italiano del patrimonio degli immigrati calabresi a Toronto in relazione alle sillabe atone (contesti di tipo (i)) e toniche (contesti di tipo (ii)), è possibile definire le dinamiche della trasmissione intergenerazionale di un unico tratto fonetico in due condizioni sociolinguistiche diverse: una in cui la variazione è sociolinguisticamente rilevante nella lingua del patrimonio, l'altra in cui la variazione è sociolinguisticamente irrilevante nella lingua del patrimonio ed è, piuttosto, indotta dal contatto con la lingua della maggioranza.

Le analisi hanno mostrato che, in effetti, un diverso status sociolinguistico può portare a differenze nel modo in cui una variabi-

le viene trasmessa nelle generazioni. Nodari *et al.* (2019) mostrano che, nel caso dell'aspirazione di tipo calabrese, vi è un significativo decremento dei valori di VOT nel confronto generazionale, e soprattutto nel parlato della terza generazione; nel caso dell'aspirazione di tipo inglese, invece, i valori non si modificano in relazione al fattore generazionale. Inoltre, nell'aspirazione di tipo calabrese si riscontra anche un'interazione tra il fattore generazionale e quello del luogo di articolazione delle consonanti occlusive: mentre i parlanti di prima generazione presentano un pattern di opposizione binaria tra consonanti non posteriori (/p/ e /t/), con VOT più breve, e la consonante posteriore /k/, con VOT più lungo, i parlanti di seconda generazione presentano invece il tipico schema tripartito /p/ < /t/ < /k/, in cui le differenze di VOT sono interamente spiegabili con le differenze articolatorie delle tre consonanti (e che si ritrova uniformemente nelle tre generazioni in corrispondenza dell'aspirazione di tipo inglese). Cosa interessante, i parlanti di terza generazione riproducono lo schema bipartito dei nonni.

Un'altra importante differenza tra i due contesti di aspirazione, che si riscontra nei dati di Nodari *et al.* (2019), è quella per cui solo l'aspirazione uditivamente percepita (un altra misura presa in considerazione in quel lavoro) in sillaba tonica (dunque, nei contesti che abbiamo definito di tipo inglese) si correla in modo statisticamente significativo con due aspetti del'orientamento etnico dei parlanti (l'uso dell'italiano e le scelte culturali), per come risulta dal questionario etnico di cui si è parlato in §3: in questo caso, più i parlanti riportano una preferenza per la lingua e la cultura d'origine, meno presente è l'aspirazione delle occlusive sorde nei contesti che risentono dell'influsso dell'inglese. Nessun fattore etnolinguistico, però, ha impatto sulla probabilità di produrre aspirazione nei contesti tipicamente calabresi.

In conclusione, le analisi condotte sull'aspirazione delle occlusive sorde hanno messo in evidenza che, rispetto a questo tratto fonologico, e diversamente da quanto riscontrato per il *pro-drop*, l'italiano del patrimonio può cambiare di generazione in generazione. Il modo in cui questo cambiamento avviene, però, è determinato da un complesso rapporto tra fattori linguistici interni e fattori extra-linguistici: lo status sociolinguisticamente rilevante di una caratteristica del parlato può modificare il modo in cui essa viene mantenuta e trasmessa, rispetto a quando lo stesso tratto fonetico non è veicolo di informazioni socio-in-

dessicalmente rilevanti. Anzi, per quanto riguarda specificamente l'aspirazione percepita, i contesti di sillaba tonica non mostrano nessun cambiamento intergenerazionale, a indicare che, da questo punto di vista, la pronuncia della lingua del patrimonio non subisce necessariamente delle modifiche nel passaggio da una generazione all'altra; la connessione culturale con l'ambiente d'origine agisce apparentemente da blocco rispetto all'introduzione di tratti di pronuncia esogeni. I dati mostrano anche che gli schemi di trasmissione intergenerazionale possono essere non lineari, con la terza generazione che riproduce un *pattern* presente nel parlato dei nonni, ma non in quello dei genitori – un risultato sicuramente da approfondire con ulteriori, mirati studi.

5. *Conclusioni*

Questo contributo ha inteso mostrare l'importanza di metodologie sociolinguistiche basate su *corpora* per lo studio della variazione e del mutamento nei contesti migratori. Il paradigma di ricerca alla base del progetto HLVC permette di elicitare parlato semi-spontaneo, ed è funzionale alla costruzione di un campione stratificato socialmente e in senso generazionale. La compresenza di dati metalinguistici provenienti dal questionario etnolinguistico rappresenta una ricchezza che permette di valutare il ruolo della percezione di sé, e del proprio rapporto con la dimensione interculturale e multilingue, nel determinare i comportamenti linguistici. La raccolta di dati relativi alle varietà parlate nella terra d'origine da individui non emigrati rappresenta in prospettiva un ulteriore arricchimento funzionale soprattutto alla comprensione delle dinamiche linguistiche legate alla prima generazione di immigrati, secondo un approccio variazionista comparativo (es. Poplack & Tagliamonte 2001) che permette di confrontare due modelli (varietà pre- e post-migrazione) per verificare se i fattori che spiegano la variazione sono i medesimi oppure diversi. La trascrizione e l'annotazione del corpus in ELAN permette di analizzare eventi linguistici e paralinguistici rilevanti. Il progetto HLVC ha anche una dimensione multilingue, che permette di utilizzare metodi coerenti per lingue del patrimonio e variabili linguistiche diverse.

Ringraziamenti

Le autrici ringraziano i parlanti per la loro generosità e conoscenza, gli studenti e assistenti di ricerca che hanno reclutato e intervistato i parlanti e hanno trascritto e aiutato ad analizzare e presentare i dati. Sono elencati a https://ngn.artsci.utoronto.ca/HLVC/3_2_active_ra.php e https://ngn.artsci.utoronto.ca/HLVC/3_3_former_ra.php.

Ringraziano anche il Social Science and Humanities Research Council of Canada per il sostegno finanziario in forma di borsa 410-2009-2330 e 435-2016-1430.

Riferimenti bibliografici

Baird, Anissa & Angela Cristiano & Naomi Nagy. 2021. Apocope in Heritage Italian. *Languages* 6(3): 120.

Celata, Chiara & Nagy, Naomi. 2022. Sociophonetic variation and change in heritage languages: Lexical effects in Heritage Italian aspiration of voiceless stops. *Language and Speech*. https://doi.org/10.1177/00238309221126483

Cervantes, Eloisa & Hoffman, Michol F. & Nagy, Naomi & Walker, James A. in corso di stampa. Italians in Toronto. In Goglia, Francesco & Hajek, John (Eds.), *Italians(s) abroad: Italian language and migration in cities of the world*. De Gruyter.

Chang, Charles Bond. 2021. Phonetics and phonology of Heritage Languages. In Montrul, Silvina & Polinsky, Maria (Eds.), *The Cambridge Handbook of Heritage Languages and Linguistics,* 581-612. Cambridge, UK: Cambridge University Press.

Cristiano, Angela. 2022. (r) in Heritage Calabrese Italian: Cross-generational nativeness. (Tesi di laurea in Fonetica e Fonologia, Università di Bologna).

de Leeuw, Esther & Celata, Chiara. 2019. Plasticity of native phonetic and phonological domains in the context of bilingualism. *Journal of Phonetics* 75. 88-93.

Di Salvo, Margherita & Nagy, Naomi. 2022. Differential object marking in Heritage and Homeland Italian. In Bayley, Robert & Preston, Dennis R. & Li, Xiaoshi (eds.), *Variation in Second and Heritage Languages*. 311-336. Philadelphia: John Benjamins. DOI: 10.1075/silv.28.12dis.

Falcone, Giuseppe. 1976. *Calabria*. Pisa: Pacini.

Frascà, Paolo. 2015. Lingua cum dialectis: analisi diagenerazionale del dialetto calabrese nella conversazione. Indagine introduttiva preliminare. *Cultura e Comunicazione* VI(6): 15-20.

Harrison, Brian. 2000. Passing on the language: Heritage language diversity in Canada. *Canadian Social Trends* 58. https://www150.statcan.gc.ca/n1/en/pub/11-008-x/2000002/article/5165-eng.pdf?st=JqjkY5jr

Labov, William. 1984. Field methods of the project on linguistic change and variation. In Baugh, John & Scherzer, Joel (Eds.), *Language in use: Readings in sociolinguistics*, 28-66. Englewood Cliffs: Prentice Hall.

Mayr, Robert & Siddika, Aysha. 2016. Inter-generational transmission in a minority language setting: Stop consonant production by Bangladeshi heritage children and adults. *International Journal of Bilingualism* 22(3). 255–284.

Montrul, Silvina. 2012. Bilingualism and the Heritage Language Speaker. In Bhatia, Tej K. & Ritchie, William C. (Eds.), *The Handbook of bilingualism and multilingualism,* 168–189. Oxford: Blackwell.

Nagy, Naomi. 2011 A multilingual corpus to explore geographic variation. *Rassegna Italiana di Linguistica Applicata* 2. 65–84.

Nagy, Naomi. 2015. A sociolinguistic view of null subjects and VOT in Toronto heritage languages. *Lingua* 164. 309–327.

Nagy, Naomi. 2017. Documenting variation in (endangered) heritage languages: how and why?. *Language Documentation and Conservation* SP13.

Nagy, Naomi. 2018. Linguistic attitudes and contact effects in Toronto's heritage languages: A variationist sociolinguistic investigation. *International Journal of Bilingualism* 22(4). 429–446.

Nagy, Naomi & Aghdasi, Nina & Denis, Derek & Motut, Alexandra. 2011. Null Subjects in Heritage Languages: Contact effects in a cross-linguistic context. *University of Pennsylvania Working Papers in Linguistics* 17(2). Article 16.

Nagy, Naomi & Brook, Marisa. 2020. Constraints on speech rate: A heritage-language perspective. *International Journal of Bilingualism* 19.

Nagy, Naomi & Chociej, Joanna & Hoffman, Michol. 2014. Analyzing Ethnic Orientation in the quantitative sociolinguistic paradigm. In Lauren Hall-Lew & Malcah Yaeger-Dror, (Eds.), Special issue of *Language and Communication: New perspectives on the concept of ethnolect*. 35: 9-26.

Nagy, Naomi & Gadanidis, Timothy. 2022. Looking for Covariation in Heritage Italian in Toronto. In Beaman, Karen & Guy, Gregory (Eds.) *The Coherence of linguistic communities: Orderly heterogeneity and social meaning*. 107-126. Routledge.

Nagy, Naomi & Kochetov, Alexei. 2013. Voice Onset Time across the generations: A cross-linguistic study of contact-induced change. In Siemund,

Peter & Gogolin, Ingrid & Schulz, Monika Edith & Davydova, Julia (Eds.), *Multilingualism and Language Diversity in Urban Areas,* 19-38. Amsterdam: John Benjamins.

Nodari, Rosalba. 2017. Indexicality and aspiration in Calabrian Italian: a sociophonetic approach. *XIII Convegno Nazionale Associazione Italiana Scienze della Voce*, Pisa. (Poster.) https://www.researchgate.net/publication/321975484_Indexicality_and_aspiration_in_Calabrian_Italian_a_sociophonetic_approach

Nodari, Rosalba. 2022. *L'identità linguistica regionale degli adolescenti: aspirazione delle occlusive sorde in Calabria e percezione della varietà locale*. Roma: Aracne.

Nodari, Rosalba & Celata, Chiara & Nagy, Naomi. 2019. Socio-indexical phonetic features in the heritage language context: VOT in the Calabrian community in Toronto. *Journal of Phonetics* 73. 91-112.

Poplack, Shana & Tagliamonte, Sali. 2001. *African American English in the Diaspora*. Oxford: Blackwell.

Statistics Canada. 2017. *Toronto [Census metropolitan area], Ontario and Ontario [Province]* (table). *Census Profile*. 2016 Census. Statistics Canada Catalogue no. 98-316-X2016001. Ottawa. (Released November 29, 2017. https://www12.statcan.gc.ca/census-recensement/2016/dp-pd/prof/index.cfm?Lang=E accesso il 28 marzo 2022).

Wiley, Terrence G. 2001. On defining heritage languages and their speakers. In Peyton, Joy Kreeft & Ranard, Donald A. & McGinnis, Scott (Eds.), *Heritage languages in America: Preserving a national resource*, 29–36). McHenry, IL: Center for Applied Linguistics and Delta Systems.

DOI: 10.17469/O2106SLI000016

RACHELE SPRUGNOLI, MATTEO PELLEGRINI,
MARCO PASSAROTTI, FLAVIO M. CECCHINI

EvaLatin 1.0: un Corpus per la Valutazione delle Tecnologie del Linguaggio Applicate al Latino

Questo articolo presenta il corpus EvaLatin 1.0, sviluppato per la prima campagna di valutazione di strumenti di Trattamento Automatico del Linguaggio per il latino. La campagna si è concentrata su due analisi linguistiche, ovvero la lemmatizzazione e l'annotazione delle parti del discorso. Particolare attenzione è stata rivolta alla costruzione del corpus in modo da affrontare problematiche di variabilità di genere e diacronica del latino.[1]

Parole chiave: latino, risorse linguistiche, corpus, annotazione linguistica, trattamento automatico del linguaggio.

1. *Introduzione*

Negli ultimi anni, in seguito alla crescente disponibilità di testi in formato digitale per le lingue antiche, e soprattutto per il greco antico e il latino, si sta assistendo anche a un incremento delle risorse linguistiche e degli strumenti di Trattamento Automatico del Linguaggio (TAL) ad esse relativi (ad es. Bouma & Adesam 2017; Sprugnoli & Passarotti 2020). Dato che l'affidabilità dei risultati delle ricerche condotte con l'ausilio di tali risorse e strumenti dipende in maniera cruciale dalla loro qualità, emerge l'esigenza di una valutazione sistematica.

È a questa necessità che si propone di far fronte EvaLatin, la prima campagna di valutazione di strumenti di TAL interamente dedicata alla lingua latina, che si inserisce in un'ampia tradizione di eventi di

[1] La responsabilità principale delle Sezioni dell'articolo va attribuita come segue. Rachele Sprugnoli: §2., §3., §3.1; Flavio M. Cecchini: §3.2; Marco Passarotti: §5. Le Sezioni §1 e §4 sono da ascrivere a tutti gli autori.

valutazione tramite *shared task*; si veda ad es. SemEval (https://alt.qcri.org/semeval2018/), CoNLL (https://www.conll.org/) e – specificamente sull'italiano – EVALITA (http://www.evalita.it/). In uno *shared task*, strumenti diversi di TAL devono risolvere uno specifico compito, come la lemmatizzazione, utilizzando dati comuni a tutti i partecipanti sia in fase di addestramento che di valutazione.

Per rendere possibile una campagna di valutazione, è necessario innanzitutto mettere a disposizione dei partecipanti i dati annotati su cui addestrare i propri strumenti di TAL (*training set*), le cui prestazioni verranno poi valutate su dati diversi (*test set*). La risorsa linguistica presentata in questo articolo, denominata EvaLatin 1.0, contiene sia i testi in prosa classica del *training set*, sia quelli del *test set* – relativi in parte agli stessi autori classici del *training set*, in parte a testi di epoca medievale e di poesia classica.

L'articolo è strutturato come segue: la Sezione 2 descriverà la prima edizione della campagna EvaLatin, mentre la Sezione 3 fornirà dettagli circa i dati soffermandosi sulla loro composizione e annotazione linguistica. I risultati della campagna di valutazione saranno presentati nella Sezione 4 insieme a un'analisi di alcune importanti caratteristiche dei dati. La Sezione 5 raccoglierà una discussione finale e presenterà un breve sguardo sulla successiva edizione di EvaLatin.

2. *EvaLatin*

I risultati della prima edizione sono stati presentati al "1st Workshop on Language Technologies for Historical and Ancient LAnguages" (LT4HALA 2020: https://circse.github.io/LT4HALA/2020/), nell'ambito della conferenza internazionale "Language Resources and Evaluation".

Nell'organizzare una campagna di valutazione di strumenti di TAL per il latino, va tenuto conto del fatto che sotto la comune etichetta di "latino" ricadono testi relativi a epoche e generi letterari diversi, risultando quindi in una notevole variazione diacronica e stilistica. Ciò tende ad impattare negativamente sui risultati dell'applicazione di un modello addestrato a testi relativi ad un'altra epoca o ad un altro genere rispetto a quelli del *training set* (Ponti & Passarotti 2016). Per poter valutare l'impatto di questo problema, i due task proposti nella prima edizione di EvaLatin – lemmatizzazione e Part-of-Speech

(PoS) tagging, cioè riconoscimento delle parti del discorso – sono divisi ciascuno in tre sotto-task, denominati rispettivamente "Classical" (i dati del *test set* sono della stessa epoca e dello stesso genere letterario di quelli del *training set*), "Cross-genre" (i dati del *test set* sono di un altro genere letterario) e "Cross-time" (i dati del *test set* sono relativi a un'altra epoca).

I testi di epoca classica sono stati tratti dalla Perseus Digital Library (Smith *et al.* 2000). L'annotazione dei lemmi e delle parti del discorso è stata ottenuta applicando ai testi alcuni modelli automatici addestrati usando UDPipe (Straka & Straková 2017) sui dati annotati manualmente del corpus sviluppato presso il centro di ricerca LASLA (Verkerk *et al.* 2020); il risultato di questa analisi automatica è stato quindi controllato e corretto manualmente da due annotatori, con eventuali dubbi risolti da un terzo annotatore. I testi di epoca medievale sono stati invece annotati manualmente nell'ambito del progetto Index Thomisticus Treebank (Passarotti 2019).

3. *Dati*

I testi forniti come dati di addestramento sono di cinque autori classici: Cesare, Cicerone, Seneca, Plinio il Giovane e Tacito. Per ogni autore abbiamo selezionato circa 50.000 token annotati, per un totale di quasi 260.000 token. Ogni autore è rappresentato da testi in prosa: trattati nel caso di Cesare, Seneca e Tacito, discorsi pubblici per Cicerone e lettere per Plinio il Giovane. La Tabella 1 riassume la composizione dei dati di addestramento.

Tabella 1 - *Composizione dei dati di addestramento*

Autori	*Testi*	*# token*
Cesare	De Bello Gallico	44.818
Cesare	De Bello Civili (libro II)	6.389
Cicerone	Philippicae (libri I-XIV)	52.563
Seneca	De Beneficiis	45.457
Seneca	De Clementia	8.172
Plinio il Giovane	Epistulae (libri I-VIII)	50.827
Tacito	Historiae	51.420
TOTALE TOKEN		259.646

Per quanto riguarda i dati di test, nel sotto-task Classical abbiamo distribuito testi in prosa degli stessi autori presenti nei dati di addestramento, selezionando circa 11.000 token per ciascuno. Per il sotto-task Cross-genre, invece, abbiamo usato i Carmina di Orazio, mentre per quello Cross-time, una parte del libro IV della Summa Contra Gentiles di Tommaso d'Aquino. In altre parole, abbiamo un testo in poesia e uno di epoca medievale. La composizione dei dati del sotto-task Classical è presentata nella Tabella 2, mentre dettagli sui dati relativi agli altri due sotto-task sono riportati nelle Tabelle 3 e 4, rispettivamente.

Tabella 2 - *Composizione dei dati di test: sotto-task Classical*

Autori	*Testi*	*# token*
Cesare	De Bello Civili (libro I)	10.898
Cicerone	In Catilinam	12.564
Seneca	De Vita Beata	7.270
Seneca	De Providentia	4.077
Plinio il Giovane	Epistulae	9.868
Tacito	Agricola	6.737
Tacito	Germania	5.513
TOTALE TOKEN		56.927

Tabella 3 - *Composizione dei dati di test: sotto-task Cross-genre*

Autore	*Testo*	*# token*
Orazio	Carmina	13.290

Tabella 4 - *Composizione dei dati di test: sotto-task Cross-time*

Autore	*Testo*	*# token*
Tommaso d'Aquino	Summa Contra Gentiles (parte del IV libro)	11.556

Il corpus è rilasciato nel formato standard CoNLL-U (https://universaldependencies.org/docs/format.html), adottato nel progetto Universal Dependencies (Nivre *et al.* 2016; de Marneffe *et al.* 2021). Secondo il formato CoNLL-U, ogni testo corrisponde ad un file in cui ogni frase è rappresentata da una struttura a 10 colonne separate da tabulazioni. Una riga vuota segna la divisione tra una frase e l'altra. Nei dati di EvaLatin 2020, solo le prime 4 colonne (di

seguito dettagliate) presentano del contenuto, mentre le altre sono riempite da un trattino basso:

1. identificatore numerico per ogni token, ovvero un numero intero che riparte da 1 ad ogni frase;
2. forma della parola così come appare nel testo;
3. lemma, ovvero forma di citazione;
4. etichetta della parte del discorso "universale" (Universal Part Of Speech, o UPOS; vedi Sezione 3.2).

Inoltre, ogni frase è preceduta da due linee di commento che iniziano con un carattere di cancelletto: una linea riporta il numero identificativo della frase, l'altra il testo. Un esempio è fornito nella Figura 1.

Figura 1 - *Il formato dei dati: una frase di esempio*

```
# sent_id = 306
# text = Debere se suspicari simulata Caesarem amicitia quod
exercitum in Gallia habeat sui opprimendi causa habere
1   Debere      debeo     VERB  _  _  _  _  _  _
2   se          sui       PRON  _  _  _  _  _  _
3   suspicari   suspicor  VERB  _  _  _  _  _  _
4   simulata    simulo    VERB  _  _  _  _  _  _
5   Caesarem    Caesar    PROPN _  _  _  _  _  _
6   amicitia    amicitia  NOUN  _  _  _  _  _  _
7   quod        quod      SCONJ _  _  _  _  _  _
8   exercitum   exercitus NOUN  _  _  _  _  _  _
9   in          in        ADP   _  _  _  _  _  _
10  Gallia      Gallia    PROPN _  _  _  _  _  _
11  habeat      habeo     VERB  _  _  _  _  _  _
12  sui         sui       PRON  _  _  _  _  _  _
13  opprimendi  opprimo   VERB  _  _  _  _  _  _
14  causa       causa     NOUN  _  _  _  _  _  _
15  habere      habeo     VERB  _  _  _  _  _  _
```

3.1 Dettagli sulla lemmatizzazione

La lemmatizzazione è il processo di riconduzione di ogni forma di parola alla sua forma di citazione corrispondente all'entrata del dizionario (cioè al lemma). Le convenzioni che abbiamo seguito sono riassunte di seguito.

– I verbi sono lemmatizzati sotto la prima persona singolare del presente indicativo attivo (o passivo, nel caso dei deponenti): es. PRS.ACT.INF *accingere* 'cingere' → *accingo*; FUT.PASS.IND.1.SG/PRS.PASS.SBJV.1.SG *sequar* 'seguirò/(che io) segua' → *sequor*.
– Le abbreviazioni sono espanse: es. *L.* → *Lucius*; *s.* → *salus*.
– Il lemma dei numeri romani (es. *ccc*, *CCCXVIII*) è *numerus_romanus*. Il lemma dei numeri arabi (es. *12*, *53*) è *num_arab*. Il lemma delle parole greche (es. *εἰσήλασαν*) è *uox_graeca*.

– Le espressioni multi-parola non sono combinate in un singolo token ma ogni loro parte è analizzata separatamente: *res publica* 'repubblica/stato' è formata da due token con due lemmi e due categorie grammaticali.
– I clitici non sono separati dal token: *exercitumque* ('esercito.ACC.SG=e') ha come lemma *exercitus* e il clitico *-que* non viene analizzato separatamente.

3.2 Dettagli sull'annotazione delle parti del discorso

Nel corpus, ogni token è annotato con la propria parte del discorso. Le etichette adottate per l'annotazione sono quelle dello schema di annotazione di UD, per cui si rimanda alle linee guida del progetto.[2] Si noti che sono usate tutte le etichette a eccezione di PUNCT e SYM, indicanti rispettivamente punteggiatura e simboli, che non sono presenti nel corpus.

Qui ci limitiamo a segnalare la caratteristica di UD per cui le categorie si presentano "in coppia", distinguendo classi funzionali e lessicali: per esempio, alla classe lessicale ADJ degli aggettivi (es. *agrestis* 'agreste') corrisponde quella funzionale DET dei determinanti (es. *hic* 'questo'), che a sua volta contempla la sottoclasse dei numerali NUM (es. *mille* 'mille'). Si hanno così NOUN/PROPN (es. *mater* 'madre', *Hercules* 'Ercole') rispetto a PRON (es. *ego* 'io'), e VERB (es. *rapiebat* 'rapiva') rispetto a AUX (che include anche la copula: nel corpus solo SUM 'essere' e, solo in alcune particolari costruzioni, EO 'andare', come in *datum iri … facultatem* 'si sarebbe data la possibilità', da Cesare, Bellum Civile, fr. 448). L'unica classe lessicale invariabile in latino è quella ADV degli avverbi (es. *certe* 'certamente'); fra le restanti, tutte funzionali, notiamo la differenza fra congiunzioni coordinanti CCONJ (es. *et* 'e') e subordinanti SCONJ (es. *quod* 'che'), e che le negazioni *non*, *ne* e *haud* ricevono la classe PART (e quindi non ADV). Infine, X è una classe residuale usata per quei token a cui per vari motivi (parole straniere, passaggi lacunosi, …) non è possibile assegnare un'analisi nel sistema latino.

Segnaliamo inoltre che l'annotazione nel nostro corpus differisce in modo rilevante dagli standard di UD riguardo al verbo *sum*: anziché essere annotato uniformemente come AUX anche laddove si com-

[2] https://universaldependencies.org/u/pos/index.html.

porta come verbo copula, questa etichetta è usata solamente quando *sum* compare come ausiliare in tempi perifrastici o composti (es. ***sunt*** *demonstratae* 'sono state dimostrate'), mentre in tutti gli altri casi (es. *paratiores* ***essent*** 'fossero più preparati', *mihi animi* ***sit*** '(io) abbia l'intenzione' lett. 'mi sia d'animo di') viene analizzato come VERB.

4. *Risultati e analisi*

EvaLatin 2020 ha visto la partecipazione di 5 gruppi, di cui 3 hanno preso parte sia alla lemmatizzazione che al riconoscimento delle parti del discorso mentre 2 al solo riconoscimento delle parti del discorso. Tutti i gruppi partecipanti erano affiliati a istituzioni di ricerca non italiane (canadesi, ceche, tedesche e statunitensi). La Tabella 5 mostra l'accuratezza raggiunta dal sistema risultato migliore, ovvero una versione appositamente creata per EvaLatin di UDPipe, strumento di analisi linguistica automatica sviluppato dall'Università Carolina di Praga (Straka & Straková 2020).[3]

In generale, i testi più semplici da elaborare per tutti i sistemi automatici partecipanti sono stati "In Catilinam" di Cicerone e "De Bello Civili" di Cesare. Al contrario, i testi che hanno registrato più errori sono "Germania" di Tacito e "De Vita Beata" di Seneca. Come previsto, tutti i sistemi, compreso UDPipe, subiscono un calo nelle prestazioni quando applicati a un genere o a un periodo temporale diverso da quello dei dati di addestramento: tale calo può arrivare anche a 10 punti percentuali. In particolare, si nota una migliore accuratezza sui testi medievali di Tommaso d'Aquino che sulla poesia classica per quanto riguarda la lemmatizzazione, mentre il contrario avviene per il riconoscimento delle parti del discorso: si ha un'accuratezza maggiore sulla poesia di Orazio che sui testi di Tommaso d'Aquino.

Tabella 5 - *Accuratezza del sistema con le migliori prestazioni*

	Classical	*Cross-genre*	*Cross-time*
Lemma	96,19 %	87,13 %	91,01 %
PoS	96,74 %	91,11 %	87,69 %

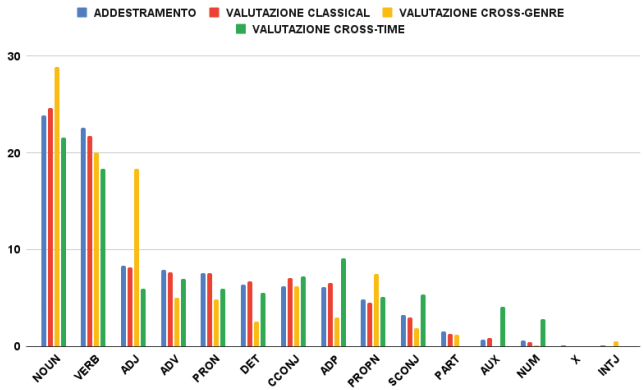
[3] L'accuratezza è il rapporto tra predizioni corrette del sistema e predizioni.

Le ragioni alla base di queste differenze possono essere ricercate in alcune caratteristiche linguistiche dei testi presi in esame. Nello specifico, per quanto riguarda la lemmatizzazione, Tommaso d'Aquino presenta un vocabolario meno ricco e vario rispetto a Orazio: il rapporto tra lemmi e token in Orazio è 0,26, mentre in Tommaso d'Aquino e nei dati di valutazione Classical lo stesso rapporto è di 0,09. Un'altra difficoltà dei Carmina risiede nel fatto che contengono un maggior numero di lemmi non presenti nei dati di addestramento (i cosiddetti lemmi *out-of-vocabulary*). Tali lemmi (ad es. *arcus* 'arco') coprono il 29% dei lemmi totali nel testo di Orazio: i lemmi *out-of-vocabulary* di Tommaso d'Aquino sono, invece, il 26% (ad es. *Christus* 'Cristo') e il 14% nei dati di valutazione del sotto-task Classical (ad es. *Archippus* 'Archippo').

Anche dal punto di vista delle parti del discorso, notiamo una distribuzione differente delle etichette nei dati, come rappresentato graficamente nella Figura 2. Se la distribuzione percentuale non varia sensibilmente tra dati di addestramento (in blu) e dati di valutazione Classical (in rosso), lo stesso non si può dire per i dati di valutazione Cross-genre (in giallo) e Cross-time (in verde). Nel testo di Orazio, infatti, abbiamo meno verbi (-3%) ma più nomi comuni (+5%), nomi propri (+3%) e aggettivi (+10%). L'alto numero di nomi propri in Orazio è stato osservato in vari studi e in letteratura non mancano indagini di natura etimologico-onomastica (Bo 1967; Roncali 2013; Paradisi 2019). Per quanto riguarda gli aggettivi, un'analisi quantitativa nei testi poetici presenti nel corpus LASLA ha rivelato che la percentuale di aggettivi rispetto al numero totale di parole è maggiore in poesia rispetto che in prosa. Infatti, se in prosa la loro percentuale media si aggira intorno all'8%, nei testi poetici del periodo classico questa è sempre superiore al 10% con evidenti punte in due delle opere di Orazio: 15,8% negli Epodi e 18,3% nei Carmina. Più preposizioni (+3%), ausiliari (+3%), congiunzioni subordinanti (+2%) e numeri (+2%) rispetto ai dati di addestramento sono invece presenti nel testo di valutazione Cross-time. Queste differenze sono da collegarsi sia a caratteristiche specifiche della prosa medievale di Tommaso d'Aquino che ad alcune discrepanze nei criteri di annotazione. Per quanto riguarda la prima motivazione, Tommaso d'Aquino tende: (i) ad usare sintagmi preposizionali laddove il latino classico userebbe la flessione dei casi per indicare il ruolo sintattico di

un sintagma nominale (Palmer 1954); (ii) a sostituire la costruzione *accusativus cum infinitivo* con proposizioni subordinate introdotte da congiunzioni subordinanti come *quia*/*quod*/*ut*; (iii) a citare frequentemente la Bibbia riportando il numero dei versi. Il maggior numero di ausiliari è, invece, da attribuire ad una discrepanza nei criteri di annotazione in quanto nel testo di Tommaso d'Aquino, tratto da un corpus annotato in altro contesto (vedi Sezione 3), sono annotate come ausiliari anche le copule verbali, cosa che non avviene negli altri testi usati in EvaLatin 2020 (Bamman 2008).

Figura 2 - *Confronto della distribuzione percentuale delle etichette relative alle parti del discorso nei dati*



È infine utile soffermarsi rapidamente sulle tipologie di errori che il sistema risultato vincitore della competizione commette nei due task proposti.

Per quanto riguarda la lemmatizzazione, la Tabella 6 mostra i 5 lemmi che presentano il maggior numero di discrepanze tra l'annotazione proposta dal sistema di Straka & Straková (2020) e quella dei dati del test set relativamente al sotto-task Classical.

Tabella 6 - *Errori di lemmatizzazione più frequenti in Straka & Straková (2020)*

Lemma (Gold)	*N. Errori*
qui	78
quod	58
quis	40
numerus_romanus	26
bonum	24

Si può notare che tra i lemmi che presentano un maggior numero di errori compaiono il pronome relativo *qui* e il pronome indefinito/interrogativo *quis*, che in molte forme flesse sono indistinguibili dal punto di vista formale e vengono dunque spesso confusi l'uno con l'altro. Inoltre, alcune forme sono identiche ad elementi invariabili e dunque lemmatizzate come tali, in particolare il nominativo/accusativo neutro *quod* come SCONJ, l'ablativo maschile/neutro *quo* come SCONJ o come ADV, e l'accusativo femminile *quam* come ADV. Queste stesse ragioni motivano, nel senso opposto, la frequente confusione del lemma *quod* come SCONJ con l'omonima forma flessa del pronome relativo (PRON).

Compare poi il nome *bonum* 'bene', che altro non è che il neutro sostantivato dell'aggettivo *bonus* 'buono', con cui condivide molte forme flesse e con cui è dunque sovente confuso. Infine, molti numeri romani non vengono riconosciuti come tali e lemmatizzati di conseguenza, in alcuni casi anche per via di omografie con altre forme (ad es. *ii* confuso con la forma di nominativo plurale maschile del pronome dimostrativo *is*).

I problemi evidenziati riguardo alla lemmatizzazione si intrecciano con l'analisi dal punto di vista morfosintattico delle parti del discorso. Basandoci sulla capacità del sistema di Straka & Straková (2020) di riconoscerle correttamente e misurandone la precisione relativa a ciascuna parte del discorso, rileviamo quattro principali nuclei di confusione, tutti con punteggi inferiori al 95%: uno che coinvolge i numerali (NUM; precisione del 66,9%), uno le congiunzioni subordinanti (SCONJ; 90,9%), uno gli ausiliari (AUX; 92,3%), e infine uno gli aggettivi (ADJ; 92,3%).

Per quanto riguarda le SCONJ (e in misura minore anche gli avverbi, che si attestano sotto il 96% di precisione, mentre ADP, CCONJ, PART e INTJ sono tutte sopra il 99%), trattandosi di una categoria composta da parole morfologicamente invariabili, l'assegnazione della parte del discorso va di pari passo con la lemmatizzazione, e si ripresentano così le problematiche appena discusse per quest'ultima.

Osserviamo tuttavia altri due fattori che hanno apparentemente influito sugli errori del modello di UDPipe in oggetto: da una parte alcune oscillazioni nell'annotazione dell'insieme di addestramento e/o test, e dall'altra la non contiguità di token appartenenti allo stesso sintagma.

Il primo fattore è ben illustrato dagli errori nella classe dei numerali (NUM): qui la precisione molto bassa è fortemente influenzata dalle molte occorrenze di forme del lemma *unus* 'uno', che, mentre nell'insieme di addestramento sono uniformemente annotati come NUM, nel test compaiono (93 occorrenze) solo come DET, rimanendo però altrimenti indistinguibili: il sistema li etichetta tutti come NUM, tranne 1 come ADV, così sbagliando.

Il secondo fattore può essere invece esemplificato da alcuni errori rispetto alla classe degli aggettivi (ADJ). Qui osserviamo come siano coinvolte soprattutto le categorie NOUN e VERB, sia in un verso che nell'altro, cioè queste tre etichette vengono (relativamente) di frequente scambiate fra loro. Infatti, se, su 4.636 ADJ, 169 sono stati etichettati come NOUN e 111 come VERB, fra i NOUN le maggiori incertezze insorgono soprattutto con ADJ e VERB (rispettivamente 202 e 93 su 14.019 token), e fra i VERB con ADJ e NOUN (90 e 80 su 12.359). In latino la flessione aggettivale è quasi indistinguibile dalla prima, seconda o terza declinazione dei nomi (a seconda di classe e/o genere), e i contesti sintattici in cui queste due categorie lessicali compaiono sono estremamente simili (ad es., sia un aggettivo che un nome possono fungere da testa di un sintagma nominale). Il sistema si trova così a dover operare scelte su criteri molto sottili e a volte contrastanti con quello che ha appreso durante l'addestramento. Per esempio, osserviamo che tre volte il sistema etichetta una forma *medio* o *medium* afferente a un ADJ *medius* 'medio' come un NOUN con lemma *medium* (2 volte) o *medius* (1 volta): nei dati d'addestramento osserviamo che effettivamente le forme *medio* e *medium* si dividono fra ADJ (lemma *medius*; 13 e 11 occorrenze a testa) e NOUN (lemma

medium 'il mezzo'; 15 e 10 occorrenze), con solo una lieve preferenza per l'aggettivo, e capiamo che il sistema propende per la seconda opzione quando interpreta tale forma come testa di un sintagma in quanto non adiacente a un nome di cui potrebbe essere attributo, ad es. *in medium reciperent agmen* '[che] li salvaguardavano nel mezzo dello schieramento [lett. nel medio schieramento]' (Cesare, Bellum Civile, fr. 503), dove *medium* modifica *agmen*, ma ne è separato dal predicato. Similmente, nella categoria delle SCONJ troviamo vari *cum* 'quando' erroneamente etichettati come adposizioni (ADP), e in quella degli ADV *modo* 'soltanto' analizzato come una forma del NOUN *modus* 'maniera'.

5. *Conclusioni e lavori futuri*

In questo articolo è stato presentato il corpus EvaLatin 1.0, sviluppato per la prima campagna di valutazione di strumenti di TAL per il latino e concentrata sulla lemmatizzazione e l'annotazione delle parti del discorso.

La campagna EvaLatin è stata avviata innanzitutto per fare il punto in merito allo stato delle prestazioni degli strumenti di TAL per la lingua latina; tale esigenza riflette il fatto che il latino sia da considerarsi una lingua per cui sono ormai disponibili numerose risorse linguistiche di diverso tipo, che iniziano a garantire una copertura testuale e lessicale sufficiente dell'ampio spettro diacronico e diatopico lungo cui questa lingua si estende. In un circolo virtuoso, proprio la disponibilità crescente di corpora annotati metalinguisticamente che raccolgono testi latini di diversa epoca, provenienza e genere consente di addestrare strumenti di TAL di tipo probabilistico capaci di garantire buoni valori di accuratezza e, quindi, di facilitare lo sviluppo di ulteriori corpora annotati.

Certamente restano molte le questioni da risolvere a livello di TAL del latino e una campagna di valutazione come EvaLatin serve proprio a farle emergere. Su tutte si impone la difficoltà di portabilità dei modelli addestrati lungo l'arco diacronico e stilistico dei testi latini. Se questa è una sfida aperta per il mondo del TAL, essa si configura anche come una solida fonte d'informazioni per chi si occupa di questioni linguistiche e letterarie del latino. A tal proposito, un'indagine dettagliata non solo della quantità e distribuzione, ma anche del tipo

di errori di lemmatizzazione e attribuzione delle parti del discorso commessi dagli strumenti di TAL può fornire indizi in merito alle differenze lessicali e morfosintattiche tra i testi usati in fase di addestramento di un modello e quelli su cui il modello è stato applicato.

La svolta empirista nel mondo del TAL, il cui stato dell'arte consiste in strumenti probabilistici addestrati sulla base di evidenza empirica, si accosta dunque fertilmente alla grande disponibilità di testi latini su supporto digitale che oggi è nelle mani dei ricercatori. Da sempre gli studiosi di lingue classiche hanno avuto un rapporto stretto con il dato testuale, unica voce che ancora risuona di lingue che non hanno più parlanti nativi, ma mai come ora quel dato è stato disponibile tanto facilmente, velocemente e ampiamente: ciò solleva la necessità di strumenti che lo possano analizzare in modo automatico, valorizzando così una svolta massivamente empirista anche negli studi classici.

Il corpus EvaLatin 1.0 aspira ad essere il primo di una serie, dal momento che è prevista l'organizzazione di nuovi *shared task* per gli strumenti di TAL del latino, di volta in volta dedicati a diversi livelli di annotazione metalinguistica. La seconda edizione di EvaLatin si è tenuta il 25 giugno del 2022 a Marsiglia nell'ambito del "2nd Workshop on Language Technologies for Historical and Ancient LAnguages" (LT4HALA 2022: https://circse.github.io/LT4HALA/2022/). Oltre alla lemmatizzazione e all'annotazione delle parti del discorso, la campagna di valutazione si è concentrata sul trattamento automatico dei tratti morfologici.

Ringraziamenti

Il progetto "LiLa: Linking Latin" è finanziato dal Consiglio Europeo della Ricerca (ERC) nell'ambito del programma di ricerca e innovazione European Union's Horizon 2020 – Grant Agreement No. 769994.

Riferimenti bibliografici

Bamman, David & Passarotti, Marco & Crane, Gregory. 2008. A Case Study in Treebank Collaboration and Comparison: Accusativus cum Infinitivo and Subordination in Latin. *The Prague Bulletin of Mathematical Linguistics* 90, 109–122.

Bo, Domenico. 1967. *L'uso dei nomi propri greci come parametro del progresso artistico di Orazio*. Torino: Giappichelli.

Bouma, Gerlof & Adesam, Yvonne (a cura di). 2017. *Proceedings of the NoDaLiDa 2017 Workshop on Processing Historical Language, Gothenburg, 22-24 maggio 2017*. Gothenburg: Linköping University Electronic Press.

Chiari, Isabella & De Mauro, Tullio. 2014. The New Basic Vocabulary of Italian as a linguistic resource. In Basili, Roberto & Lenci, Alessandro & Magnini, Bernardo (a cura di), *Proceedings of the First Italian Conference on Computational Linguistics (CLiC-it 2014)*, 113–116. Pisa: Pisa University Press.

de Marneffe, Marie-Catherine & Manning, Christopher D. & Nivre, Joakim & Zeman, Daniel. 2021. Universal dependencies. *Computational linguistics* 47(2): 255–308.

Palmer, Leonard Robert. 1954. *The Latin language*. London: Faber and Faber.

Paradisi, Patrizia. 2019. Donne oraziane: onomastica e identità. *il Nome nel testo. Rivista internazionale di onomastica letteraria*, 155–167.

Passarotti, Marco. 2019. The Project of the Index Thomisticus Treebank. In Berti, Monica (a cura di), *Digital Classical Philology. Ancient Greek and Latin in the Digital Revolution*, 299–319. Berlino-Boston, De Gruyter GmbH: 299-319.

Petrov, Slav & Das, Dipanjan & McDonald, Ryan. 2012. A Universal Part-of-Speech Tagset. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12),* 2089–2096. Istanbul: European Language Resources Association (ELRA).

Ponti, Edoardo Maria & Passarotti, Marco. 2016. Differentia compositionem facit. A slower-paced and reliable parser for Latin. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, 683–688. Portorož: European Language Resources Association (ELRA).

Roncali, Renata. 2013. Orazio. In Roncali, Renata (a cura di), I *classici nella storia della letteratura latina. I poeti*, 269–340. Bari: Edizioni di Pagina.

Smith, David A. & Rydberg-Cox, Jeffrey A. & Crane, Gregory R. 2000. The Perseus Project: A digital library for the humanities. *Literary and Linguistic Computing* 15(1): 15–25.

Sprugnoli, Rachele & Passarotti, Marco (a cura di). 2020. *Proceedings of LT4HALA 2020 – 1st Workshop on Language Technologies for Historical and Ancient Languages.* Marsiglia, European Language Resources Association (ELRA).

Straka, Milan & Straková, Jana. 2017. Tokenizing, POS Tagging, Lemmatizing and Parsing UD 2.0 with UDPipe. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, 88–99. Vancouver: Association for Computational Linguistics.

Straka, Milan & Straková, Jana. 2020. UDPipe at EvaLatin 2020: Contextualized Embeddings and Treebank Embeddings. In Sprugnoli & Passarotti (a cura di), 124–129.

Verkerk, Philippe & Ouvrard, Yves & Fantoli, Margherita & Longrée, Dominique. 2020. LASLA and Collatinus: a convergence in lexica. *Studi e Saggi Linguistici*, 58(1), 95–120.

DOI: 10.17469/O2106SLI000017

MIRKO TAVONI

Allestimento, fruizione e prospettive di *DanteSearch*

La relazione illustra le funzionalità, la storia e le attuali linee di sviluppo della risorsa *DanteSearch* (https://dantesearch.dantenetwork.it/), corpus completo delle opere volgari e latine di Dante con annotazione linguistica in formato XML-TEI. Nato all'Università di Pisa all'inizio degli anni Duemila, *DanteSearch* si è sviluppato attraverso successivi progetti di ricerca e mette a disposizione funzionalità uniche di ricerca morfologica e sintattica. *DanteSearch* è una risorsa utilizzata nell'ambito di progetti di ricerca contigui quali il *Vocabolario Dantesco* (http://www.vocabolariodantesco.it/), il *Vocabolario Dantesco Latino* (http://www.vocabolariodantescolatino.it/) e il progetto ERC *LiLa-Linking Latin* (https://lila-erc.eu/#page-top). Nell'ambito del progetto *HDN-Hypermedia Dante Network* (https://hdn.dantenetwork.it/) *DanteSearch*, insieme con *DanteSources* (https://dantesources.dantenetwork.it/), viene ripensato in ottica di web semantico, al fine di costituire una base di conoscenza fondata su logiche calcolabili (RDF, OWL) superando i limiti della singola struttura gerarchica imposta da XML.

Parole chiave: Dante Alighieri, linguistica dei corpora, storia della grammatica italiana, italiano antico, latino medievale.

1. *Introduzione*

Sono onorato dell'invito a tenere questa relazione plenaria e felice di avere con ciò l'occasione di presentare anche a linguisti non italianisti uno strumento di ricerca – *DanteSearch: corpus delle opere volgari e latine di Dante con annotazione morfologica e sintattica* – a cui hanno lavorato con me nell'arco di vent'anni giovani ricercatori di varie generazioni: un'impresa in divenire passata attraverso diverse stagioni tecnologiche che si sta evolvendo nella logica del web semantico.

DanteSearch ha la sua ormai lontana origine nell'ambito del progetto di ricerca di interesse nazionale (PRIN 1999) di cui sono stato coordinatore nazionale, intitolato *La memoria testuale. Edizioni,*

studi e strumenti per l'analisi computazionale del patrimonio italiano, che portò alla creazione di molte collezioni e strumenti di filologia digitale presso diverse Università italiane, fra cui il nucleo di quella che diventerà poi la Biblioteca italiana, la più grande biblioteca digitale di testi della letteratura italiana (http://www.bibliotecaitaliana.it/).

L'archivio testuale dantesco prodotto allora (una prima presentazione in Tavoni 2005) presentava diversi motivi di interesse: il corpus era completo; includeva le opere latine, molto meno reperibili on line di quelle volgari; i testi digitali garantivano un'affidabilità ben diversa dalla massa dei testi incontrollatamente reperibili in rete; e soprattutto erano stati non solo lemmatizzati ma anche arricchiti di una marcatura morfologica molto capillare, che consentiva di svolgere sull'*opera omnia* di Dante ricerche linguistiche fino ad allora impossibili.

Quella prima versione è stata successivamente ampliata e perfezionata da diversi punti di vista, in particolare nell'ambito del PRIN 2009 *Morfosintassi e corpora informatici dell'italiano antico* coordinato da Lorenzo Renzi e successivamente da me.

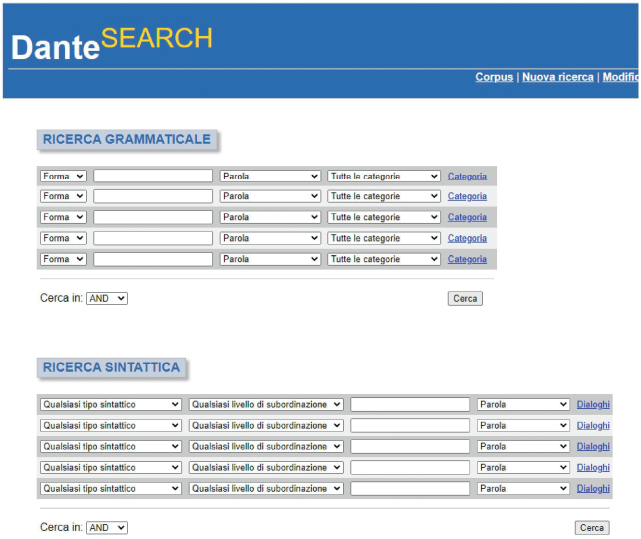
La squadra dei codificatori grammaticali del progetto originario era coordinata da Samuela Brunamonti, e il motore di ricerca XCDE (XML Compressed Document Engine: http://pages.di.unipi.it/ferragina/Libraries/xcde/index.htm) era stato ideato da Paolo Ferragina, brillante algoritmista, allora giovane ricercatore di Informatica e docente di Information retrieval nel neonato corso di laurea in Informatica umanistica.

È ovviamente superflo ricordare che la TEI-Text Encoding Initiative (http://www.tei-c.org/index.xml), «is a consortium which collectively develops and maintains a standard for the representation of texts in digital form». La conformità allo standard internazionale XML-TEI è – o almeno era, all'epoca – un requisito imprescindibile per qualunque impresa di filologia digitale: per rendere le codifiche indipendenti dalle tecnologie, hardware e software, in continua evoluzione, e così renderle permanenti nel tempo; scambiabili fra progetti di ricerca diversi che usano tecnologie diverse; riusabili anche a fini diversi da quelli per i quali sono state prodotte. La codifica grammaticale secondo lo standard XML-TEI fu curata da Elena Pierazzo, che aveva appena terminato il suo dottorato in Filologia italiana alla Scuola Normale Superiore ed era una colonna del nostro già ricordato corso di laurea in Informatica umanistica, ed era destinata a una bril-

lante carriera nel Regno Unito (King's College London) e in Francia (Université Grenoble 3, Université de Tours-CESR), e ad assumere ruoli di responsabilità, appunto, nel Consorzio TEI, e a divenirne, dal 2012 al 2014, Chair.

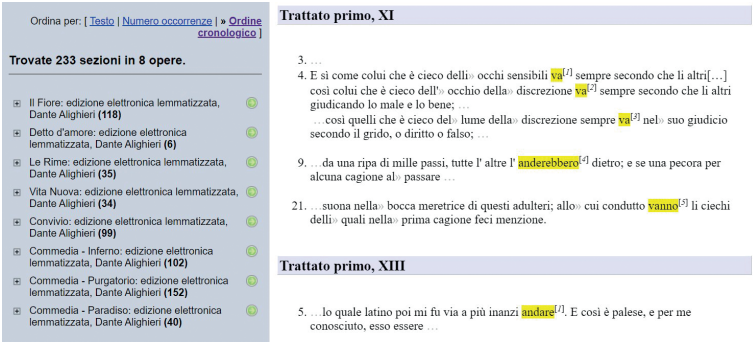
Chi si colleghi oggi al sito di *DanteSearch* (https://dantesearch.dantenetwork.it/), dopo aver cliccato su "Nuova ricerca" si trova davanti a una schermata divisa in due parti: "Ricerca grammaticale" (nel senso di morfologica) e "Ricerca sintattica" (Fig. 1).

Figura 1 - *Maschera di ricerca iniziale di DanteSearch*



Nella "Ricerca grammaticale" il primo campo permette di scegliere tra ricerca per "Forma" o per "Lemma"; il secondo tra ricerca per "Parola" (s'intende parola intera), "Sottostringa" (cioè qualunque parte di parola), "Prefisso" o "Suffisso" (naturalmente non in senso morfologico, ma nel senso di parte iniziale o finale di parola), o "Espressione regolare", o "Tutte le occorrenze". Per esempio, posso digitare nel primo campo la stringa 'fortuna' come "Forma" e come "Parola", il che mi darà come risultato 44 occorrenze della parola intera *fortuna* in 11 opere, sia in volgare sia in latino. Posso digitare 'andare' come "Lemma" e come "Parola", ottenendo 586 occorrenze in 8 opere, evidentemente solo in volgare (Fig. 2).

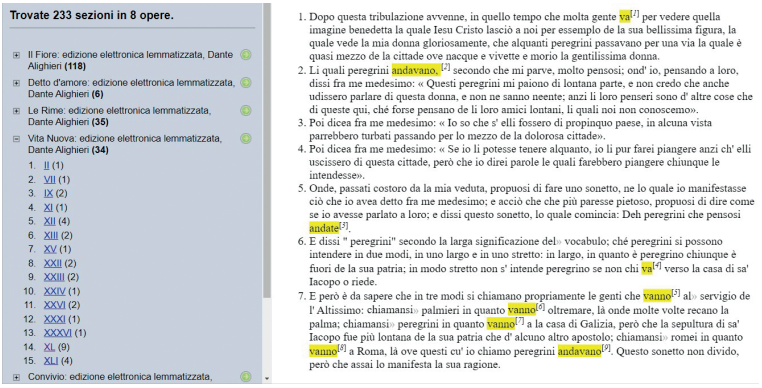
Figura 2 - *Risultati di ricerca per il lemma* andare*, visualizzazione ristretta*



E posso digitare parti di parole (parti qualunque, o iniziali, o finali), come forme o come lemmi. Il risultato di una ricerca di questo tipo appare inizialmente, come vediamo nella Fig. 2, in forma di elenco dei testi che contengono la parola, o la parte di parola, o l'insieme di parole richiesti, col relativo numero di occorrenze; questo elenco può essere ordinato, cliccando sulla riga in alto, o per "Testo" (cioè in ordine alfabetico per titolo di testo), o per "Ordine cronologico" (cioè secondo la successione delle date di composizione dei testi), o per "Numero occorrenze" (cioè ordinando i testi per numero decrescente di occorrenze contenute in ciascun testo). A partire da un elenco di questo tipo, si possono scegliere due diverse visualizzazioni delle occorrenze. La prima si ottiene cliccando sul tondino con freccetta a fianco del singolo testo, e consiste nella visualizzazione di tutte le occorrenze entro il testo, in ordine di testo, evidenziate in giallo entro contesti di poche righe ritagliati dal testo, come vediamo nella Fig. 2.

La seconda visualizzazione, che vediamo nella Fig. 3, si ottiene cliccando su un titolo di testo dell'elenco. Si apre allora nel menu a tendina l'articolazione interna delle "Sezioni" (cioè canti, capitoli, ecc.) di quel testo in cui è presente almeno una occorrenza della stringa ricercata: p.es. qui i capitoli della *Vita nova*. Cliccando su una di queste, nella finestra a destra compare il testo integrale della sezione, nel quale le occorrenze della stringa ricercata sono evidenziate in giallo, come vediamo.

Figura 3 - *Risultati di ricerca per il lemma* andare*, visualizzazione estesa*

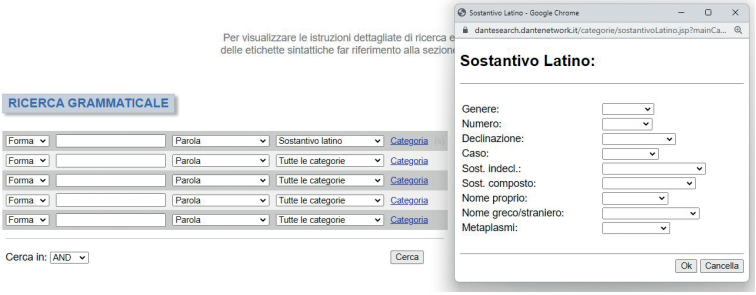


Dunque, tornando alla maschera di ricerca iniziale (Fig. 1), i primi due campi nella riga di ricerca (Lemma/Forma e Parola/Sottostringa, ecc.) servono per la ricerca lessicale. Il terzo campo introduce la vera novità: la ricerca per categorie grammaticali. Ogni occorrenza di ogni parola nel testo, infatti, è stata etichettata come appartenente a una parte del discorso, e più specificamente come caratterizzata dai corrispondenti possibili tratti morfologici. Il terzo campo della riga, che appare inizialmente occupato dalla dizione "Tutte le categorie", permette di scegliere fra Volgare e Latino e all'interno dell'uno o dell'altro fra Verbo, Sostantivo, Aggettivo, ecc.; operata questa selezione, cliccando sul link a destra "Categoria" si apre una finestra che consente di selezionare i tratti morfologici propri di quella categoria ovvero parte del discorso.

Per esempio, nella Fig. 4 si vedono quali sono i tratti morfologici selezionabili per il "Verbo volgare"; nella Fig. 5 quali sono i tratti morfologici selezionabili per il "Sostantivo latino".

Figura 4 - *Tratti morfologici ricercabili per "Verbo volgare"*

Figura 5 - *Tratti morfologici ricercabili per "Sostantivo latino"*



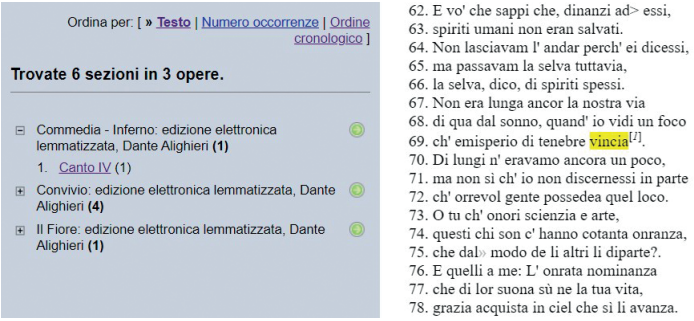
La ricerca lessicale e quella grammaticale possono essere attivate l'una indipendentemente dall'altra, oppure in combinazione. Se voglio cercare tutte le occorrenze che rispondono a una certa definizione grammaticale, senza porre nessuna restrizione di tipo lessicale, sceglierò, nel secondo campo, "Tutte le occorrenze", il che neutralizza la selezione lessicale. Per esempio posso cercare tutti i pronomi dimostrativi volgari (che sono 2.995 in 8 opere); se seleziono solo i femminili singolari si riducono a 550; solo i maschili plurali 306, e così via.

Oppure posso porre a questa ricerca grammaticale una restrizione lessicale. In questo caso non selezionerò, nel secondo campo, "Tutte le occorrenze" ma per esempio "Prefisso", per restringere la ricerca dei pronomi dimostrativi a quelli che iniziano con – poniamo – *cost-*: che risultano essere 148 (*costui*, *costu'*, *costei*, *coste'*, *costoro*, *costor*) in 7 opere.

Un esempio euristico di ricerca combinata per tratti grammaticali e stringa lessicale consiste nella ricerca delle III persone singolari degli imperfetti indicativi della III, o in alternativa della II coniugazione,

con la restrizione che si tratti di forme terminanti in *-ia* (restrizione che si ottiene digitando *ia* come "Forma" e come "Suffisso", cioè semplicemente come parte finale della parola). Ne risulta che gli esempi della III coniugazione sono molto numerosi: 82 occorrenze distribuite su tutte le opere volgari tranne il *Detto d'Amore*. Invece gli esempi della II coniugazione (Fig. 6) risultano essere pochissimi.

Figura 6 - *III p.s. imperfetto indicativo II coniugazione in* -ia

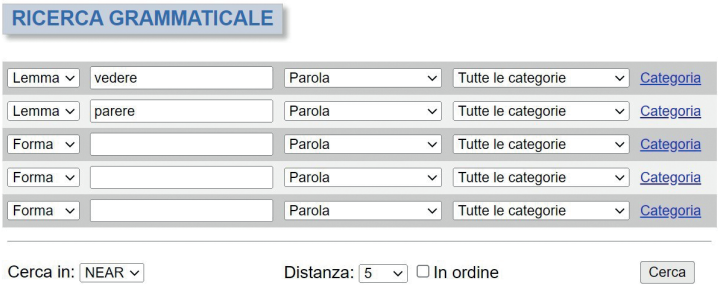


L'unica occorrenza nella *Commedia* è in *If* IV 69 «ch'emisperio di tenebre vincia» (R: *tuttavia* : *via*), ma solo perché il codificatore, seguendo la maggioranza degli interpreti, ha interpretato *vincia* come voce del verbo *vincere* e non di *vincire*, 'avvincere', interpretazione minoritaria. Ma questo quadro quantitativo d'insieme depone fortemente a favore di questa seconda interpretazione, perché non c'è in tutta la *Commedia*, oltre a questo caso dubbio, neanche un caso sicuro di imperfetto della II coniugazione in *-ia*, per sicilianismo (Tavoni 2011: 592-593; 2020b: 163-168).

Fin qui abbiamo visto singole ricerche, di natura lessicale e/o grammaticale, esprimibili su una sola riga della maschera di ricerca. Ma questa, come si vede in Fig. 1, ne annovera ben cinque. Su ogni riga si può formulare una ricerca, e due o tre o anche quattro o cinque ricerche si possono combinare con i classici operatori booleani AND, OR, NOT, NEAR. Da notare che la ricerca AND dà come risultato qualunque compresenza di parole all'interno della "Sezione" dell'opera (un intero canto della *Commedia*, un intero capitolo del *Convivio*, ecc.). Per cercare, come accade più spesso di voler fare, risultati compresenti a breve distanza, bisogna usare l'operatore NEAR. Il quale, una volta selezionato, fa aprire il box "Distanza", cioè a quante

parole di distanza al massimo devono cooccorrere le parole ricercate, con la possibilità di indicare se devono comparire nell'ordine in cui è stata formulata la richiesta oppure no. Per esempio, nell'ambito di una ricerca sulla fenomenologia della visione, mi può interessare cercare quanto spesso cooccorrono, e quanto da vicino, e in quali opere, i verbi *vedere* e *parere* (che in italiano antico ha notoriamente un valore più forte, fisico: non 'sembrare', ma 'apparire', con impatto visivo). Scriverò dunque questi due verbi, come "Lemma", in due righe successive, e li cercherò con la funzione NEAR (Fig. 7). Posso ripetere la ricerca variando la distanza, e troverò che le cooccorrenze sono numerosissime, e in tutte le opere volgari: a distanza 20 sono 322, a distanza 10 sono 201, a distanza 5 – come qui – sono 114. In latino, invece, i lemmi *appareo* e *video* non cooccorrono mai, nemmeno a distanza 20. Questa "mutua attrazione" del vedere e dell'apparire è il portato di un'ossessione stilnovistica, cavalcantiana, che resta viva per tutta la carriera poetica, ma solo poetica, di Dante, e non compare mai in un universo di discorso profondamente diverso come quello della trattatistica latina.

Figura 7 - *Ricerca di* vedere *e* parere *cooccorrenti entro distanza di 5 parole*



Passiamo ora alla ricerca sintattica. La strada per questa risorsa, completamente nuova, è stata aperta da Sara Gigli, che alla codifica sintattica esaustiva della Commedia, con tutti i problemi interpretativi che essa evidentemente comporta, ha dedicato la propria tesi di dottorato (Gigli 2004; e cfr. 2003, 2007, 2015). Sara Gigli, prendendo a riferimento l'impianto teorico-descrittivo della *Grande grammatica italiana di consultazione* di Renzi-Salvi-Cardinaletti (1988-1995), quindi della *Grammatica dell'italiano antico* di Salvi-Renzi (2010), si è posta il pro-

blema di come applicare quell'apparato di categorie e nozioni sintattiche al testo della *Commedia*, e di come tradurlo in un sistema di codifica capace di descrivere il testo per intero, in tutta la sua complessità e varietà di aspetti. Con la collaborazione di Elena Pierazzo per quanto riguarda il formalismo XML-TEI, Sara Gigli ha portato a termine in modo eccellente un compito tanto oneroso quanto irto di problemi interpretativi, filologici e linguistici, e con ciò ha messo a disposizione della comunità scientifica uno strumento unico e sofisticatissimo.

Come si vede nei campi 3 e 4, anche la ricerca per categorie sintattiche può essere combinata con restrizioni di tipo lessicale. La differenza è che il terzo campo ammette solo la ricerca per "Forma"; che può essere specificata, nel quarto campo, come "Parola" o "Sottostringa". Se invece voglio fare una ricerca esclusivamente per categorie sintattiche, senza restrizioni lessicali, anche qui, come nella ricerca grammaticale, nel quarto campo selezionerò "Tutte le occorrenze". I campi per la ricerca sintattica sono il primo, dedicato al tipo di frase, e il secondo, dedicato al grado di subordinazione.

Nel primo campo, il sistema permette di interrogare il corpus su due livelli: per tipi sintattici (una trentina, articolati in frasi principali e coordinate a una principale, subordinate e coordinate a una subordinata, psudocoordinate, parentetiche e cordinate a una parentetica) e per sottotipi (più di 300: dichiarativa illocutiva, coordinata congiuntiva esclamativa, interrogativa di tipo x, ecc. ecc.). La prima ricerca, per tipi, consente di ricercare (e conteggiare) tutte le relative insieme, tutte le consecutive insieme, tutte le ipotetiche insieme, ecc.; la seconda consente di ricercare sottotipi di frase in modo estremamente analitico.

Il secondo campo permette di specificare il grado di subordinazione al quale si vuole restringere la ricerca del sottotipo di frase definito nel campo precedente. Per esempio voglio vedere le interrogative alternative, ma solo quelle che siano subordinate di III grado. Ce ne sono solo 3, di cui una nell'*Inferno*: «Dintorno mi guardò, come talento / avesse di veder s'altri era meco», (X 55-56). Si aggiunga la possibilità di usare gli operatori AND e NEAR, che consentono di ricercare cooccorrenze di tipi di frase, a una determinata distanza l'uno dall'altro ed eventualmente nell'ordine voluto – per esempio una causale immediatamente seguita da una finale (come «Qual si lamenta perché qui si moia / per viver colà su, non vide quivi...», *Pd* XIV 25-26) – e si percepirà il numero vertiginoso di combinazioni di

ricerca sintattica che si possono costruire, per rispondere ai più vari percorsi mentali del ricercatore, alimentando una catena virtuosa di risposte a curiosità di ricerca, a loro volta passibili di stimolare nuove intuizioni e nuove domande.

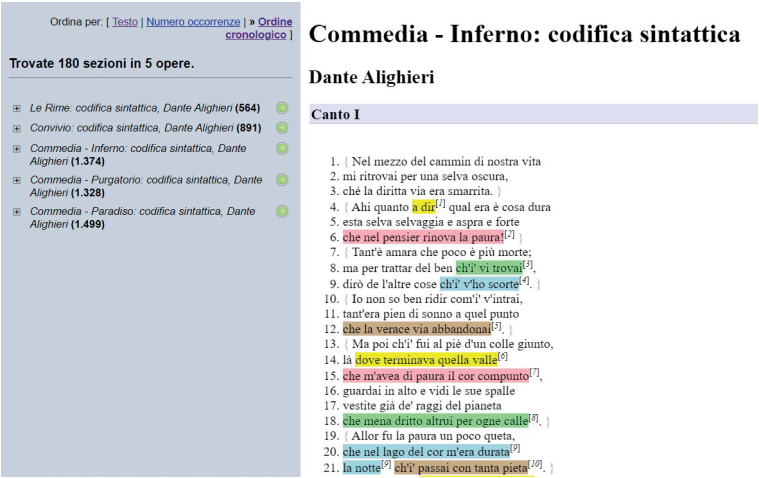
Vediamo ora un solo esempio di una ricerca sintattica generalissima, finalizzata a una verifica molto specifica. In un mio ormai antico lavoro (Tavoni 2002) ho proposto un argomento sintattico che non era mai stato addotto a proposito della vessatissima questione di «forse cui Guido vostro ebbe a disdegno» (*If* X 63): cioè il principio che i grammatici generativi chiamano "insularità delle relative", secondo il quale nessun elemento appartenente a una frase relativa può essere dislocato a sinistra del pronome relativo che la introduce – e quindi *forse* NON può appartenere alla relativa, cioè significare '...a colui che forse il vostro Guido ebbe a disdegno'. Questa era l'interpretazione preferita da Contini, ma a mio giudizio è agrammaticale, dunque impossibile. All'epoca non disponevo della ricerca sintattica di *DanteSearch*. Oggi è possibile sottoporre la mia tesi a una verifica esaustiva, semplicemente richiamando insieme tutte le proposizioni relative in tutta la *Commedia*. Lancio dunque la ricerca sintattica: "Relativa" (senza ulteriore specificazione) – "Tutte le occorrenze".

Il risultato è un totale di 4.201 frasi relative nella *Commedia*. Purtroppo, al momento non è possibile chiedere al sistema di fare anche l'ultima operazione materiale per noi, cioè chiedergli di estrarre da queste 4.201 frasi quelle in cui eventualmente il pronome relativo non occupi la prima posizione. Non resta dunque che scorrere tutte queste frasi, dalla prima all'ultima: operazione che comunque non richiede più di qualche ora. Nella Fig. 8 appaiono, nel corso di un tale scorrimento di tutti i contesti dell'*Inferno*, quelli compresi nei primi 21 versi del I canto.

Come abbiamo già notato, nella maschera della ricerca sintattica il terzo campo prevede solo la possibilità di ricercare per forma, non per lemma. Questo significa che la lemmatizzazione non è fruibile all'interno della ricerca sintattica, come non lo è la marcatura morfologica, perché la lemmatizzazione-marcatura morfologica da una parte, e la marcatura sintattica dall'altra, sono state realizzate su due distinte copie (in partenza identiche) dello stesso corpus testuale. Sarebbe stato impossibile sovrapporre le due marcature l'una sull'altra: è questo un limite del formalismo di codifica XML che rende impossibile, allo

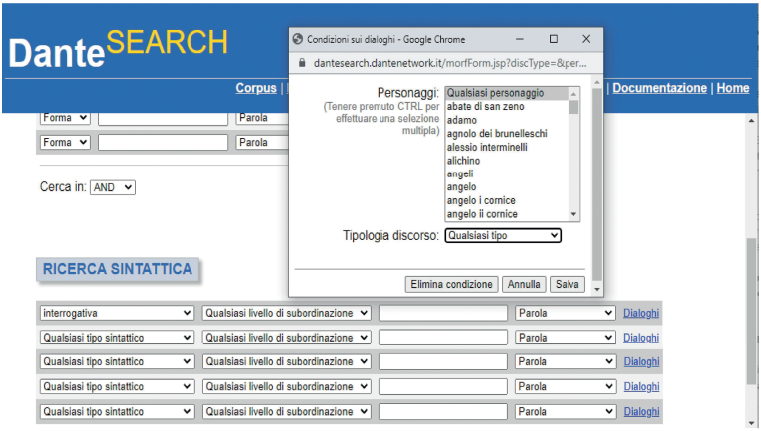
stato dell'arte, interrogare insieme per lemmi, per categorie morfologiche e per categorie sintattiche. Sarebbe invece molto interessante poter combinare questi diversi tipi di ricerca. Come superare questo limite di XML costituisce precisamente il problema che l'attuale fase di sviluppo di *DanteSearch* ha davanti, e su questo mi soffermerò alla fine della mia relazione.

Figura 8 - *Le frasi relative nel I canto dell'*Inferno



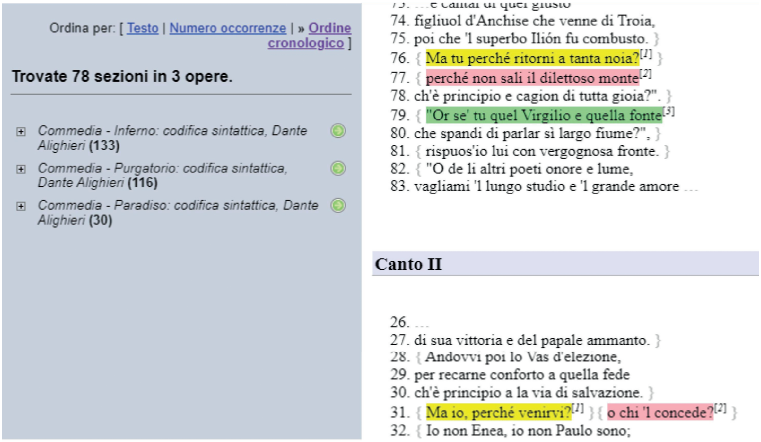
Infine, la ricerca sintattica sul parlato dei personaggi. Per dare un'idea delle possibilità offerte dalla ricerca sintattica limitata ai discorsi o pensieri pronunciati o pensati da personaggi della *Commedia*, codifica introdotta da Marta D'Amico sulla base della sua tesi di laurea magistrale dal suggestivo titolo *La sintassi dell'aldilà*, vediamo anzitutto che, cliccando su "Dialoghi" nella maschera di ricerca, si apre un box che permette di selezionare i "Personaggi" (tutti i personaggi della *Commedia* vi compaiono in ordine alfabetico) e/o la "Tipologia di discorso" (Fig. 9). Lanciamo dunque per prima cosa una ricerca sintattica scegliendo, nella Finestra "Dialoghi", "Qualsiasi personaggio" e "Qualsiasi tipologia di discorso", per andare a verificare la presenza e distribuzione di un certo tipo di frase – per esempio le principali interrogative – all'interno delle parti mimetiche del poema.

Figura 9 - *Ricerca sul parlato dei personaggi della* Commedia



Ed ecco il risultato (Fig. 10): 133 frasi principali interrogative nell'*Inferno*, 116 nel *Purgatorio*, 30 nel *Paradiso*. La sproporzione fra la terza cantica e le prime due è clamorosa, e rende evidente (ciò che resta invece invisibile "a occhio nudo") una diversa "condizione cognitiva" dei beati rispetto ai dannati e ai penitenti. Mentre a destra compare un esempio della ricorrenza di interrogative nei dialoghi fra I e II canto dell'*Inferno*.

Figura 10 - *Le frasi principali interrogative nel parlato della* Commedia



È solo un esempio. La ricerca di Marta D'Amico (2009 e 2015) esplora una grande messe di dati desumibili dal sistema di interrogazione e li combina con altri dati, come gli introduttori, i tempi e i modi verbali, la semantica delle frasi, ricavabili per altra via, traendone conclusioni molto interessanti circa la caratterizzazione linguistica in rapporto all'identità storica dei personaggi, individualmente e per categorie in senso lato sociali; circa i fenomeni di mimesi del parlato che l'autore sfrutta graduandoli sapientemente (su questo cfr. anche Tavoni 2020a); e circa i diversi, appunto, stili cognitivi che riflettendosi in strutture sintattiche privilegiate qualificano i discorsi delle tre cantiche.

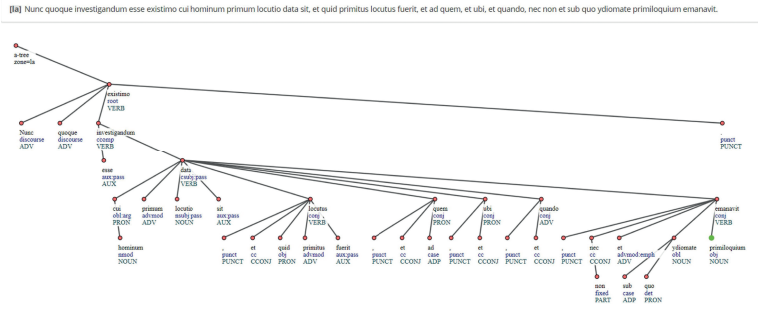
DanteSearch è dunque un sistema aperto. E per indicare in quali direzioni è attualmente aperto, indicherò tre progetti di ricerca con i quali collabora, con reciproca utilità, e la propria attuale prospettiva di sviluppo in direzione del web semantico.

I due primi progetti con cui *DanteSearch* collabora sono il *Vocabolario Dantesco* (http://www.vocabolariodantesco.it/), impresa congiunta dell'Accademia della Crusca e dell'Istituto CNR OVI – Opera del Vocabolario Italiano diretta da Paola Manni e Lino Leonardi, e il *Vocabolario Dantesco Latino* (http://www.vocabolariodantescolatino.it/), progetto parallelo e strettamente collegato al *Vocabolario Dantesco*, realizzato, oltre che dalla Crusca e dall'OVI, dalla Fondazione Franceschini e dalla SISMEL, dalla Società Dantesca Italiana, dall'ISTI-CNR e dal Dip. di Filologia Letteratura e Linguistica dell'Università di Pisa, e coordinato da Gabriella Albanese. I due progetti hanno il fine di arrivare a dare una rappresentazione lessicografica completa della cultura bilingue di Dante, e *DanteSearch* costituisce un ovvio strumento di lavoro quotidiano dei redattori di entrambi.

Un terzo progetto con cui *DanteSearch* collabora è il progetto ERC *LiLa-Linking Latin* (https://lila-erc.eu/#page-top), Principal Investigator Marco Passarotti, che si propone di «connect and ultimately exploit the wealth of linguistic resources and NLP tools for Latin created so far», in ottica di web semantico. *Vocabolario Dantesco Latino* e *DanteSearch* collaborano con *LiLa* alla codifica sintattica delle opere latine di Dante secondo lo standard del progetto mondiale Universal Dependencies (https://universaldependencies.org/). Ecco per esempio, alla Fig. 11, la rappresentazione sintattica di una frase del *De vulgari eloquentia*, codificata da Giulia Pedonese (la *Monarchia* è

codificata da Federica Favero, le *Epistole* da Elena Vagnoni, le *Egloge* da Veronica Dadà).

Figura 11 - *Codifica sintattica di* De vulgari eloquentia*, I iv 1*

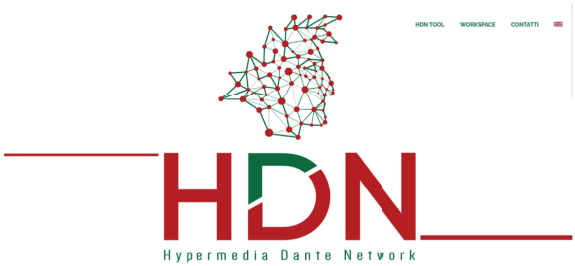


La codifica sintattica delle opere latine realizzata secondo questa grammatica a dipendenze integra così la codifica sintattica delle opere volgari secondo la grammatica a costituenti inserita in *DanteSearch*.

Lo sviluppo di DanteSearch in direzione del web semantico si realizza nell'ambito del PRIN *HDN-Hypermedia Dante Network* (https://hdn.dantenetwork.it/), Principal Investigator Michelangelo Zaccarello, professore di Filologia italiana dell'Università di Pisa; responsabile della realizzazione informatica Carlo Meghini, dirigente di ricerca dell'ISTI-CNR (Fig. 12).

Questa realizzazione informatica, in sintonia con il progetto *LiLa*, ha preso il nome di *LiDa-Linking Dante*, con l'obiettivo di digitalizzare le opere dantesche e la conoscenza a esse relativa, con ciò intendendosi anzitutto i commenti, e in particolare il secolare commento alla *Commedia*, creando una biblioteca digitale che rispetti i principi FAIR-Findable, Accessible, Interoperable and Reusable (https://www.go-fair.org/fair-principles/), e su cui si possano formulare e testare ipotesi scientifiche, eventualmente anche attraverso agenti digitali (Bartalesi-Meghini-Andriani-Tavoni 2015).

Figura 12 - *Homepage del progetto HDN-Hypermedia Dante Netwo*



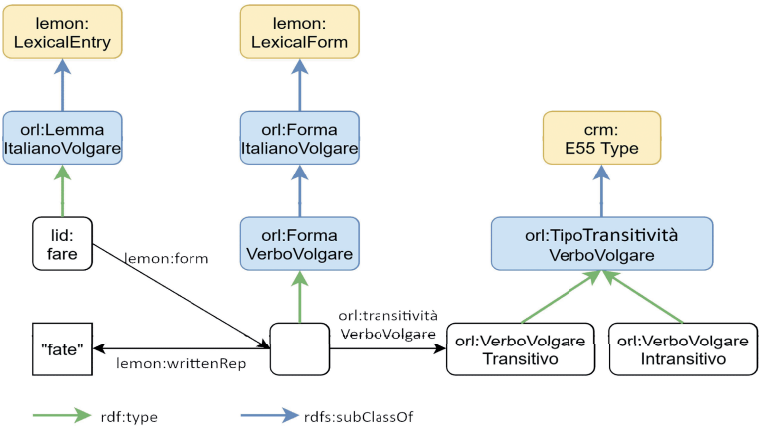
A questo punto bisogna almeno nominare il progetto *DanteSources* (https://dantesources.dantenetwork.it/), base di conoscenza sulle fonti citate nelle opere di Dante, che negli anni scorsi abbiamo sviluppato in parallelo a *DanteSearch* (vedi Bartalesi-Meghini-Metilli-Andriani-Tavoni 2017, Tavoni-Andriani-Meghini-Bartalesi-Metilli 2017). L'ambizioso obiettivo di Carlo Meghini e dei suoi collaboratori presso l'ISTI-CNR (Cesare Concordia, Chiara Paolini, Daniele Metilli, Luca Trupiano) è ora arrivare a creare una unica base di conoscenza aperta e sempre implementabile nella quale confluiscano sia *DanteSearch* sia *DanteSources*: cioè sia l'informazione linguistica (morfologica, sintattica, ma anche retorica) sia tutta l'informazione di diversissima natura sulle fonti e su quant'altro registrato dai commenti, adottando il paradigma del web semantico, e con ciò uscendo dai limiti, notati prima, della codifica in XML, perché l'eXtensible Markup Language (XML) non è adeguato alla doppia integrazione delle opere con i commenti e dei commenti tra di loro, dal momento che si fonda su una singola struttura gerarchica.

LiDa, dunque, farà uso di logiche calcolabili, che usano il Resource Description Framework (RDF) come linguaggio di base e ruotano intorno all'Ontology Web Language (OWL) per la codifica formale della conoscenza in ambiente web. L'uso di queste logiche permette di superare il problema della singola struttura gerarchica posto dall'XML, e di rappresentare così in modo formale non solo il testo dantesco, ma anche la sua esegesi e i legami che testo e commenti hanno fra loro.

In particolare, la reimplementazione di *DanteSearch* muove dal fatto che la lemmatizzazione e la marcatura morfologica non sono fruibili all'interno della ricerca sintattica, dato che XML non permet-

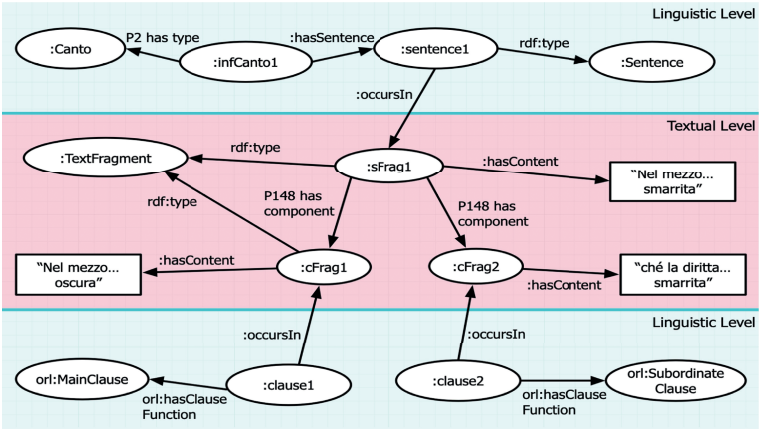
te di sovrapporre le due marcature, e si propone di rendere la lemmatizzazione fruibile non solo all'interno della ricerca sintattica ma anche di qualsiasi altra ricerca che riguardi un aspetto affrontato dal progetto *LiDa*. Per raggiungere questo scopo, le due marcature alla base di *DanteSearch* si trasformano in un unico grafo RDF, quindi in un insieme di triple, conformi a corrispondenti ontologie: un'ontologia morfologica e un'ontologia sintattica. La Fig. 13 rappresenta l'ontologia morfologica, e la Fig. 14 rappresenta l'ontologia sintattica.

Figura 13 - *L'ontologia morfologica del nuovo* DanteSearch



La trasformazione richiede dunque il prioritario sviluppo di dette ontologie, e successivamente lo sviluppo di un algoritmo di parsing dell'XML e di successiva emissione delle triple. Al momento, siamo in fase di testing dell'ontologia sintattica e di messa a punto dell'algoritmo di trasformazione della marcatura sintattica, mentre l'ontologia morfologica e la trasformazione della corrispondente marcatura è stata completata.

Figura 14 - *L'ontologia sintattica del nuovo* DanteSearch



Una volta completata anche la trasformazione della marcatura sintattica sarà possibile reimplementare le operazioni offerte dall'interfaccia *DanteSearch* come interrogazioni SPARQL al grafo risultante, sostituendo il back-end attuale di *DanteSearch* con il nuovo back-end. In questo modo, gli utenti di *DanteSearch* potranno continuare a usare la stessa interfaccia avendo in più la possibilità di combinare in modo arbitrario i due criteri di ricerca.

Queste parole di Carlo Meghini e collaboratori, che vi ho riportato, oltrepassano i limiti delle mie competenze. Hanno per me il sapore della Teoria del Tutto, che io conosco non da Stephen Hawking ma da Sheldon Cooper. E questo, in realtà, mi dà un grande senso di appagamento, perché è arrivato il momento in cui giovani informatici umanisti, che tu hai contribuito a formare, ti hanno superato, e questo ti dà la certezza che il tuo passaggio di alcuni decenni dall'Università non è stato inoperoso e non è stato inutile.

Riferimenti bibliografici

Bartalesi, Valentina & Meghini, Carlo & Andriani, Paola & Tavoni, Mirko. 2015. Towards a Semantic Network of Dante's Works and Their Contextual Knowledge. *Digital Scholarship in the Humanities*, 30(1). 28-35.

Bartalesi, Valentina & Meghini, Carlo & Metilli, Daniele & Andriani, Paola & Tavoni, Mirko. 2017. DanteSources: a Digital Library for Studying Dante Alighieri's Primary Sources. *Umanistica Digitale* 1. 119-128.

D'Amico, Marta (a cura di). 2015. *Sintassi dell'italiano antico e sintassi di Dante. Atti del seminario di studi, Pisa, 15-16 ottobre 2011*. Pisa: Felici.

D'Amico, Marta. 2009. *La sintassi dell'aldilà. Studio sulla sintassi periodale dei discorsi diretti delle anime della* Commedia *di Dante.* Università di Pisa. (Tesi di laurea specialistica in Lingua e letteratura italiana.)

D'Amico, Marta. 2015. Le interrogative dirette non canoniche nei discorsi dei personaggi della *Commedia*: pragmatica e testualità. In D'Amico, Marta (a cura di), *Sintassi dell'italiano antico e sintassi di Dante. Atti del seminario di studi (Pisa 15/16 ottobre 2011)*, 125-139. Pisa: Felici.

Gigli, Sara. 2003. *Le proposizioni consecutive nella* Commedia*: osservazioni stilistiche. In Battaglia Ricci, Lucia (a cura di), Leggere Dante,* 329-344. Ravenna: Longo.

Gigli, Sara. 2004. *Codifica sintattica della* Commedia *dantesca*, Università di Pisa, Scuola di dottorato in Studi italianistici.

Gigli, Sara. 2007. Le subordinate concessive nella *Commedia* dantesca, *Studi Linguistici Italiani*, XXXIII(2). 161-190.

Gigli, Sara. 2015. *La codifica sintattica della* Commedia *di Dante*, in D'Amico (a cura di), *Sintassi dell'italiano antico e sintassi di Dante. Atti del seminario di studi (Pisa 15/16 ottobre 2011)*, 81-95. Pisa: Felici.

Meghini, Carlo & Tavoni, Mirko & Zaccarello, Michelangelo. 2021. Mapping the Knowledge of Dante Commentaries in the Digital Context: A Web Ontology Approach. *Romanic Review*, 112/1 (*The Pleasure of Dante's Text / Il piacere del testo dantesco*, H. Wayne Storey, Guest Editor). 138-157.

Renzi, Lorenzo & Salvi, Giampaolo & Cardinaletti, Anna (a cura di). 1988-1995. *Grande grammatica italiana di consultazione*, 3 voll., Bologna: il Mulino.

Salvi, Giampaolo & Renzi, Lorenzo (a cura di). 2010. *Grammatica dell'italiano antico*, Bologna: il Mulino.

Tavoni, Mirko & Andriani, Paola & Meghini, Carlo & Bartalesi, Valentina & Metilli, Daniele. 2017. L'esplorazione delle fonti dantesche attraverso la biblioteca digitale *DanteSources*. In Persico, Thomas & Viel, Riccardo (a cura di), *Sulle tracce del Dante minore. Prospettive di ricerca per lo studio delle fonti dantesche*, 29-52. Bergamo: Sestante Edizioni.

Tavoni, Mirko. 2002. Contributo sintattico al 'disdegno' di Guido (*If* X 61-63). Con una nota sulla grammaticalità e la leggibilità dei classici, *Nuova Rivista di Letteratura Italiana*, V(1). 51-80.

Tavoni, Mirko. 2005. Un nuovo strumento informatico per lo studio di Dante (con una proposta interpretativa per *Inf.* IV 69). In De Matteis, Giuseppe (a cura di), *Dante in lettura*, 217-229. Ravenna: Longo.

Tavoni, Mirko. 2011. *DanteSearch*: il corpus delle opere volgari e latine di Dante lemmatizzate con marcatura grammaticale e sintattica. In Cerbo, Anna (a cura di), *Lectura Dantis 2002-2009. Omaggio a Vincenzo Placella per i suoi settanta anni*, t. II: *Lectura Dantis 2004 e 2005*: 583-608. Napoli: Università degli Studi di Napoli L'Orientale – Officine Grafico-Editoriali "il Torcoliere".

Tavoni, Mirko. 2015. *DanteSearch*: istruzioni per l'uso. Interrogazione morfologica e sintattica delle opere volgari e latine di Dante. In D'Amico (a cura di), *Sintassi dell'italiano antico e sintassi di Dante. Atti del seminario di studi (Pisa 15/16 ottobre 2011)*, 59-79. Pisa: Felici.

Tavoni, Mirko. 2020a. Lingua parlata e lingua scritta in Dante: appunti metalinguistici e linguistici. In Orletti, Franca & Albano Leoni, Federico (a cura di), *L'antinomia scritto/parlato*, 89-115. Città di Castello: I libri di Emil.

Tavoni, Mirko. 2020b. Lessicografia ed esegesi dantesca. In Manni, Paola (a cura di), *«S'i'ho ben la tua parola intesa». Atti della giornata di presentazione del* Vocabolario dantesco. *Firenze, Villa Medicea di Castello, 1° ottobre 2018*: 157-168. Firenze: Accademia della Crusca (Quaderni degli *Studi di lessicografia italiana*).

DOI: 10.17469/O2106SLI000018

PAOLA MANNI, ROSSELLA MOSTI

Per Dante. Il *VD* e i corpora dell'italiano antico[1]

The *Vocabolario Dantesco* (*VD*) www.vocabolariodantesco.it, created by Accademia della Crusca and Opera del Vocabolario Italiano (OVI-CNR), proposes to gather the lexicon contained in Dante's vulgar works starting from *Commedia* of which it has already published over 1000 entries. Paola Manni, director of the *VD* for Accademia della Crusca, describes its inspiring reasons and methodological criteria, focusing on the new solutions required for the lessical analysis of a single – and singular – author as Dante. Rossella Mosti, Editorial Coordinator of the *Tesoro della lingua italiana delle Origini* (*TLIO*) and member of the directive Committee of the *VD*, enters in the synergetic relationship between the two lexicographic works. In this way she demonstrates how the *corpora* and *sottocorpora* of ancient Italian, already the foundation of the *TLIO*, are utilised and interact in the environment of a new initiative that finally frames Dante's words in their historical context and underlines his extraordinary creativity.

Keywords: Dante Vocabulary, Dante Studies, Lexicography of Ancient Italian, Corpus linguistics, Digital Humanities.

1. *Il VD: principi ispiratori e fondamenti metodologici*

Non ci sono dubbi che l'ultimo quarto del XX secolo ha rappresentato un'epoca di straordinario progresso negli studi sulla lingua dantesca con iniziative che, grazie alla tecnologie informatiche, hanno veramente aperto nuovi orizzonti e gettato luce su aspetti che precedentemente si sottraevano a indagini sistematiche e esaurienti. Alla prova esemplare che ne ha ora dato Mirko Tavoni illustrando l'impresa del *DanteSearch*, si aggiunge l'enorme contributo che sul versante lessicologico e lessicografico è pervenuto dalle esperienze che fanno capo all'OVI – Opera del Vocabolario Italiano (CNR).

[1] Nel rispetto delle due Istituzioni che collaborano al *VD* (Accademia della Crusca e CNR-Opera del Vocabolario Italiano), la stesura del paragrafo 1 si deve a Paola Manni, quella del paragrafo 2 a Rossella Mosti.

Altrove ho richiamato l'attenzione sul vistoso scarto cronologico che separa le due voci corradicali *dantista* e *dantismo*[2]: se il *dantista*, ovvero lo studioso della *Commedia*, accanitamente dedito a coglierne i significati più sottili e reconditi, è figura nota fin dal Trecento, solo nel Novecento appare la voce *dantismo* nel senso di «parola coniata o introdotta da Dante» (*GRADIT*), in cui si coagula la coscienza dell'indice di creatività insito nel lessico dantesco e del suo lascito nell'italiano. Per capire i motivi di questa diffrazione, ricorriamo ancora una volta alle parole di Ernesto Giacomo Parodi che, alla fine dell'Ottocento, denunciava la mancanza, entro la pur sconfinata bibliografia dantesca, di «un lavoro complessivo sulla lingua di esso [del poema], che ci esponga con precisione scientifica in quanta parte Dante attinse al tesoro comune della lingua del tempo e in quanta parte fu innovatore» (Parodi 1896: 202). È una lacuna, questa, che per quanto concerne il lessico, si è protratta per buona parte del Novecento, e non è stata colmata neppure dall'*Enciclopedia Dantesca* che comunque ha rappresentato un notevolissimo avanzamento anche nello studio della lingua di Dante, sia per quanto essa offre negli articoli dedicati alle singole voci, sia per i grandi saggi di approfondimento linguistico che la corredano. È evidente tuttavia che per restituire con pienezza alla parola dantesca il suo spessore storico dovevamo attendere i grandi *corpora* dell'italiano antico approntati in seno all'OVI e la pubblicazione del *TLIO*, strumenti che ci consentono di ricostruire un quadro affidabile e compiuto dell'italiano dei primi secoli su cui proiettare le scelte dantesche. E si capisce anche quanto questa comparazione possa avvantaggiarsi se accanto al *TLIO* e in sinergia col *TLIO* – opera che costituzionalmente rappresenta l'italiano antico nella sua dimensione "plurale" (volutamente aperta a considerare la produzione delle origini senza preclusioni o privilegi di genere testuale, area di provenienza o livello stilistico) – si ha a disposizione uno strumento che, con adeguata metodologia, metta a fuoco le scelte "individuali" dantesche. La felice idea di riprodurre, in questo anno del Centenario dantesco, nella pagina d'ingresso dell'OVI, un medesimo lemma nella versione del *TLIO* e in quella del *VD* ne ha dato una prova semplice e diretta. E la seconda parte di questa relazione, tenuta da Rossella Mosti, membro del Comitato di Direzione

[2] Cfr. Manni 2018: 417-418.

del *VD* per l'OVI e Coordinatore della Redazione del *TLIO*, tratterà questo tema di fondamentale importanza.

D'altro lato, è risaputo che negli ultimi decenni, proprio dal tronco di quella lessicografia dell'uso, che a partire dalla fine dell'Ottocento aveva drasticamente reciso il secolare legame che la univa al canone degli autori, e anzi da quello che ne costituisce il prodotto più autorevole e avanzato – e qui penso naturalmente al *GRADIT* – si è fatto strada un nuovo, potente e per certi versi sorprendente richiamo a Dante, che con le sue parole non solo entra nella tessitura di quel vocabolario con oltre 2000 citazioni (2174 per la precisione), ma acquista un ruolo centrale nella riflessione teorica dedicata da Tullio De Mauro al processo di formazione e stabilizzazione del lessico italiano contemporaneo. Le conclusioni di De Mauro sono ampiamente note[3] e io non le ripeterò se non per ricordarne il nucleo essenziale, che sta nell'aver richiamato l'attenzione, col supporto dei dati statistici fino ad allora disponibili (dichiaratamente parziali e provvisori), sulla "funzione" che spetta a Dante nel processo di costituzione e assestamento del lessico italiano.

Intenti e metodi che abbiamo fin qui messo in evidenza agganciano saldamente il *VD* a un'impostazione lessicografica, differenziandolo dai diversi vocabolari danteschi che si sono fin qui editi, sempre ispirati a finalità esegetiche. E ancora rispondendo a un'esigenza da tempo affermatasi nell'ambito della lessicografia storica[4] prende vita un'altra prerogativa del *VD*, che si qualifica come nuova. Alludo al recupero della variantistica, nella sua parte lessicalmente significativa: una scelta che peraltro, nel caso di Dante e della *Commedia* in particolare, è resa indispensabile dal vivace dibattito in corso sulle più recenti edizioni del poema, che ci invitano a guardare al dettato dantesco in una prospettiva non rigidamente circoscritta entro un testo di riferimento prescelto (da noi comunque identificato nell'edizione Petrocchi 1994), ma aperta verso quanto ha successivamente prodotto e va producendo la ricerca filologica.

Il *VD*, nato dalla collaborazione dell'Accademia della Crusca con l'OVI – Opera del Vocabolario Italiano (CNR), è offerto alla consultazione libera e gratuita al sito di rete www.vocabolariodantesco.it. La

[3] Per un riepilogo degli interventi di De Mauro dedicati a questo tema, che acquista un crescente rilievo negli studi degli ultimi suoi anni, cfr. ancora Manni 2018: 420-421.

[4] Basterà ricordare Nencioni 1961 [1983] e, da ultimo, Coluccia 2020.

sua metodologia, l'allestimento della scheda-tipo, la realizzazione e le modalità della fruizione in rete hanno richiesto una lunga gestazione e sono il frutto di una strettissima collaborazione fra competenze linguistiche, filologiche e informatiche. Avvalendomi di alcune immagini, cercherò ora di delineare le linee portanti della sua struttura. Resta inteso che per una trattazione più ampia e soddisfacente, finalizzata alla consultazione, ci si rivolgerà alle pagine introduttive del *VD*, mentre per un approfondimento critico della teoria e della prassi che ispirano il progetto si rimanda al volume degli Atti della giornata di presentazione[5], avvenuta il 1° ottobre 2018, data che coincide anche con l'avvio della pubblicazione dei primi lemmi della *Commedia*.

Fig. 1 - *Pagina d'entrata del sito www.vocabolariodantesco.it*

Nella Fig. 1 vediamo la pagina d'entrata del *VD*, con la sua grafica sobria ed elegante ispirata ai disegni che Lorenzo di Pier Francesco de' Medici, cugino del Magnifico, commissionò a Sandro Botticelli. Assimilato a un periodico in aggiornamento continuo, il *VD* è diretto dalla sottoscritta (per l'Accademia della Crusca) e da Lino Leonardi (per l'OVI), coadiuvati da un Comitato di Direzione formato da: Giancarlo Breschi, Rosario Coluccia, Giovanna Frosini, Aldo Menichetti, Alessandro Pancheri e Mirko Tavoni (per l'Accademia della Crusca); e da Rossella Mosti, Zeno Verlato e Giuseppe Marrani (per l'OVI). La squadra dei redattori, che ha subito diversi assesta-

[5] Cfr. Manni 2020.

menti nel tempo, è attualmente formata da Barbara Fanini, Chiara Murru, Francesca De Cianni, Elena Felicani e Paolo Rondinelli. A Salvatore Arcidiacono, ricercatore dell'OVI, si deve l'allestimento informatico, innovativo rispetto a quello già in uso per il *TLIO* (e messo poi al servizio sia del *TLIO* stesso, sia di altre recenti imprese lessicografiche) che consta di una piattaforma unica capace di integrare *back-end* e *front-end* del vocabolario, ossia di gestire sia l'inserimento dei dati e la redazione delle singole voci, sia la pubblicazione e la consultazione di quest'ultime da parte degli utenti esterni[6].

Fig. 2 - *Es. di scheda lessicografica del VD*

Nella Fig. 2 vediamo come si presenta, in apertura, la scheda lessicografica, suddivisa in tre campi:

1. Il **lemma di entrata** con indicazione della relativa categoria grammaticale.
2. Un **pannello di approfondimento**, parte dinamica della scheda, che consente l'accesso a una serie di sezioni da cui si evincono ulteriori notizie sull'uso della voce nel contesto dantesco (*Frequenza*, *Index locorum*, *Locuz. e fras.*) e sulla sua vitalità nella tradizione linguistica italiana. A quest'ultimo scopo è deputata l'importantissima sezione delle *Corrispondenze* che, come si vede nella Fig. 3, è suddivisa in due parti: *a) Testi italiani antichi*, da cui si accede al *Corpus OVI*, al *DiVo*, al *LiriO*, e ai due

[6] Cfr. Arcidiacono 2020.

sottocorpora creati appositamente dal *Corpus OVI*: la *Prosa fior. sec. XIII* (testi in prosa fiorentini del secolo XIII) e *Petrarca e Boccaccio* (opere volgari dei due massimi autori trecenteschi); *b) Vocabolari*, da cui si accede al *TLIO*, alla *Crusca in rete* (che consente il monitoraggio della voce nelle cinque impressioni del Vocabolario della Crusca che vanno dal sec. XVII al XX) e alla sempre utile *Enciclopedia Dantesca*.

Fig. 3 - *Sezione "Corrispondenze" della scheda lessicale di febbre*

Dalla sezione *Corrispondenze* è anche possibile rimandare, se esiste, alla corrispettiva voce presente nelle opere latine di Dante, grazie al collegamento con il *Vocabolario Dantesco Latino* (VDL) alla cui realizzazione collaborano, insieme ad altre istituzioni, sia l'Accademia della Crusca che l'OVI (CNR)[7]. Completano il pannello la sezione dedicata alle *Varianti* (qui assente perché la voce *febbre* manca di varianti significative), una *Nota* di riepilogo e il nome del redattore che ha compilato la scheda. Infine il tasto *Tutto/stampa*, consente di visualizzare la scheda nella sua interezza e ottenerne una versione stampabile.

3. Abbiamo infine la **struttura semantica della voce**, nucleo fondante dell'articolo lessicografico. Le definizioni sono organizzate in griglie semantiche che mirano a registrare in modo rigoroso ed esaustivo tutte le accezioni attestate con i relativi esempi. Parte in-

[7] Per approfondimenti su quest'ultima iniziativa cfr. Albanese 2020.

tegrante dell'atto definitorio, le marche semantiche e d'uso intervengono sistematicamente a segnalare le molteplici componenti (filosofiche, scientifiche, settoriali, ecc.) che alimentano il lessico dantesco e il continuo germogliare di usi metaforici, estensivi, metonimici, ecc. Aspetti, questi, che si colgono in misura limitata nella voce qui riprodotta che pure, nella sua semplice struttura semantica, evidenzia un'accezione e una sottoaccezione di ambito medico, e un uso figurato. Da notare che le marche d'uso e semantiche sono codificate informaticamente; pertanto, attraverso una maschera di ricerca, si potranno effettuare delle interrogazioni mirate a richiamare le diverse categorie di termini.

L'architettura descritta e la centralità e la cura riservata alla struttura semantica delle voci basteranno a giustificare il titolo dell'opera che, pur dedicata al lessico di un autore, non si propone come *Glossario* ma come *Vocabolario*, denominazione che si sostanzia peraltro nell' attributo *dantesco*, che rimanda a un autore che come nessun altro ha messo alla prova e potenziato le risorse del proprio volgare nativo facendone una lingua capace di esprimere un'universalità di temi e di modularsi in una straordinaria varietà di registri espressivi. Se le oltre mille voci della *Commedia* fin qui pubblicate non consentono ancora dei bilanci sicuri, esse sono però già capaci di arricchire in modo significativo la nostra conoscenza del lessico dantesco non solo per quanto possono dirci sulle singole parole, ma anche per quanto testimoniano circa l'atteggiamento del poeta di fronte al lessico. Esse ci mostrano con sempre maggiore larghezza e oggettività di riscontri un Dante radicato nella lingua del suo tempo ma anche intensamente proteso a dilatarne i confini e a investire la parola di una potenza creativa – una «forza rifondante» potremmo dire con Riccardo Viel[8] – che continua a stupire e chiama il lessicografo ad un'ardua (ma esaltante) sfida[9].

[8] Cfr. Viel 2021.

[9] Per l'impasse lessicografico legato al trattamento di alcune voci polisemiche cfr. Fanini & Manni 2021.

2. *Il VD e i corpora dell'OVI*

Il *Vocabolario Dantesco* ha diverse peculiarità che si rivelano indubbiamente punti di forza rispetto ai precedenti repertori danteschi: una di queste è l'impostazione prettamente lessicografica della scheda, per cui ogni singolo lemma viene analizzato con rigore e sistematicità di metodo nei suoi diversi aspetti di interesse linguistico-lessicologico, che implicano *in primis* un'analisi della voce nel suo spessore storico, peraltro indispensabile per restituire al dettato dantesco la pienezza dei suoi valori semantici e stilistici, e colmare quella lacuna degli studi lucidamente colta alla fine dell'800 da Parodi già evidenziata da Paola Manni nella sua relazione. È evidente però che una simile contestualizzazione del lessico dantesco nella temperie linguistica della sua epoca può realizzarsi solo in questa fase storica della ricerca, avendo a disposizione innanzitutto il *Corpus OVI* (gattoweb.ovi.cnr.it), vale a dire la raccolta completa dei testi pubblicati dell'italiano antico, databili entro la fine del secolo XIV, che l'Istituto del CNR "Opera del Vocabolario Italiano" rende liberamente accessibile alla consultazione pubblica, affiancato dal *Corpus TLIO per il vocabolario* (tlioweb.ovi.cnr.it), la banca dati informatizzata e interrogabile per forme e per lemmi sulla quale si redige specificamente il Vocabolario, ma anche le altre banche dati create in seno all'OVI, strumenti imprescindibili per ricostruire un quadro completo e affidabile dell'italiano dei primi secoli. Il ricorso al *Corpus OVI* è utile per ricavare sì informazioni di tipo storico-linguistico, ma in questa sede vorrei rimarcare quelle di tipo metrico, destinate a impreziosire la sezione discorsiva della scheda lessicale (il "Campo Nota"), e questo grazie a raffinate funzioni di ricerca rese possibili dal *software* lessicografico GATTO[10] di proprietà dell'OVI, come quella del RIMARIO che per l'appunto permette di interrogare il lessico in rima.

[10] GATTO (Gestione degli Archivi Testuali del Tesoro delle Origini) è un *software* per l'indicizzazione e interrogazione degli archivi testuali ideato e sviluppato da Domenico Iorio Fili presso l'OVI e gestito ora da Andrea Boccellari. Sugli ultimi sviluppi della versione *on-line*, denominata GATTOweb, cfr. Boccellari 2019.

Fig. 4 - *Funzione "RIMARIO" nel Corpus OVI*

Grazie a tale opzione, che si attiva impostando una ricerca per forme e digitando la rima d'interesse, preceduta da un asterisco, verifichiamo che è proprio Dante a introdurre per primo la rima in *–olfo*, attestata nel poema solo nell'ottavo canto del *Paradiso* (nella serie *golfo* : *solfo* : *Ridolfo*), ma che troviamo successivamente in diversi autori del '300 quali Fazio degli Uberti (*Dittamondo*), Antonio Pucci (*Centiloquio* e *Guerra tra' Fiorentini e' Pisani*), Niccolò Cicerchia (*Risurrezione*) e Francesco di Vannozzo (*Rime*), che utilizzano i soliti rimanti danteschi, in particolare la coppia *solfo* : *golfo*, variandone talvolta solo l'ordine e la combinazione con l'antroponimo (Aristolfo, Astolfo, Landolfo, Pandolfo).

La possibilità offerta da GATTO di visualizzare gli esempi in modalità *kwic* (keyWord In Context) consente un immediato riscontro dei rimanti utilizzati, grazie alla parola chiave incolonnata in posizione centrale (mentre i relativi testi da cui sono estratti gli esempi sono rappresentati a sinistra da una stringa formata da caratteri alfanumerici seguìta dal riferimento organico e topografico)[11].

[11] Nel caso specifico: **fq** = *Commedia*; **hj** = *Dittamondo*; **qo** = *Centiloquio*; **qn** = *Guerra tra' Fiorentini e' Pisani*; **np** = *Risurrezione*; **ggu** = *Rime*.

Fig. 5 - *Visualizzazione degli esempi in modalità kwic relativi alla rima in –olfo*

Corpus TLIO per il vocabolario: risultati della ricerca - (intero corpus)

Contesti standard | Selezione | Annulla selezione | Salva | Vai a.. | Riavvia GattoWeb | Guide..

1 fq 3. 128.1	che caliga / tra Pachino e Peloro, sopra 'l	**golfo**	/ che riceve da Euro maggior briga, / non per Tifeo
2 fq 3. 128.3	Euro maggior briga, / non per Tifeo ma per nascente	**solfo**,	/ attesi avrebbe li suoi regi ancora, / nati per me
3 fq 3. 128.5	li suoi regi ancora, / nati per me di Carlo e di	**Ridolfo**,	/ se mala segnoria, che sempre accora / li popoli
4 hj 158.2	Puglia e tutto il Regno / per forza vinse e prese	**Pandolfo**,	/ che ne la Magna tenne poi per pegno. / Costui,
5 hj 158.4	poi per pegno. / Costui, veggendo tra' cherici il	**zolfo**	/ acceso per tre papi, ne fe' uno, / cacciando quei
6 hj 158.6	papi, ne fe' uno, / cacciando quei tre via per ogni	**golfo**.	/ Cinque con cinque e sette anni aduno / che questo
7 hj 455.11	la terra aperse, / né quanto dal ciel piovve foco e	**solfo**,	/ né tutte le città ch'al fondo amerse. / Ma se di
8 hj 455.13	ch'al fondo amerse. / Ma se di là andremo, vedrai il	**golfo**	/ dispettoso a mirar, che manifesta / se 'l miracolo
9 qo 1. 18.29	di Dio erronico, / nel settecencinquanta, e non fu	**ciolfo**	/ niente, perocch'ei morì Monaco. / E poi regnando
10 qo 1. 19.1	perocch'ei morì Monaco. / E poi regnando il Fratello	**Aristolfo**	/ di Santa Chiesa nemico mortale, / più che alla
11 qo 1. 19.3	mortale, / più che alla paglia non è il fuoco al	**zolfo**,	/ e' prese Roma, e lo spirituale / arse, e rubò; e
12 qo 1. 226.29	ma men d' un lupino / vi diede il Papa, e confermò	**Ridolfo**,	/ siccome Imperador verace, e fino. / Perchè
13 qo 1. 227.1	verace, e fino. / Perchè mostrava d'ogni virtù	**golfo**,	/ e promise venire, e poi non venne, / e trattò il
14 qo 1. 227.3	e poi non venne, / e trattò il Papa peggio, che	**micciolfo**;	/ che la promessa fatta non attenne; / ch' avie
15 qo 2. 76.11	tempo d' Imperio Campione / eletto nella Magna fue	**Astolfo**,	/ ma non pervenne alla 'ncoronazione, / perocchè
16 qo 2. 76.13	alla 'ncoronazione, / perocchè dal figliuol del Re	**Ridolfo**	/ in battaglia fu morto. Or mi diletta / di mutar
17 qo 2. 76.15	morto. Or mi diletta / di mutar cibo per istar più	**golfo**.	/ Nel dett' anno di Giugno, per vendetta / del Ponte
18 qo 3. 173.11	io t' ho contato, / tenea con gli altri, e non era	**micciolfo**,	/ ma d'ogni astuzia bene ammaestrato, / ed ebbe
19 qo 3. 173.13	astuzia bene ammaestrato, / ed ebbe nome Piero di	**Landolfo**,	/ il qual si tornò a Roma Cavaliere / del Popol
20 qo 3. 173.15	tornò a Roma Cavaliere / del Popol di Firenze molto	**golfo**.	/ Ma per gli modi, ch'e' tenne in primiere, / si
21 qn 212.10	'l Parmigiano / in quel di Pisa portò il fuoco, e 'l	**zolfo**,	/ poi seguitai del Cavalier sovrano, / di Camerin
22 qn 212.12	del Cavalier sovrano, / di Camerin Signor, Messer	**Ridolfo**,	/ ed or diren del terzo Capitano, / che fra'
23 qn 212.14	del terzo Capitano, / che fra' valenti non parve	**misciolfo**,	/ Messer Pier, ch'è Campione di San Piero, / e
24 np 403.23	tormento, / perché riman co' suo nell'aspro	**golfo**,	/ ov' è stridor di denti, fuoco e solfo. /
25 np 403.24	nell'aspro golfo, / ov' è stridor di denti, fuoco e	**solfo**.	/
26 ggu 5045.35	ogni barcuzo, / con magior puzo - che non getta el	**solfo**,	/ con tue galee de colfo / che te fanno gagliarda, /
27 ggu 5045.36	puzo - che non getta el solfo, / con tue galee de	**colfo**	/ che te fanno gagliarda, / facendo guarda - a Berta

Naturalmente per voci di bassa frequenza è sufficiente impostare una ricerca per lemmi nel *corpus* lemmatizzato, il *Corpus TLIO*, per ricavare già tutte le informazioni rilevanti di natura metrica come avviene per es. nel caso del verbo *abbicare*, parasintetico da *bica*, che ricorre per la prima volta a *Inf.* 9.78, col valore pronominale di 'aggrapparsi strettamente a qsa formando una sporgenza', e successivamente sempre in poesia e in posizione di rima eccezion fatta per i commentatori danteschi (Maramauro e Francesco da Buti).

Fig. 6 - *Risultato di una ricerca per lemmi di abbicare nel Corpus TLIO*

Relativamente alle poliremati che (locuzioni, espressioni fraseologiche e proverbiali) presenti nel poema, il *Corpus TLIO* si rivela uno strumento potente per individuare riscontri anteriori o evidenti richiami danteschi grazie alla ricerca "per cooccorrenze": valga su tutti l'esempio dell'espressione *Mettere in borsa*, attestata una volta nel poema (*Inf.* 19.72) ma presente già in un testo documentario della fine del '200 col significato di 'appropriarsi di denaro, intascare'.

Fig. 7 - *Risultato di una ricerca per cooccorrenze nel Corpus TLIO di "mettere in borsa"*


1 Doc. fior., 1277-96 389.13 TS mettere v.
dal fondacho, che nne rendei alla bottegha di Chalemala li diecie, e li tre **misi in borsa** per mie spese. E de dare lb. XJ e s. IIIJ a fiorini in

2 Dante, Commedia, a. 1321 Inf. 19 v.72 1, 321.6
orsa, / cupido sì per avanzar li orsatti, / che sù l'avere e qui me **misi in borsa**. / Di sotto al capo mio son li altri tratti / che precedetter me simoneggiando, / per

3 <Ottimo, Inf., a. 1334 (fior.)> c. 19 350.10 mettere(-si) v.
che fu sì avaro per avanzare a' suoi, che su, cioè in vita, si **misse in borsa** li danari, e misse sè qui, cioè in Inferno, che è etterna dannazione. Onde

4 <Ottimo, Purg., a. 1334 (fior.)> c. 19, proemio 336.5
la pecunia non si pone dove è la sete della avarizia, però che si **mette in borsa**, o in arca, e l' avarizia è nell' animo. La VIJ è, che lla

5 Giovanni Villani (ed. Porta), a. 1348 (fior.) L. 12, cap. 106 3, 216.13
quali in prima com'erano eletti, erano i loro nomi iscritti in polizze, e **messe in borse**, e per sesti. A' tempi, quando si traieno per detti ufici, si rimettieno in

6 Doc. fior., 1348-50 59.4
II di luglio, anno [M]CCCXLVIII, ebene contanti fior. cinquanta d'oro, i quali si **mise in borsa** per farne male spese per pasare tenpo. Lbr. LXXII s. X. E dè dare,

7 Doc. fior., 1348-50 87.40
due furono per comperare orzo per lo cavallo e fior. uno d'oro per **mettersi in borsa** per spese. A ffior. lbr. IIII s. VII. E deono dare, di primo d'

8 Doc. fior., 1348-50 97.13
desse a [...]. E fior. I d'oro e s. XXVII d. VI piccioli si **mise in borsa** per sue propie spese. E fior. I d'oro disse rendé a Francesco di

9 Maramauro, Exp. Inf., 1369-73 (napol.>pad.-ven.) cap. 19 317.9
nel mondo, *e qui*, *idest* in questo loco, se imborsa me, *idest* me se **mete in borsa**». Di sotto al capo mio son li altri tratti / che precedetter me simonegiando, / per

Ovviamente, essendo il *TLIO* ormai completo per il 73% del suo lemmario[12], gran parte delle informazioni si ricavano già dalle corrispondenti voci del Vocabolario (e in effetti nella voce *borsa* del *TLIO* troviamo registrata puntualmente l'espressione *Mettere in borsa* opportunamente definita e corredata della prima attestazione dei *Doc. fior.*, 1277-96), ma normalmente gli esempi citati nella voce non esauriscono tutta la documentazione presente nel *corpus* per cui una ricerca nella base di dati più generale dell'OVI (che conta oggi[13] 3261 testi appartenenti a tutte le varietà del sistema linguistico italiano) può rivelarsi fonte preziosa di informazioni aggiuntive e offrire un'occasione di lettura favorevole per interessanti spunti di riflessione. Nel caso specifico, il redattore della relativa scheda del *VD* (Cristiano Lorenzi Biondi) coglie opportunamente che l'espressione fraseologica assume in Dante un'accezione moralmente negativa di 'ottenere e accumulare guadagni in maniera illecita'.

Talvolta il ricorso al corpus testuale dell'OVI conferma la provenienza di un vocabolo alla base di una coniazione dantesca, come

[12] Ad oggi [marzo 2022, al momento di consegnare il contributo per gli Atti], il numero delle voci pubblicate del *TLIO* ha raggiunto quota 42.066 voci su un totale previsto di 57.800.

[13] Secondo i dati dell'ultimo aggiornamento del 10 gennaio 2022.

vediamo nel caso della scheda *ringavagnare* del *VD*, attestazione unica in *Inf.* 24.12 col valore figurato di 'recuperare (la speranza)'. Il verbo, che mostra una limitata vitalità (il *GDLI* lo attesta fino a Bernardo Davanzati), è un parasintetico di *gavagna* 'cesta', parola – secondo Nencioni[14] – conosciuta dal poeta a Lucca all'epoca dell'esilio, «penetrata in Lucchesia e fino nel pistoiese dalla Liguria e in genere dall'Italia settentrionale». E in effetti le due occorrenze rintracciate nel *Corpus OVI* e incluse nella voce *cavagno* del *TLIO* firmata dalla sottoscritta confermano la provenienza settentrionale del vocabolo di base (e precisamente il maschile *cavagno* è documentato nell'Anonimo Genovese mentre il femminile plurale *cavagne* è attestato nella *Parafrasi pavese del Neminem laedi* col valore figurato di 'custode'). Tali informazioni sono opportunamente citate nella relativa scheda dantesca, firmata da Fiammetta Papi.

A volte invece l'indagine condotta nella banca dati dell'OVI smentisce le affermazioni rese dalla precedente critica dantesca, come avviene per es. nel caso dell'aggettivo *probo*, latinismo da *probus*, attestato unicamente in *Par.* 22.138. La redattrice della scheda del *VD*, la stessa Fiammetta Papi, definisce così l'occorrenza aggettivale del poema: 'che possiede virtù e buoni costumi; che pensa e opera rettamente', legando quindi il vocabolo al complesso di virtù morali espresso già dal sostantivo femminile *probitate* di *Purg.* 7.122, e rigettando in pratica l'ipotesi di Parodi nella sua recensione a Isidoro Del Lungo (ipotesi appoggiata da Domenico Consoli che ha firmato la scheda relativa dell'*Enciclopedia Dantesca*), secondo cui *probo* è da ricollegare unicamente al cavalleresco *prode*: «il probo di Dante e il *probus* del latino medievale, nonostante le fallaci apparenze, non hanno quasi nulla a che fare col *probus* dei classici. Nel lat. mediev. dei paesi romanzi [[...]] il cavalleresco prode, discendente legittimo di prode prodis, per la solita tendenza ad etimologizzare [[...]] fu reso con probus, che gli somigliava di suono e ne conteneva l'idea che pareva fondamentale. Dante poi ritradusse in volgare quella singolar traduzione»[15]. La definizione resa da Papi, e avallata dal comitato di Direzione del *VD*, si appoggia principalmente sulla documentazione del *Corpus OVI*, in cui *probo* «è attestato spesso come sinon. di 'buono', 'virtuoso', 'retto',

[14] Nencioni 1989 [2000]: 6.
[15] Parodi 1898: 18.

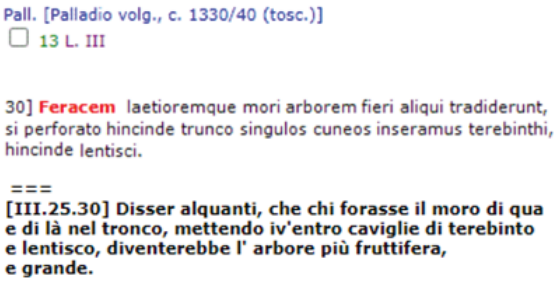
senza che la semantica dell'agg. debba ridursi al solo valore militare (pur compreso tra le possibili accezioni)».

Leggendo la sezione Nota delle 900[16] voci pubblicate del *VD* osserviamo tra l'altro che tra le risorse informatiche dell'OVI risulta particolarmente ricco di spunti per il *VD*, nel caso specifico dei latinismi danteschi, il *Corpus CLaVo* (clavoweb.ovi.cnr.it), che raccoglie le opere latine tradotte dai volgarizzamenti inclusi nel *Corpus DiVo* (divoweb.ovi.cnr.it), strumento quest'ultimo originariamente concepito per la redazione delle voci del *TLIO* ma ampiamente sfruttato ai fini del *VD*, e per questo già incluso insieme al *Corpus OVI* e al *Corpus LirIO* (lirioweb.ovi.cnr.it), altro repertorio che integra il *corpus* principale di riferimento per quanto riguarda i testi lirici, nella sezione "Corrispondenze" della scheda lessicale del *VD*.

Se il *Corpus OVI* è lo strumento più accreditato per valutare la reale portata dei latinismi danteschi dal momento che ci permette di verificare con certezza se la prima attestazione di un latinismo risalga proprio al divino poeta, grazie al Corpus *CLaVo* i redattori del *VD* accertano la rarità dell'uso dantesco di un vocabolo anche attraverso il testo latino di partenza dei volgarizzamenti medievali italiani di opere classiche e tardoantiche. A tale archivio attinge infatti Cristiano Lorenzi Biondi, nel caso dell'aggettivo *ferace* 'che produce benefici effetti', latinismo mai attestato in volgare prima di Dante, e presente solo in due volgarizzamenti, quello del Palladio e di Piero de' Crescenzi probabilmente per contatto diretto col latino, ricavando grazie proprio alla presenza del testo volgare associato, che nei testi toscani del '300 esso viene tradotto per lo più come *abbondevole* o *abbondante*, *fruttifero* e *fruttevole*.

[16] Alla data del convegno (settembre 2021). Oggi (marzo 2022) le schede pubblicate del *VD* ammontano a 1052.

Fig. 8 - *Es. di testo volgare associato nel Corpus CLaVo*



Se la consultazione del corpus *CLaVo* è di indubbia utilità per il *VD*, essa può risultare proficua per gli studi della lessicografia latina: mi riferisco naturalmente al *Vocabolario Dantesco Latino* (www.vocabolariodantescolatino.it), progetto dedicato al latino di Dante, diretto da Gabriella Albanese, Paolo Chiesa e Mirko Tavoni. In effetti, la stessa Albanese, in un contributo pubblicato per gli Atti della giornata di presentazione del Vocabolario Dantesco volgare, ne evidenzia il proficuo uso complementare, per la lessicografia latina[17].

Da quanto detto, emerge chiaramente che il *Corpus OVI* e le altre risorse informatiche dell'Istituto, e in particolare la funzione di ricerca che permette la creazione e la modifica di *sottocorpora* adattati in base alle esigenze dell'utente, si rivela quanto mai efficace ai fini del progetto del *VD*, e a tal proposito sono lieta che una proposta operativa avanzata dalla sottoscritta nell'ottobre 2018 in sede di convegno, in occasione della giornata di presentazione del Vocabolario Dantesco volgare[18], si sia felicemente concretizzata all'interno dell'OVI con l'allestimento di un *corpus* relativo ai commenti più antichi della Commedia, denominato CoDa "Commenti Danteschi" (consultabile attualmente all'indirizzo codaweb.ovi.cnr.it), ricavato agevolmente a partire da una ricerca per *sottocorpora* condotta nel campo "Genere", selezionando la sigla Comm[enti] (che per quanto riguarda

[17] Cfr. Albanese 2020: 183: L'archivio digitale del CLaVo «offre specifica possibilità d'indagine sul lessico dei volgarizzamenti dei classici latini, evidenziando la genesi e l'eventuale esistenza dei latinismi nel patrimonio lessicale coevo a Dante, con specifica pertinenza all'area dei traduttori dei classici latini».

[18] Su cui cfr. Mosti & Verlato 2020: 101.

Dante abbraccia anche le chiose poetiche e le rubriche) e restringendo ulteriormente l'indagine a quelli esclusivamente danteschi.

Fig. 9 - *Corpus CoDa*

Tale strumento critico, che per ovvie ragioni al momento è limitato ai soli commenti volgari del '300, è destinato ad ampliarsi notevolmente nell'immediato futuro: è prevista infatti la realizzazione di un progetto in collaborazione con l'Università degli Studi di Napoli Federico II per la costituzione di un *corpus*, realizzato sempre col software GATTO, esteso fino a Cristoforo Landino e allargato ai commenti latini.

Di tale percorso di ricerca si avvarrebbero non solo i redattori del gruppo del *VD* che comunque consultano già obbligatoriamente come fonte il *Dartmouth Dante Project*, banca dati *on-line* che rende disponibile gratuitamente la consultazione di più di 70 commenti antichi e recenti della *Divina Commedia*, ma ne trarrebbe giovamento anche il dantista che sfrutterebbe così tale materiale esegetico per una immediata interlocuzione critica. La chiamata in causa dei commentatori antichi, e nello specifico l'analisi lessicale delle chiose dantesche, fornisce sempre allo storico della lingua, al lessicografo, eccezion fatta per qualche caso di genericità di significato e di confusione, materiale prezioso per testare la vitalità del lessico dell'epoca, lessico che offre talvolta un certo grado di originalità e inventiva.

Insomma, l'importanza delle risorse informatiche dell'OVI per il *VD* e in generale per gli studi dell'italiano antico è evidente, e per

ribadirlo, a conclusione della mia breve relazione, mi piace citare le parole di Luca Serianni che in un contributo relativo ai latinismi nei testi dei primi secoli, pubblicato negli Atti del convegno conclusivo del progetto DiVo, scrive: «Lavorare avendo a disposizione archivi come il *Corpus OVI*, il *DiVo* e il *CLaVo* significa sottrarre l'indagine al rischio dell'impressionismo stilistico fondato sul singolo testo o sul singolo autore o traduttore; mettere a fuoco il contesto latino di partenza [[...]]; valorizzare parallelamente il contesto diacronico successivo»[19]. Di ciò ne sono ben consapevoli i redattori del *VD* che lavorano a un progetto che è un fulgido esempio di lessicografia informatizzata (visto che non solo lo si consulta in rete ma lo si redige in rete grazie a una piattaforma lessicografica ad esso dedicata sviluppata all'OVI da Salvatore Arcidiacono: *PlutoVD*)[20], che ha accolto fin dalla sua idea iniziale una collezione di *corpora* dell'OVI per la contestualizzazione storica del lemma dantesco, e che prevede tutta una serie di procedure apposite in cui gli stessi *corpora* dialogano attivamente con le voci.

Riferimenti bibliografici

Albanese, Gabriella. 2020. Per il Vocabolario latino di Dante. In Manni, Paola. 2020 (a cura di), 169-185.

Arcidiacono, Salvatore. 2020. «Forse tu non pensavi ch'io löico fossi!»: metodi computazionali al servizio del *VD*. In Manni, Paola. 2020 (a cura di), 81-92.

Boccellari, Andrea. 2019. Corpus OVI: GattoWeb. In Leonardi, Lino & Squillacioti, Paolo (a cura di), *Italiano antico, italiano plurale. Atti del convegno internazionale in occasione delle 40.000 voci del TLIO*, Supplementi al Bollettino dell'Opera del Vocabolario Italiano, Firenze 13-14 settembre 2018, Villa Reale di Castello, 71-81. Alessandria: Edizioni dell'Orso.

Coluccia, Rosario. 2020. Cosa le varianti della Divina Commedia possono insegnare alla storia della lingua e alla lessicografia italiana. In Manni, Paola. 2020 (a cura di), 141-156.

Dartmouth Dante Project. Banca dati dei commenti danteschi realizzata dal Dartmouth College in collaborazione con la Princeton University, consultabile in rete all'indirizzo http://dante.dartmouth.edu/.

[19] Serianni 2017: 141.

[20] Su tale dispositivo lessicografico realizzato presso l'OVI al servizio del *VD* cfr. Arcidiacono 2020.

Enciclopedia Dantesca. Diretta da Umberto Bosco. Roma: Istituto della Enciclopedia Italiana, 1984[2], consultabile in rete all'indirizzo http://www.treccani.it/enciclopedia/elencoopere/Enciclopedia_Dantesca.

Fanini, Barbara & Manni, Paola. 2021. Il "Vocabolario Dantesco" (relazione presentata al Convegno «A guisa d'uom che 'n dubbio si raccerta». *Vecchie questioni e nuove prospettive per la biografia e l'opera dantesca*, Verona, 5-7 luglio 2021), in corso di stampa.

GDLI = *Grande Dizionario della Lingua Italiana*, fondato da Salvatore Battaglia, poi diretto da Giorgio Bàrberi Squarotti, Torino, utet, 1961-2002, 21 voll.

GRADIT = *Grande Dizionario Italiano dell'uso*, ideato e diretto da Tullio De Mauro. Torino: UTET, 6 voll.

Manni, Paola. 2018. Da Dante a noi. Parole dantesche nel lessico italiano. In D'Onghia, Luca & Tomasin, Lorenzo (a cura di), *Etimologia e storia delle parole*. *Atti del XII Convegno ASLI – Associazione per la Storia della Lingua Italiana*, Firenze 3-5 novembre 2016, 417-432. Firenze: Franco Cesati Editore.

Manni, Paola. 2020 (a cura di). «*S'i' ho ben la parola tua intesa*». *Atti della giornata di presentazione del Vocabolario Dantesco*, Firenze, Villa Medicea di Castello, 1° ottobre 2018, Quaderni degli «Studi di Lessicografia Italiana». Firenze: Accademia della Crusca.

Mosti, Rossella & Verlato, Zeno 2020. Le Corrispondenze del VD. TLIO, lessicografia storica, corpora dell'OVI. In Manni, Paola. 2020 (a cura di), 93-121.

Nencioni, Giovanni. 1961 [1983]. Filologia e lessicografia a proposito della «variante». In Nancioni, Giovanni (a cura di), *Di scritto e di parlato. Discorsi linguistici*, 57-66. Bologna: Zanichelli 1983.

Nencioni, Giovanni. 1989 [2000]. Il contributo dell'esilio alla lingua di Dante. In Nencioni, Giovanni (a cura di), *Saggi e memorie*, 3-121. Pisa: Scuola Normale Superiore.

Parodi, Ernesto Giacomo. 1896. La rima e i vocaboli in rima nella «Divina Commedia». In Folena, Gianfranco (a cura di), *Lingua e letteratura. Studi di Teoria linguistica e di Storia dell'italiano antico*, con un saggio introduttivo di Alfredo Schiaffini, 2 voll., II, 203-284. Venezia: Neri Pozza.

Parodi, Ernesto Giacomo. 1898. Recensione a Isidoro Del Lungo, Dal secolo e dal poema di Dante. *Bullettino della Società Dantesca Italiana* n.s. 6. 1-19. Bologna: Zanichelli.

Petrocchi, Giorgio. 1994. *La Commedia secondo l'antica vulgata*. Firenze: Le Lettere, 4 voll (2° ed.).

Serianni, Luca. 2017. Per una tipologia dei latinismi nei testi dei primi secoli. In Guadagnini, Elisa & Vaccaro, Giulio (a cura di), *Rem tene, verba sequentur*. *Latinità e medioevo romanzo: testi e lingue in contatto. Atti del convegno conclusivo del progetto DiVo – Dizionario dei Volgarizzamenti*, Firenze, 17-18 febbraio 2016, 125-141. Bollettino dell'Opera del Vocabolario Italiano (Supplementi), 6, 125–156. Alessandria: Edizioni dell'Orso.

Viel, Riccardo. 2021. La parola creatrice: mitopoiesi della terzina dantesca (relazione presentata al Convegno «Dietro al mio legno che cantando varca», ri-scritture dantesche, Bari, 13-14 maggio 2021), in corso di stampa.

DOI: 10.17469/O2106SLI000019

GIULIO VACCARO

Rappresentatività e bilanciamento in un corpus di italiano antico: appunti sul *Corpus TLIO*

Il contributo propone alcune riflessioni sull'uso del *Corpus TLIO* come strumento per l'analisi semantica e per la ricostruzione lessicale dell'italiano (antico). Gli incrementi realizzati a partire dal 2016 ne hanno modificato infatti profondamente la struttura e hanno sensibilmente mutato, conseguentemente, i risultati che la ricerca consente di ottenere. Si propongono qui alcune considerazioni sulla composizione di un corpus ai fini della ricostruzione complessiva del lessico dell'italiano antico, per cui sarebbe necessario un triplice bilanciamento. Il principale è quello dall'appartenenza a singoli generi (anche molto ampi) o a tradizioni discorsive insufficientemente rappresentate (si pensi ai manuali di medicina, ai ricettari di cucina, ai libri di viaggio...), incrociata con i dati geografici e con quelli cronologici.

Parole chiave: italiano antico, corpus TLIO, tradizioni discorsive, bilanciamento, rappresentatività.

1. *Il Corpus TLIO: nascita e sviluppo.*

Come è abbondantemente noto, gli studiosi della storia linguistica italiana godono di un privilegio pressoché unico tra i colleghi romanisti e, generalmente, studiosi di lingue moderne: quello di avere (e di avere avuto fin dalla fine degli anni Novanta) a disposizione in rete un corpus di enormi dimensioni per descrivere la fase antica della lingua. Si tratta ovviamente della galassia dei corpus realizzati dall'Opera del Vocabolario Italiano del CNR (OVI), un sistema che ha conosciuto, nel corso del quarto di secolo di servizio, varie fasi. Da ultimo è stata creata una separazione tra il *Corpus OVI dell'italiano antico*, destinato a contenere tutti i testi antichi editi in modo affidabile, e il *Corpus TLIO*, su cui si fonda il vocabolario. I due corpus sono stati fondati, fino al 2016, sulle stesse basi teoriche ma con una diversa opera di marcatura dei lemmi; a partire da questa data l'aggiornamento del *Corpus TLIO* è stato vincolato a alcuni criteri, esplicitati nel sito

dell'OVI (<www.ovi.cnr.it>): principalmente la datazione dei testi entro la fine del Duecento e l'«appartenenza ad aree linguistiche scarsamente documentate», ma anche l'«eccezionale rilevanza lessicale e/o culturale» e le scritture femminili[1].

Le radici del *Corpus TLIO* (che è, ovviamente, il nocciolo del *Corpus OVI dell'italiano antico*) vanno ricercate negli anni Sessanta, dunque agli albori della linguistica computazionale e della linguistica dei corpus: ciò ha fatto sì che proprio all'interno dell'OVI si siano affrontate per la prima volta – in un'epoca in cui la riflessione teorica si trovava ancora a dover lottare con i limiti di operatività imposti dal *calcolatore* – questioni metodologiche rilevantissime, dettate soprattutto dalla considerazione che il corpus non fosse uno strumento in sé conchiuso ma che esso dovesse servire come base per la realizzazione della prima sezione cronologica del *Vocabolario Storico Italiano*[2]. Insomma, il corpus non poteva "limitarsi" a essere grande (soprattutto per la nebulosità dei confini) ma – come sosteneva nell'autunno del 1974 l'allora direttore dell'OVI, d'Arco Silvio Avalle – doveva avere due caratteristiche principali e coesistenti per poter procedere alla redazione del vocabolario: la prima era la stabilità dell'insieme; la seconda la sua rappresentatività[3]. Le ragioni portate da Avalle erano molto semplici: solo un insieme stabile avrebbe infatti potuto garantire un vocabolario che descrivesse coerentemente uno stato di lingua; solo un insieme rappresentativo delle diverse realizzazioni testuali avrebbe fatto sì che il vocabolario non fosse afflitto da quel "morbo della letterarietà" che era stato il portato tradizionale della lessicografia fiorentina.

Con il primo argomento, tra l'altro, Avalle individuava un problema che è oggi infinitamente più complesso vista l'estrema facilità (tecnica) per ciò che riguarda la possibilità di aggiornamento degli strumenti *online*. Se infatti è vero che far crescere un corpus è relativamente semplice – basti pensare che il primo corpus si cui si fondava il *TLIO* era di circa 15 milioni di occorrenze e quello attuale di oltre 23 milioni –, l'altra faccia della medaglia è che, relativamente alla redazione di un vocabolario di uno stato di lingua, un corpus incre-

[1] Per un panorama generale sugli strumenti si vedano gli interventi raccolti in Leonardi & Squillacioti 2019.

[2] Per la storia dell'Opera del Vocabolario Italiano dalla fondazione fino al 1992 (anno di inizio della direzione di Pietro Beltrami), cfr. Vaccaro 2013.

[3] Le considerazioni di Avalle si leggono in Vaccaro 2013: 355-363.

mentato rispetto agli inizi di circa il 53% (o, reciprocamente, in cui la parte su cui sono state redatte le prime voci è pari al 65% dell'attuale) darà inevitabilmente risposte diverse. La rilevanza di questa criticità non vale tanto per il più epidermico dei problemi, ossia l'aggiunta, sempre possibile con facilità in uno strumento *online*, di singole voci nel vocabolario. Se si guarda agli ultimi incrementi lessicali si noterà infatti che essi pertengono in misura quasi esclusiva all'aspetto glossaristico ineliminabile nel *TLIO*, che redige una voce per ciascun lemma attestato nel *Corpus TLIO*: di certo però voci come *ammanifestare*, *balievole* o *canesco* rappresentano prima ancora formazioni desultorie che lessemi periferici nella lingua.

L'aspetto invece più rilevante è che questo 53% aggiunto fatalmente documenterà in modo diverso la distribuzione del lessico innanzitutto sotto il profilo delle tradizioni discorsive (che è la cosa più rilevante sul piano lessicografico), e in subordine anche sull'asse della diacronia (e varrebbe qui la pena aprire una riflessione sul concetto di *ultima attestazione* proposto da D'Achille 2020) e – ancorché in misura minore e meno rilevante – della diatopia.

È insomma concreto il rischio, già additato da Guadagnini (2016), che nel passaggio dall'attestazione all'analisi semantica un corpus avente per obiettivo precipuo la descrizione semantica (cosa ovvia per il *Corpus TLIO*) che si arricchisca di testi scelti in base a criteri diversi da quello lessicale finirà per arricchirsi di *hapax* provenienti dal novero degli occasionalismi, delle autoschediastiche formazioni deverbali o denominali, di latinismi marginali; di una serie di parole, dunque, che non hanno avuto alcun peso nella storia dell'italiano (antico), che vanno ulteriormente ad ampliare quella quota già altissima di *hapax* o di lessemi monotestuali che costituisce già oggi quasi i due quinti del lessico descritto nel *TLIO*: per avere un dato numerico basti considerare che nelle 42.006 voci pubblicate nel *TLIO* gli *hapax* attestati nel corpus sono circa 12.000 e i lessemi documentati solo fuori corpus (per la gran parte anch'essi monoattestati) circa 5000. Ciò vuol dire che circa il 40% del *TLIO* è costituito da lessemi episodici nella storia dell'italiano antico.

Questa prima questione si intreccia strettamente con la seconda che poneva Avalle, quella della rappresentatività, che incrocia tuttavia un tema di portata maggiore. La domanda di base è infatti: è possibile realizzare un corpus rappresentativo di una varietà linguistica di cui

non conosciamo l'estensione e in cui il peso dei testi letterari è soverchiante? Conseguentemente, è possibile per un corpus sincronico rispetto a una fase antica di lingua o per un corpus diacronico caratterizzarsi non solo per la finitezza dell'insieme, ma anche per selettività e rappresentatività?

2. *Distorsione e bilanciamento: il problema della diatopia*

Ritornando al *Corpus TLIO*, Avalle riteneva che quanto raccolto rappresentasse ormai in modo adeguato le tradizioni testuali e la parte omessa fosse trascurabile o «inerte» (cfr. Vaccaro 2013: 361). Ciò implica che qualunque ampliamento, qualunque *rimeditazione* (questo il termine che usa a De Robertis 1985: 445) non può che partire da una riflessione preliminare sulla qualità di quel lavoro e sulle eventuali necessità e sull'opportunità di rimodularne il bilanciamento all'interno delle singole tradizioni discorsive.

Di fatto, la massiccia inclusione nel *Corpus TLIO* di testi aggiunti sulla base di una selezione in cui diacronia e diatopia fanno aggio su qualunque altra valutazione ha cessato dichiaratamente di fare del *Corpus TLIO* un corpus stabile ma non ne ha fatto e non ne fa in automatico un corpus bilanciato. In ultima analisi, se il vecchio *Corpus TLIO* poteva fondare una descrizione lessicografica che fosse un *vocabolario della lingua* e non solo il glossario del corpus stesso, il nuovo *Corpus TLIO* che prevede solamente aggiunte casuali qualitativamente può essere solamente un glossario, per quanto ampio, di un conglomerato, per quanto numeroso, di testi: la questione, ovviamente, non si pone sul dipolo della presenza/assenza del singolo testo, bensì su una valutazione qualitativa complessiva che investe l'idea stessa di descrizione lessicografica di una lingua. Preliminare rispetto alla "rimeditazione" di un corpus che voglia essere di base per una descrizione lessicale (sia essa dell'italiano antico o di qualunque varietà storica di una lingua) dovrebbe essere una riflessione di fondo: un qualunque corpus che si prefigga di andare al di là del mero dato di attestazione non può essere una semplice somma di testi (anche là dove – proiettando la questione su un piano meramente teorico – la somma dei testi inclusi coincida col totale dei testi editi, e quest'ultimo coincida a sua volta con il totale dei testi conservati) ma si deve fondare, comunque, su un principio di valutazione qualitativa delle

testimonianze, e dunque sulla selezione di quelle ritenute significative ai fini della determinazione dell'insieme, e del bilanciamento di queste testimonianze sull'asse delle variazioni interne (tradizioni discorsive, tipologie testuali) e esterne alla lingua (diacronia e diatopia). Il solo aggiungere testi al fine di incrementare numericamente un corpus, senza una seria riflessione preliminare che investa il piano complessivo del lessico dell'italiano antico e della sua rappresentazione, porta come unico risultato quello di mettere a disposizione un corpus più grande sotto il profilo quantitativo, ma non migliore qualitativamente. La valutazione dei testi si dovrà basare, dunque, non su piani contingenti (lingua del testo, tipologia dell'edizione, ecc.), bensì sul piano della storia della lingua e su un'analisi dei dati fondata su una "filologia dei grandi numeri". È quanto esplicitato con estrema chiarezza da Burgassi & Guadagnini (2017), che rappresenta il più ampio e intelligente uso di macrodati ricavabili dal *Corpus OVI*:

> "riteniamo che una corretta interpretazione dei dati restituiti dal *Corpus OVI* risulti dall'applicazione di una filologia 'dei grandi numeri': con questa denominazione indichiamo un tipo di analisi (ma prima ancora un punto di vista) che si attua su un piano supra-testuale e che consiste nell'osservare la testimonianza lessicale in prospettiva contrastiva, vale a dire interpretando la documentazione alla luce della caratterizzazione diatopica e diastratica, e delle diverse tipologie di documento e di tradizioni discorsive" (p. 11).

Poco rileva o almeno poco dovrebbe rilevare, quindi, quanto un'edizione sia commentata o quanto essa sia affidabile dal punto di vista fonologico (e in subordine morfologico) rispetto al manoscritto o ai manoscritti di base: una banca dati testuale, come qualunque opera umana, è costruita e progettata per una finalità specifica, e dunque la sua qualità e la sua efficacia vanno misurate rispetto a quella medesima finalità. Per quanto dirlo possa parere lapalissiano, il *Corpus OVI* è stato realizzato per essere la base su cui costruire il vocabolario dell'italiano antico. Dunque il principio interno che ne guida la costituzione e su cui se ne deve verificare la tenuta e la qualità è l'affidabilità del lessico testimoniato dalle edizioni. Ciò, ovviamente, non esclude che il corpus possa essere usato anche per altre ricerche su piani linguistici diversi dal lessico, e dunque sulla grafia, sulla fonologia, sulla morfologia e sulla sintassi; ma tutte queste ricerche richiedono un alto

grado di attenzione, di selezione e di (pre-)analisi dei dati da parte dell'utente.

Nel passaggio dal dato dell'attestazione a quello dell'analisi, un corpus destinato all'analisi semantica che si arricchisca di testi scelti in base a criteri diversi da quello lessicale (compreso un corpus esaustivo, e a prescindere dal fatto che sia esaustivo rispetto all'edito, al noto o al conservato) o sposerà una prospettiva glossaristica o, mantenendo la prospettiva del vocabolario, necessiterà di correttivi nella redazione delle voci (esclusione degli *hapax*, vaglio dell'attestazione, ecc.). Per contro, un corpus destinato alla redazione di un vocabolario in cui i testi siano aggiunti con criteri casuali dal punto di vista del lessico rappresentato finirà per fornire una descrizione casuale della lingua, in cui sono solo i grandi numeri a garantire la plausibilità dei risultati.

Al contrario, solo in un corpus bilanciato tra tutte le variabili è possibile individuare la posizione di un lessema (o – per essere più precisi – la posizione di ciascuno dei significati di un lessema) in una delle fasce lessicali. Poco (se non nulla) si può derivare dalla sola presenza di un alto numero di attestazioni, che è una condizione necessaria, ma sicuramente non sufficiente. Per usare le parole di Cosimo Burgassi e Elisa Guadagnini «una stima puramente numerica [...] non renderebbe giustizia dell'essenza del lessema, che acquista spessore dalla sinergia dei numeri e dalla valutazione dei contesti nei quali i numeri si collocano» (Burgassi & Guadagnini 2017: 69).

Il punto di riflessione necessario e preliminare all'inclusione (o no) di un testo in un corpus che abbia come obiettivo precipuo il lessico non può che essere, banalmente, quanto il testo contribuisca al dettaglio di determinate zone lessicali in rapporto alla documentazione già nota. Per fare un solo esempio, la versione italiana del *Lancelot en prose* (Cadioli 2016) rappresenta senza dubbio uno snodo fondamentale per comprendere lo sviluppo della diffusione e della ricezione della materia arturiana in Italia, ma lessicalmente testimonia esclusivamente un'ampia gamma di gallicismi che non sono tuttavia mai usciti dal manoscritto in cui erano contenuti. Al contrario testi più tardi e certamente meno interessanti sotto il profilo culturale, come per esempio il *Trattato dell'arte del vetro* di Benedetto di Baldassarre Obriachi (Milanesi 1864: 69-109), sono latori di documentazione altrimenti inattingibile di "lessico di bottega", quindi di lessico tecnico-specialistico, ma anche di tutto quel lessico materiale che è tipicamente

composto di quella parte della lingua fatta di parole che si usano ma, tendenzialmente, non si scrivono o si scrivono poco: in ultima analisi, quello che – in un corpus di lingua contemporanea – definiremmo lessico "ad alta disponibilità". Proprio la tendenziale scarsa attestazione di questi vocaboli, che sono perlopiù quelli maggiormente soggetti alla variazione geografica, rende, di fatto, inapplicabile un'analisi del lessico (e, conseguentemente, poco utile un incremento del corpus) sotto il profilo della diatopia.

Il dato della distribuzione linguistica dei testi, infatti, si scontra con due problemi: il primo, storico, è che in Toscana, nel Medioevo, si è scritto in volgare più che altrove e, soprattutto, molto più si è conservato, e – per ragioni storiche – molto di più si è edito; il secondo, contingente, è che la pubblicazione di singoli documenti locali è stata spesso delegata (altrove più che in Toscana) a eruditi del luogo in pubblicazioni a diffusione estremamente marginale e in edizioni spesso tutt'altro che impeccabili. Infine, l'immissione di testi sulla base di considerazioni di ordine diatopico può essere un criterio solo e esclusivamente nell'ottica di documentare il numero più elevato possibile di "punti" in cui un lessema è diffuso. È d'altronde un dato evidente che la massima diversificazione nell'analisi semantica del lessico avviene sulla base delle tipologie testuali, delle tradizioni discorsive e dell'uso linguistico (ovvero in diafasia e in diastratia), in misura minore in diacronia (ma tale variabilità tende a scemare in una descrizione in sincronia quale è quella del *TLIO*) e solo in minima parte in diatopia.

L'enorme squilibrio dei dati numerici tra l'area toscana (da cui proviene sì "solo" il 54% dei testi, ma ben l'80% delle occorrenze) e il resto del dominio italo-romanzo non consente infatti, partendo dall'incrocio dei dati sulla presenza/assenza di un lessema in una data area – se non in numeratissimi casi (perlopiù in cui il referente sia già geograficamente referenziato: per esempio i nomi di pesi e misure o di imposte) – di inferire nulla sulla distribuzione di quella determinata parola in italiano antico, e dunque sul tasso di "regionalità" di un singolo vocabolo. Il solo Leonardi (2019: 26) si è soffermato sulla questione, proprio partendo da questi dati:

> "Se andiamo a verificare nel vocabolario la presenza di voci fondate su documentazione esclusivamente non toscana, cioè fondata su quel 20% di testi [*rectius*: occorrenze] (...), il dato appare compatibile con quella dimensione, attestandosi attorno al 17%, a fronte

> dell'83% corrispondente alle voci per le quali sussiste una documentazione toscana (...)".
>
> "Ma la documentazione non toscana è presente in un numero significativo di queste voci, attestate *anche* in Toscana: queste voci a "Documentazione mista" (...) sono pari a un 28% del totale, per cui le voci ad attestazione solo toscana (...) nel *TLIO* riusltano pari al 55% del totale, una percentuale ben più ridotta rispetto al peso dei testi [*rectius*: occorrenze] toscani sull'insieme del corpus. È un dato che può sembrare irrilevante, nella sua astrattezza, ma è un dato da cui si può partire per un'analisi del lessico italiano delle origini che tenti di valutare l'effettiva consistenza e il peso relativo delle varietà regionali dell'italiano scritto".

Se la comparazione tra le percentuali dei *token* delle singole aree linguistiche e quella dell'attestazione dei singoli lessemi non finisce di convincermi (vista l'ovvia considerazione che la gran parte delle occorrenze in ciascuna area sarà riconducibile ai lessemi costituenti il lessico fondamentale e che tale lessico è privo di variazioni sostanziali in diatopia), i dati che emergono paiono additare piuttosto la difficoltà di un'analisi delle varietà regionali dell'italiano antico. La struttura stessa delle voci del *TLIO*, in cui il lemma è individuato sulla base dell'etimo indipendentemente dalle realizzazioni fonetiche, ci consente di dire solamente che (almeno) un 28% dei lessemi diffusi in più di un'area linguistica dell'Italia medievale ha un etimo comune: un dato che, in numeri assoluti, è pienamente sovrapponibile in una comparazione tra le diverse lingue romanze.

Che la presenza di un determinato lessema in un'area non sia segno di regionalità vale in modo massimo per la Toscana (ossia: un termine attestato solo in Toscana non è di necessità un toscanismo), ma vale anche in senso inverso, e vale anche nel caso in cui vi sia una comunanza di più aree contro la Toscana. Come hanno già rilevato Burgassi & Guadagnini (2017: 22), quando si verifica quest'ultima condizione «tale lessema è nella maggioranza dei casi un forte latinismo lessicale, e ciò descrive un fatto culturale più che linguistico».

Viceversa, il dato essenziale per un corpus di italiano antico è la sua capacità di documentare adeguatamente, e possibilmente in modo bilanciato, le diverse tipologie lessicali, in modo tale da rappresentare il lessico riducendo per quanto possibile il rischio di sovrarappresentare determinate aree a detrimento di altre. Si tratta di un problema in

realtà comune nella lessicografia italiana (storica e – per quanto possa sembrare paradossale – anche dell'uso), in cui si largheggia spesso nella documentazione di forme della lingua poetica minore e minima del Due e Trecento, ossia di uno dei settori più largamente indagati dal punto di vista filologico e letterario, mentre intere quote di lessico rimangono marginali. È il caso, per esempio, dei suffissati in *-anza* tipici della lirica duecentesca e primo trecentesca: su 359 voci contenute nel *TLIO* circa i due quinti (153) sono costituiti di attestazioni uniche di lessemi che sono, evidentemente, tipici di una tradizione discorsiva ma scarsamente rappresentativi della lingua e della sua evoluzione.

3. *Un triplice bilanciamento*

Pare dunque evidente che ai fini della ricostruzione complessiva del lessico dell'italiano antico sia necessario un corpus con un triplice bilanciamento. Il principale è quello dall'appartenenza a singoli generi (anche molto ampi) o a tradizioni discorsive insufficientemente rappresentate (si pensi ai manuali di medicina, ai ricettari di cucina, ai libri di viaggio...), incrociata con i dati geografici e con quelli cronologici.

Una rappresentazione dei testi provenienti da tutte le aree italiane, infatti, è sicuramente indispensabile, là dove si tenga presente però che non basta che un testo provenga da una determinata area per farne, *sic et simpliciter*, un testo lessicalmente affidabile per quella specifica area. La constatazione è banale e la facevano per esempio già Ferdinando Gabotto e Delfino Orsi pubblicando il primo (e purtroppo unico) volume dedicato ai laudari piemontesi: «rilevare l'influsso umbro in queste laude pare perfino ozioso» (Gabotto & Orsi 1891: XII). I laudari di Bra e di Carmagnola rappresentano infatti senza dubbio dei monumenti della letteratura piemontese del Quattrocento, ma ciò non implica (o almeno: non implica *necessariamente*) che essi siano anche dei monumenti del volgare piemontese del Quattrocento, poiché essi rientrano pienamente in un modello letterario che è a un tempo anche un modello linguistico. Ciò non vuol dire, naturalmente, riproporre il mito della genuinità del testo di carattere pratico contro l'artificiosità del testo letterario, visto che anche il volgare notarile è per sua natura esposto all'interferenza, in particolare del modello latino, e anche la scrittura dei mercanti riflette al suo interno le dinamiche del contatto linguistico con le zone in cui si commercia (si pensi a quanto accade

nei documenti conservati nell'archivio del mercante pratese Francesco di Marco Datini) o con le aree con cui si ha contatto.

Del pari è necessario rappresentare in modo adeguato anche la variazione (almeno: la possibile variazione) nella microdiacronia: privilegiare i testi duecenteschi – rimasti salvi nell'Ottocento dalla filologia che perseguiva il mito del secolo d'oro della lingua – vuol dire comunque sottorappresentare la prima grande età della scrittura in volgare, in cui – soprattutto per Firenze e la Toscana – si scriveva effettivamente di tutto in volgare.

L'incrocio di questi tre dati, tuttavia, non può prescindere da un quarto: quello dell'analisi critica preliminare del lessico di un testo rimane lo strumento fondamentale nelle mani dello studioso.

Proprio quest'aspetto è quello più significativo: fare un corpus di una fase storica di una lingua non è un qualcosa di meccanico, di banausico o di ancillare. L'auspicio è che gli aspetti testuali e lessicali vengano sviluppati in un nuovo *Corpus TLIO* che risponda in prima battuta a criteri di bilanciamento (che potrebbero utilmente essere quelli già previsti per *MIDIA*). Questo non tanto per contribuire alla realizzazione, ormai prossima, del *TLIO*, quanto per poter legare un corpus di lingua antica agli stati di lingua successivi, il cui pilastro finale potrebbe auspicabilmente essere il *VoDIM* e il cui primo stato potrebbe essere il *TLIQ* (*Tesoro della lingua italiana del Quattrocento*) suggerito in occasione del convegno per il trentennale dell'OVI da Tullio De Mauro (Leonardi & Maggiore 2016: 256).

Riferimenti bibliografici

Burgassi, Cosimo & Guadagnini, Elisa. 2017. *La tradizione delle parole: Sondaggi di lessicografia storica*, Strasbourg: ELiPhi.

Cadioli, Luca (a cura di). 2016. *Lancellotto. Versione italiana inedita del «Lancelot en prose»*. Firenze: Edizioni del Galluzzo per la Fondazione Ezio Franceschini.

Corpus TLIO per il vocabolario, diretto da Larson, Pär, Artale, Elena & Dotto, Diego, consultabile online all'indirizzo <http://tlioweb.ovi.cnr.it>.

Corpus OVI dell'italiano antico, diretto da Larson, Pär, Artale, Elena & Dotto, Diego, consultabile online all'indirizzo <http://gattoweb.ovi.cnr.it>.

D'Achille, Paolo. 2020. «A te l'estremo addio»? Il problema dell'ultima attestazione nella linguistica e nella lessicografia italiana. *Studi di lessicografia italiana* 37. 333–355.

De Robertis, Domenico. 1985. L'Ufficio filologico dell'Opera del Vocabolario, il suo impianto, il suo lavoro. In *La Crusca nella tradizione letteraria e linguistica italiana*. *Atti del Congresso Internazionale per il IV centenario dell'Accademia della Crusca (Firenze, 29 settembre-2 ottobre 1983)*, 444–451. Firenze: presso l'Accademia.

Gabotto, Ferdinando & Orsi, Delfino (a cura di). 1891. *Le laudi del Piemonte*. Bologna: Romagnoli.

Guadagnini, Elisa. 2016. Lessicografia, filologia e «corpora» digitali: qualche considerazione dalla parte dell'OVI. *Zeitschrift für romanische Philologie* 132(3). 755–792.

Leonardi, Lino. 2019. Filologia e lessicografia digitali: l'Opera del Vocabolario Italiano a quota 40.000. *Bollettino dell'Opera del Vocabolario italiano. Supplementi* 7. 15-31.

Leonardi, Lino & Maggiore, Marco (a cura di). 2016. *Attorno a Dante, Petrarca, Boccaccio: la lingua italiana. I primi trent'anni dell'Istituto CNR Opera del Vocabolario Italiano*. *Convegno internazionale (Firenze, 16-17 dicembre 2015)*. Alessandria: Edizioni dell'Orso.

Leonardi, Lino & Squillacioti, Paolo (a cura di). 2019. *Italiano antico, italiano plurale. Testi e lessico del Medioevo nel mondo digitale. Atti del convegno internazionale in occasione delle 40.000 voci del TLIO (Firenze, 13-14 settembre 2018)*. Alessandria: Edizioni dell'Orso.

MIDIA. Morfologia dell'Italiano in DIAcronia, coordinato da D'Achile Paolo, consultabile online all'indirizzo <https://www.corpusmidia.unito.it/>.

Milanesi, Gaetano (a cura di). 1864. *Dell'arte del vetro per musaico. Tre trattatelli dei secoli XIV e XV*. Bologna: Romagnoli.

Vaccaro, Giulio. 2013. Veniamo da molto lontano e andiamo molto lontano. Documenti per la storia dell'Opera del Vocabolario Italiano dalle origini al 1992. *Bollettino dell'Opera del Vocabolario Italiano* 18. 277–390.

VoDIM. Vocabolario Dinamico dell'Italiano Moderno, consultabile online all'indirizzo <https://vodim.accademiadellacrusca.org/>.

PARTE SECONDA

STUDI LINGUISTICI SU CORPORA

DOI: 10.17469/O2106SLI000020

ANGELA FERRARI, LETIZIA LALA, FILIPPO PECORARI

La punteggiatura italiana attraverso i corpora. Teoria, sincronia e diacronia

Tra il 2015 e il 2020 sono stati attivi all'Università di Basilea due progetti di ricerca dedicati allo studio della punteggiatura italiana in prospettiva sincronica e diacronica. Le ricerche si sono avvalse di una metodologia *corpus-based*, fondata sull'analisi di dati estratti da corpora elettronici e raccolte di testi non annotate. L'articolo riassume dapprima i risultati principali dei due progetti, mettendo in evidenza il ruolo fondamentale svolto dai corpora nell'elaborazione di una teoria della punteggiatura e nella descrizione dell'evoluzione del sistema interpuntivo nel tempo. In un secondo momento, è proposta una riflessione metodologica su alcuni limiti che i corpora elettronici pongono all'analisi della punteggiatura: limiti connessi principalmente al trattamento di segni specifici (lineetta, punto a capo) e all'impossibilità di accedere a testi interi per l'analisi di fenomeni che coinvolgono ampie porzioni testuali.

Parole chiave: punteggiatura, punteggiatura comunicativa, linguistica dei corpora, linguistica del testo.

1. *Introduzione*[1]

Tra il 2015 e il 2020, abbiamo sviluppato due progetti[2], finanziati dal Fondo Nazionale Svizzero per la Ricerca Scientifica, incentrati sulla punteggiatura italiana, che abbiamo studiato dapprima in prospettiva sincronica[3] e poi in prospettiva diacronica[4]. Per queste ricerche ci

[1] La stesura del contributo è dovuta a Letizia Lala per i §§ 1-2 e a Filippo Pecorari per il § 3. Angela Ferrari ha diretto entrambi i progetti menzionati.

[2] Ai progetti hanno partecipato: Angela Ferrari (direttrice), Letizia Lala, Fiammetta Longo, Filippo Pecorari, Benedetta Rosi, Roska Stojmenova.

[3] *PUNT-IT – Le funzioni informativo-testuali della punteggiatura nell'italiano contemporaneo, tra sintassi e prosodia* (100012_156119, febb. 2015 – genn. 2018).

[4] *PUNT-IT2 – La punteggiatura italiana in prospettiva diacronica: dallo standard al neo-standard, e dal Cinquecento al Novecento* (100012_175741, febb. 2018 – genn. 2020).

siamo avvalsi di una metodologia *corpus-based*, analizzando raccolte di grandi quantità di dati, consultabili e confrontabili. Per il nostro tipo di analisi, incentrata sugli usi interpuntivi, abbiamo potuto utilizzare corpora elettronici annotati, disponibili alla consultazione online (e.g. CORIS, DiaCORIS, CONTRAST-IT), e raccolte di testi non annotate (e.g. PUNT-IT, Web2Corpus_it, Narrativa anni 2000, Corpus La Stampa).

L'applicazione della *corpus analysis* ci ha consentito di individuare fenomeni specifici ed è stata fondamentale per arrivare alle generalizzazioni necessarie per elaborare una teoria e tratteggiare le evoluzioni del sistema su larga scala. Ha però mostrato anche alcune criticità che ci proponiamo di commentare (§ 3) una volta illustrati i risultati delle nostre ricerche (§ 2).

2. *Risultati delle ricerche*

2.1 La prima fase: lo studio teorico e sincronico

Nella prima fase della ricerca, dedicata all'osservazione della punteggiatura nell'italiano contemporaneo, abbiamo passato in rassegna l'intero sistema interpuntivo. Scostandoci dalla vulgata grammaticale e saggistica, abbiamo mostrato come le tradizionali interpretazioni della punteggiatura approntate in termini sintattici e/o prosodici siano fuorvianti dal punto di vista teorico e in ogni caso concretamente inapplicabili, e come l'uso contemporaneo dell'interpunzione in italiano non possa che definirsi in termini comunicativo-testuali[5]. Più precisamente, la punteggiatura svolge nella scrittura contemporanea due funzioni, che possono anche intersecarsi:

a. segmenta il testo nelle sue unità comunicative costitutive e (eventualmente) le gerarchizza (è il caso del punto, della virgola, del punto e virgola, dei due punti, della lineetta, delle parentesi);
b. introduce nel testo valori interattivi: inferenze (come per i puntini di sospensione), atteggiamenti illocutivi (come per il punto interrogativo e il punto esclamativo), piani polifonici (come per le virgolette).

[5] La definizione del valore comunicativo-testuale di fondo dei segni interpuntivi è stata proposta nel volume collettivo Ferrari *et al.* (2018).

Le nostre ricerche hanno mostrato che la concezione diffusa che la punteggiatura segnali snodi sintattici o che indichi le curve intonative e prosodiche di una realizzazione orale è inadeguata non solo in base a questioni teoriche (in particolare una concezione strutturale *vs*. funzionale della lingua), ma anche in quanto solo un'interpretazione comunicativo-testuale riesce davvero a rendere conto degli usi interpuntivi nell'italiano contemporaneo. In effetti, ogni tentativo di trattare la punteggiatura in chiave sintattica ha mostrato grossi limiti, finendo per scomporsi in distinzioni e sotto-distinzioni, e contraddicendosi con una massa di eccezioni ed eccezioni alle eccezioni (Ferrari & Lala 2011 e 2013). Anche gli usi che la tradizione più stabilmente riconduce alla sintassi mostrano in realtà regolarità di carattere testuale. Si pensi al punto, a cui tradizionalmente si attribuisce la funzione di chiudere una frase sintattica, ma che in realtà segnala la chiusura di un'unità testuale (l'Enunciato) autonoma da un punto di vista illocutivo-testuale, che non è affatto obbligatorio che abbia natura sintattica frasale, come mostrano impieghi come i seguenti[6]:

(1) Ho conosciuto tua moglie. **Affascinante.**
Oggi ho guidato io. **Fino a Firenze!**

Anche la virgola ha una funzione di natura testuale, avendo il ruolo di scandire l'Enunciato al suo interno in sotto-unità (Unità Informative) la cui natura è, ancora una volta, testuale e non sintattica:

(2) Si avvicinò**, lentamente,** e sorrise.
Mi è parso triste, **e arrabbiato**.

Abbiamo studiato e chiarito anche la relazione tra punteggiatura e prosodia di lettura. Ne è emerso che, se esiste in effetti una relazione tra presenza di un segno e intonazione di lettura, essa è però indiretta, sotto-specificata e parziale. Indiretta, perché mediata dai valori comunicativi ad esso associati; sotto-specificata, in quanto l'unità delimitata da un segno può ricevere profili intonativi diversi in base alla funzione informativo-illocutiva che è chiamata a svolgere; e parziale, in quanto è determinata non solo dalla presenza del segno, ma dalla combinazione con le indicazioni date dal lessico, dalla morfosintassi e dal cotesto.

[6] Per un approfondimento del modello teorico sul quale si basano queste analisi cfr. Ferrari *et al*. (2008).

Studiando gli impieghi sui corpora analizzati siamo risaliti al valore specifico di ogni segno di punteggiatura, arrivando a definizioni, che, allontanandosi dalla vulgata, sono tutte riconducibili all'una e/o all'altra delle due funzioni generali appena proposte (cfr. *supra*) e realmente in grado di rendere conto degli impieghi nei testi.

Si prenda ad esempio il punto interrogativo, al quale viene tradizionalmente riservato un trattamento sintattico (chiuderebbe nello scritto la frase interrogativa) e/o prosodico (attribuirebbe alla sequenza che chiude un'intonazione 'interrogativa'). In realtà lo studio sui corpora ha mostrato che: (i) in moltissimi casi il punto interrogativo chiude unità che non hanno affatto una natura frasale (*Perché mai? E quindi? Giornata pesante?*); (ii) la restituzione nell'orale di unità chiuse da questo segno non indirizza verso un unico profilo intonativo, ma verso un paradigma di realizzazioni anche piuttosto diverse tra loro: *Che cosa fai?* (domanda aperta), *Che cosa fai?* (richiesta di conferma), *Che cosa fai?!* (domanda accompagnata da tono enfatico, perentorio). Scartata dunque la lettura tradizionale, il nostro studio su corpora ci ha permesso di stabilire che il vero valore del punto interrogativo è di tipo (comunicativo) interattivo: esso introduce una richiesta di reazione, linguistica o non-linguistica, rivolta all'interlocutore. In contesto monologico, l'interazione sollecitata dal segno si produce tra scrittore e lettore; in contesto dialogico, tra i partecipanti allo scambio (Lala 2018).

2.2 La seconda fase: lo studio diacronico

L'aver compreso a fondo la punteggiatura italiana contemporanea ci ha portati a interrogarci su due questioni teoriche importanti di orientamento diacronico: (i) la punteggiatura italiana è sempre stata quello che è oggi? (ii) la punteggiatura italiana negli ultimi decenni è cambiata?

Adottando un'attenta metodologia *corpus-based* abbiamo intrapreso uno studio della punteggiatura in ottica diacronica che si è posto l'obiettivo di investire in queste due direzioni: così, da una parte abbiamo studiato la storia della punteggiatura dal Cinquecento a oggi (diacronia lunga), e dall'altra le sue evoluzioni più recenti, relative all'epoca del passaggio dell'italiano dallo standard al neostandard (diacronia breve).

2.2.1 Diacronia lunga

Per il primo ambito, di diacronia lunga, si è trattato di delineare e spiegare la storia della punteggiatura italiana, della sua concezione e dei suoi impieghi, dal Cinquecento fino al Novecento.

La ricerca è stata svolta basandosi sulle più importanti grammatiche e su un ampio corpus di scritture rappresentative. Abbiamo studiato il sistema interpuntivo in generale e ogni singolo segno di punteggiatura. I risultati ottenuti sono stati significativi: è emerso come l'italiano sia passato da un uso che combinava il criterio prosodico con quello morfosintattico (Cinquecento-primo Seicento), a un uso più rigorosamente morfosintattico (Seicento-secondo Settecento), infine a un uso comunicativo (stabilizzato nel secondo Ottocento, per raffinarsi sempre di più nel corso del Novecento).

Si è dunque potuto appurare come l'evoluzione nella storia della punteggiatura italiana equivalga in buona parte al passaggio da una *ratio* morfosintattica a una *ratio* comunicativo-testuale; passaggio registrato nel secondo Ottocento dalle grammatiche, anche su spinta delle scelte manzoniane.

Questo orientamento è particolarmente visibile se si osservano i mutamenti negli usi della virgola. Come si sa, la *ratio* morfosintattica prevede che la virgola compaia ogni volta che emerge un confine tra reggente e subordinata, qualunque sia la natura di quest'ultima; e, per la coordinazione, che la virgola accompagni sempre e comunque il collegamento copulativo. La *ratio* comunicativa chiede invece la virgola solo nei casi in cui la proposizione subordinata sia autonoma dal punto di vista informativo (il che la esclude nei casi in cui la subordinata è compattata semanticamente con la principale, come nelle completive post-reggente e nelle relative restrittive); mentre per la coordinazione, la regolarità comunicativa conduce a ometterla a meno che non emergano particolari necessità di ordine comunicativo, quali la disambiguazione o la focalizzazione. Così usi come quelli che vediamo negli esempi (3) e (4), improntati a regolarità di natura morfo-sintattica, perfettamente adeguati nel XVIII secolo[7]:

(3) Di questo n'ebbi una chiara pruova, **quando** mi fu concesso l'onore dalla Maestà Vostra d'ammirare la diligenza, **che**

[7] Gli esempi sono riprodotti fedelmente, senza normalizzare eventuali devianze o peculiarità grafiche.

usate nell'esaminare i minimi oggetti col Microscopio. (Della Torre 1755, in Ferrari 2020)

(4) A me resta di supplicare in fine l'A. V., **che** le piaccia nella presente offerta risguardare con occhio clemente l'estraordinaria volontà, **che** porto di corrispondere [...] all'obbligo che tengo alla Serenissima sua Casa, **ed** al contento che sento d'esserle nato, **e** di doverle morire devotissimo suddito, **e** servitore, **e** qui con ogni umiltà me le inchino. (Molza 1750, in Ferrari 2020)

hanno ceduto il passo a impieghi legati a una concezione comunicativa della punteggiatura, che è andata stabilizzandosi nella prima metà del XIX secolo, e che è all'origine dell'alternanza di presenza *vs.* assenza di virgola nelle copulative in (5):

(5) Sia dedicato a voi questo libro, dove io cercava, come si cerca spesso colla poesia, di consacrare il mio dolore, **e** col quale al presente (nè posso già dirlo senza lacrime) prendo comiato dalle lettere **e** dagli studi. Sperai / che questi cari studi avrebbero sostentata la mia vecchiezza, **e** credetti colla perdita di tutti gli altri piaceri, di tutti gli altri beni della fanciullezza **e** della gioventù, avere acquistato un bene che da nessuna forza, da nessuna sventura mi fosse tolto. Ma io non aveva appena vent'anni, quando da quella infermità di nervi **e** di viscere, che privandomi della mia vita, non mi dà speranza della morte, quel mio solo bene mi fu ridotto a meno che a mezzo; poi, due anni prima dei trenta, mi è stato tolto del tutto, **e** credo oramai per sempre. Ben sapete che queste medesime carte io non ho potute leggere, **e** per emendarle m'è convenuto servirmi degli occhi **e** della mano d'altri. (Leopardi 1831, in Ferrari 2020)

2.2.2 Diacronia breve

Per l'analisi in diacronia a breve gittata, l'obiettivo è stato quello di verificare se negli ultimi decenni, accanto alla punteggiatura standard, se ne stesse disegnando una neo-standard, nello stesso modo in cui hanno preso forma un neo-standard morfologico, sintattico e lessicale.

Sono emersi in effetti alcuni usi significativi che vanno in questa direzione: (i) il diffondersi di uno *style coupé* creato dal susseguirsi di punti che spezzano la linearità del dettato (6); (ii) la cosiddetta *virgola passepartout*, che sta espandendo il proprio ambito d'azione andando ad occupare spazi/funzioni tradizionalmente riservati a segni più forti (7); (iii) il diffondersi della lineetta singola di origine inglese (8); (iv)

il calo significativo d'impiego dei segni semanticamente ricchi, in particolare del punto esclamativo (9) e dei due punti (10); (v) il declino del punto e virgola, segno i cui impieghi rimangono frequenti solo in alcuni generi testuali conservativi (in particolare, nel linguaggio giuridico-amministrativo, con il ruolo di scandire enumerazioni) (11):

(6) Si era messa la gonna e la giacca grigia che usava quando faceva le cose importanti. Il golf girocollo. Le perle. E le scarpe blu con i tacchi alti. (Ammaniti, in Ferrari & Lala 2021: 22)

(7) Ha scelto la batteria. Dave Grohl, ex batterista dei Nirvana, approverebbe. («La Stampa», 6 luglio 2019, in Demartini 2019)

(8) L'ultima funzione, "Poke", permetteva infine di mandare a un utente solo un segnale di interesse, del tutto privo di contenuto – il destinatario riceveva un avviso che spiegava solamente che il mittente l'aveva "toccato, stuzzicato" («poked»). (Tavosanis, in Longo 2018a: 145)

(9) Franceschino. Che bello sentire la sua voce, ora. [*vs.* !] Che sollievo pensare che lui c'è davvero, non è una questione di crederci o non crederci. [*vs.* !] Che bello sentirlo partire sparato come al solito [...]. [*vs.* !] (Ferrante, in Lala 2019: 336)

(10) Il problema era mia madre, [*vs.* :] con lei le cose non andavano mai per il verso giusto. [...] Di sicuro non era felice, [*vs.* :] le fatiche di casa la logoravano e i soldi non bastavano mai. (Ferrante, in Lala 2019: 335) [*vs.* :]

(11) La Corte: riunisce i ricorsi; dichiara ammissibile ed accoglie il ricorso principale e rigetta il ricorso incidentale; dichiara la giurisdizione del giudice italiano; cassa la sentenza impugnata e rinvia la causa al Tribunale di Genova, che deciderà anche sulle spese del giudizio di legittimità. (Cass., sez. un. civ., 10-03-1998, n. 2642, in Dell'Anna 2017: 139)

La spinta verso il cambiamento e l'affermarsi di queste tendenze proviene in buona parte da scritture marcate in diafasia, come quelle letterarie degli ultimi cinquant'anni, o marcate in diamesia, come quelle che rientrano nella *Computer-Mediated Communication*, senza dimenticare l'influenza di impieghi tipici in altre lingue.

3. *Alcuni limiti delle ricerche interpuntive su corpora*

Come si è detto in §§ 1-2, tutte le ricerche qui menzionate si sono avvalse di una metodologia *corpus-based*, senza la quale non avremmo potuto elaborare una teoria della punteggiatura né tratteggiare le evoluzioni del sistema interpuntivo. L'uso dei corpora ha però fatto emergere anche alcuni limiti che questi strumenti pongono all'analisi della punteggiatura. Va precisato che i limiti osservati intaccano solo in maniera marginale l'utilità dei corpora come strumento per un'analisi della punteggiatura nei testi: come si vedrà, si tratta di limiti che riguardano una casistica limitata di segni e di fenomeni interpuntivi. Riteniamo tuttavia che sia opportuno stimolare una riflessione di carattere metodologico, anche perché a nostra conoscenza mancano studi specifici sull'uso dei corpora come strumento per indagini sulla punteggiatura: forse, dunque, può essere utile proporre qualche osservazione cercando di far interagire il punto di vista teorico sulla punteggiatura – privilegiato nelle nostre ricerche – con quello applicato alla costruzione di corpora.

Per l'analisi abbiamo ragionato retrospettivamente su un campione di corpora dell'italiano che ci sono stati utili in diverse fasi delle nostre ricerche: i corpora bolognesi CORIS (Rossini Favretti 2000) e DiaCORIS (Onelli *et al*. 2006), il PEC-Perugia Corpus (Spina 2014), il Primo Tesoro della Lingua Letteraria Italiana del Novecento (De Mauro 2007) e i *web corpora* ospitati da Sketch Engine (Jakubíček *et al*. 2013), con particolare riferimento a itTenTen.

3.1 La non-tokenizzazione dei segni di punteggiatura

Un primo limite che abbiamo riscontrato consiste nell'assenza della tokenizzazione dei segni di punteggiatura[8]. Si tratta evidentemente di un limite pressoché insormontabile per l'analisi interpuntiva: se il corpus non riconosce i segni di punteggiatura, e non può essere interrogato su di essi, la ricerca risulta impossibile.

Nel nostro campione di riferimento, l'unico corpus che presenta questo limite è il Primo Tesoro della Lingua Letteraria Italiana del Novecento, che raccoglie i testi di 100 romanzi vincitori o partecipanti al Premio Strega tra il 1947 e il 2006. A fronte della grande raf-

[8] Cfr. Lenci *et al*. 2005 (105-107) per una presa di posizione, dalla prospettiva computazionale, a favore della considerazione dei segni interpuntivi come token indipendenti.

finatezza che il corpus consente nelle analisi lessicali (ad es. sulle marche d'uso dei lessemi), esso non ammette ricerche sulla punteggiatura. Ciò risulta piuttosto sorprendente, perché tra le categorie oggetto di annotazione morfosintattica ci sono anche quelle di "ideogramma" e di "simbolo", che comprendono segni sicuramente meno interessanti per l'analisi linguistica rispetto ai segni di punteggiatura: segni come, ad esempio, <&>, <+>, <$> per la prima categoria e <[ ]>, <|> per la seconda categoria, che annovera peraltro anche due segni interpuntivi propri di lingue diverse dall'italiano, ovvero il punto esclamativo <¡> e interrogativo <¿> capovolti usati in spagnolo.

3.2 Il trattamento della lineetta

Al netto del problema appena menzionato, rilevante ma anche molto raro, gli altri limiti che abbiamo riscontrato riguardano principalmente il modo in cui sono considerati alcuni segni di punteggiatura nella tokenizzazione del corpus.

Un primo problema riguarda il trattamento della lineetta, un segno interpuntivo «alloglotto» (Longo 2020: 232) che l'italiano ha acquisito dall'inglese verso la fine del Settecento. Il problema principale che si incontra lavorando sui corpora è l'assimilazione tra i token relativi a due segni molto diversi dal punto di vista funzionale, ovvero la lineetta <–> e il trattino <->. La lineetta, secondo il punto di vista teorico, può essere considerata un segno interpuntivo a tutti gli effetti, perché contribuisce alla costruzione del messaggio testuale fornendo indicazioni sulla segmentazione del testo; il trattino, invece, è un segno paragrafematico non interpuntivo (Longo 2020: 234), che agisce a livello lessicale segnalando una relazione tra due lessemi o lavorando all'interno del lessema, come nelle parole composte. È evidentemente molto diversa la funzione che ha la lineetta in (12), una funzione simile a quella delle parentesi, rispetto alla funzione che ha il trattino in (13):

(12) A quelle parole – nette e assertive – i mercati hanno cambiato direzione. («Corriere della Sera», in Longo 2018b: 131)

(13) Salerno-Reggio Calabria, calcio-mercato, tecnico-scientifico, eco-incentivi (in Tonani 2011)

Il trattamento della lineetta e del trattino, nella maggior parte dei corpora del campione, non prevede una distinzione dei due segni su basi funzionali. Le opzioni che si riscontrano sono due:

i. Ci sono corpora, come il CORIS e il PEC, che assimilano integralmente i due simboli grafici: nel CORIS la ricerca della lineetta non dà risultati, e tutto ciò che si trova – tanto usi come in (12) quanto usi come in (13) – compare tokenizzato come trattino e richiede di essere ricercato come tale; nel PEC la ricerca del trattino e la ricerca della lineetta danno esattamente gli stessi risultati.
ii. Ci sono corpora, come il DiaCORIS e itTenTen, che distinguono sì i due simboli grafici, ma su basi esclusivamente formali e non funzionali. Questo trattamento può porre problemi alla ricerca, perché i testi non sono sempre uniformi nella realizzazione della lineetta e del trattino con due simboli distinti, e in particolare la lineetta è spesso realizzata graficamente nella forma breve tipica del trattino. Di fatto, chi volesse esaminare le occorrenze del segno interpuntivo "lineetta" sarebbe obbligato a discriminare gli esempi manualmente, in maniera non diversa da quanto impone il caso (i).

3.3 Il trattamento dell'a capo

Un altro limite relativo al trattamento di forme specifiche è quello che riguarda gli a capo. Nelle ricerche teoriche che abbiamo condotto sulla punteggiatura (cfr. Ferrari 2018), il punto a capo è classificato come un segno a sé stante rispetto al punto fermo, perché ha la funzione di delimitare un'unità testuale specifica, ovvero il Capoverso, che può contenere al suo interno più Enunciati delimitati da punti fermi o da altri segni. Più in generale, per un'analisi *corpus-based* della punteggiatura in prospettiva testuale sarebbe importante poter accedere alla scansione in capoversi del testo, non solo in relazione agli usi del punto a capo, ma anche a quelli di altri segni accompagnati dall'a capo: ad esempio, dei due punti che introducono un elenco, oppure delle lineette usate come segnale grafico dei punti di una lista.

Nei corpora abbiamo riscontrato tre opzioni relativamente al trattamento dell'a capo:

i. Ci sono corpora, come il CORIS e il DiaCORIS, in cui la scansione del testo in capoversi è assente. In questo caso, si perde la possibilità di distinguere il punto a capo dal punto fermo, o di riflettere sull'associazione tra l'a capo e altri segni interpuntivi.

ii. In altri corpora, come il PEC, si adotta una soluzione in qualche misura opposta a quella in (i): nelle finestre di contesto mostrate dall'interfaccia tutte le frasi sono isolate da un a capo, a prescindere dalla presenza effettiva di un a capo nel testo originario. Anche in questo caso, la scansione originaria in capoversi è evidentemente non ricostruibile.
iii. Un'opzione più utile all'analisi interpuntiva è quella presente in itTenTen, che marca i confini di frase e di capoverso con due tag XML specifici: <s> per la frase (*sentence*) e <p> per il capoverso (*paragraph*). Questa scelta di annotazione consente di discriminare i punti a capo dai punti fermi non solo *a posteriori*, alla lettura delle concordanze, ma anche al momento della ricerca automatica, attraverso l'impiego di stringhe apposite[9].

3.4 L'accessibilità ai testi del corpus

A conclusione di questa breve rassegna, proponiamo un'ultima osservazione che ci porta oltre il dominio della punteggiatura, a toccare più in generale i rapporti tra corpora e linguistica del testo. Alcuni tra i fenomeni interpuntivi considerati nelle nostre ricerche coinvolgono non singoli enunciati o brevi sequenze testuali, ma blocchi di testo più estesi, quando non addirittura testi interi. È per esempio il caso della virgola *passepartout* menzionata in § 2.2.2., che può essere usata in lunghe sequenze di testo. Si tratta di un espediente tipico del testo letterario, che serve principalmente a mimare una tirata di parlato senza interruzioni da parte di un personaggio, come nell'esempio seguente:

(14) Che brutta gente – attaccò a dire senza fermarsi più fino alla metropolitana di Piazza Amedeo –, hai visto la vecchia come t'ha trattata, s'è voluta vendicare, non può sopportare che Nadia, educata apposta per essere la meglio di tutte, Nadia che doveva darle tante soddisfazioni, non combina niente di buono, s'è messa col muratore e le fa la puttana sotto gli occhi: sì, non può sopportarlo, ma tu fai male a dispiacerti, fottitene, non glielo dovevi lasciare il tuo libro, non dovevi chiedere se voleva la dedica, soprattutto non gliela dovevi fare, questa è gente che bisogna trattare a calci in culo, il tuo difetto è che sei troppo buona, abbocchi a tutto ciò che dicono quelli che

[9] Per cercare i soli punti a capo, ed escludere i punti fermi, si può usare la stringa in *Corpus Query Language* [lemma=""] <p>.

> hanno studiato come se la testa ce l'avessero soltanto loro, e invece non è così, rilassati, va', sposati, fa' il viaggio di nozze, ti sei preoccupata troppo per me, scrivi un altro romanzo, lo sai che m'aspetto da te cose bellissime, ti voglio bene. (Ferrante, in Ferrari 2017: 148)

Un altro fenomeno che coinvolge ampie porzioni testuali è l'uso dei puntini di sospensione come segno esclusivo, che si sostituisce a tutti o quasi gli altri segni segmentanti. Questo uso è piuttosto comune in rete, nei testi scritti dai non professionisti della scrittura come il seguente:

(15) Mi rivolgo a tutte le Persone che non stanno facendo altro che insultare gigio e tutta la nostra famiglia.. Gigio sin da piccolo e tifoso del Milan.. per lui giocare con la maglia del Milan e' un sogno..ha sempre onorato e dato L anima per questi colori.. ha pianto per ogni sconfitta.. fino a ieri eravate tutti con gigio.. ora senza sapere nulla state insultando tutta la famiglia, scrivendo frasi che la nostra famiglia non augura nemmeno al peggior nemico.. la nostra famiglia ha gioito e pianto con tutti voi tifosi.. il Milan ha una storia incredibile.. e nessuno può metterlo in dubbio..Per le persone che hanno scritto messaggi a favore di gigio ci tengo a dire che voi avete capito davvero che persona e' gigio.. qualunque gesto che ha fatto e qualunque frase ha detto o scritto.. L ha fatto davvero per amore del Milan.. gigio e' soprattutto un tifoso del Milan.. Come voi..e chi lo insulta non è tifoso del Milan..ora potete anche riempire di insulti questa foto.. ma la famiglia lasciatela stare.. loro ci hanno sempre insegnato i veri valori della vita... per quelli che invece continuano a dire che io devo ringraziare a gigio perché mi da i soldi.. vi dico che a me mai nessuno mi ha regalato qualcosa.. ogni anno lotto per guadagnare quello che mi merito.. grazie.. (instagram.com/antodonnarumma90, 16.06.2017, in Pecorari 2019: 161-162)

Per l'analisi di fenomeni come quelli qui esemplificati, i corpora pongono al testualista il problema – ormai annoso – dell'accessibilità ai testi contenuti nel corpus. In linea di massima, la maggior parte dei corpora non consente l'accesso ai testi nella loro interezza, ma solo a porzioni limitate. Il problema, come è noto, è essenzialmente legale, e dipende dai vincoli posti dalle norme sul copyright (cfr. Allora & Barbera 2007): i dati del corpus possono essere distribuiti al pubblico

soltanto nei limiti di una finestra di contesto di poche decine o centinaia di caratteri.

La principale soluzione pratica che abbiamo adottato nelle nostre ricerche per ovviare a questo problema è stata quella di costruire raccolte di testi *ad hoc*, progettate per essere rappresentative di un tipo testuale o di una varietà linguistica, a partire da testi disponibili pubblicamente in rete: articoli pubblicati negli archivi online dei quotidiani, saggi e articoli scientifici ad accesso libero, testi normativi e amministrativi ecc. È per esempio questo il caso del corpus PUNT-IT, che è stato il nostro principale terreno di indagine della punteggiatura negli ultimi anni[10], e anche del corpus It-Ist_CH, che ci sta sostenendo da qualche tempo nello studio dell'italiano istituzionale svizzero in prospettiva testuale[11].

I corpora così costruiti hanno tendenzialmente dimensioni più ridotte dei corpora elettronici di ultima generazione (PUNT-IT, per esempio, conta circa 500.000 parole, mentre It-Ist_CH ne comprende 2.500.000), e peraltro, secondo alcune definizioni più restrittive di "corpus", non potrebbero nemmeno essere chiamati corpora in senso stretto, dal momento che i testi non sono tokenizzati e non sono addizionati di markup[12]. Tuttavia, la possibilità che queste risorse offrono di accedere agilmente e senza alcuna limitazione ai testi interi costituisce un vantaggio pratico di enorme rilevanza per l'analisi testuale, specialmente in casi particolari come gli usi interpuntivi esemplificati in (14) e (15).

[10] Il corpus PUNT-IT è stato costruito nell'ambito del progetto omonimo e contiene al suo interno testi giornalistici, saggistici e giuridico-amministrativi scritti negli ultimi trent'anni (1985-2015), rappresentativi della scrittura funzionale contemporanea di registro medio-alto.

[11] Il corpus It-Ist_CH è stato costruito nell'ambito del progetto FNS "L'italiano istituzionale svizzero: analisi, valutazioni, prospettive" attualmente in corso ed è disponibile ad accesso libero sul sito del progetto: cfr. https://sites.google.com/view/progettoitistch/corpus e la descrizione del corpus in Ferrari *et al*. (2022).

[12] Secondo Barbera *et al*. (2007), il corpus è una «[r]accolta di testi in formato elettronico uniformemente trattati (ossia *almeno tokenizzati ed addizionati di un markup adeguato*) in modo da essere gestibili ed interrogabili informaticamente» (p. 26) [corsivo nostro].

Riferimenti bibliografici

Allora, Adriano & Barbera, Manuel. 2007. Il problema legale dei corpora. Prime approssimazioni. In Barbera, Manuel & Corino, Elisa & Onesti, Cristina (a cura di), *Corpora e linguistica in rete*, 109–118. Perugia: Guerra.

Barbera, Manuel & Corino, Elisa & Onesti, Cristina. 2007. Cosa è un corpus? Per una definizione più rigorosa di corpus, token, markup. In Barbera, Manuel & Corino, Elisa & Onesti, Cristina (a cura di), *Corpora e linguistica in rete*, 25–88. Perugia: Guerra.

Dell'Anna, Maria Vittoria. 2017. Veniamo al punto. Interpunzione e dintorni nei testi giudiziari italiani. In Ferrari, Angela & Lala, Letizia & Pecorari, Filippo (a cura di), *L'interpunzione oggi (e ieri). L'italiano e altre lingue europee*, 131–146. Firenze: Cesati.

Demartini, Silvia. 2019. I punti della situazione. Viaggio nella punteggiatura dell'italiano di oggi. 3. La virgola splice. (https://www.treccani.it/magazine/lingua_italiana/articoli/scritto_e_parlato/punteggiatura3.html) (Consultato il 02.02.2022.)

De Mauro, Tullio. 2007. *Primo Tesoro della Lingua Letteraria Italiana del Novecento*. Torino: UTET/Fondazione Bellonci.

Ferrari, Angela. 2017. Usi "estesi" del punto e della virgola nella scrittura italiana contemporanea. *La lingua italiana. Storia, strutture, testi* XIII. 137–153.

Ferrari, Angela. 2018. Il punto a capo. In Ferrari, Angela & Lala, Letizia & Longo, Fiammetta & Pecorari, Filippo & Rosi, Benedetta & Stojmenova, Roska, *La punteggiatura italiana contemporanea. Un'analisi comunicativo-testuale*, 95–107. Roma: Carocci.

Ferrari, Angela. 2020. Note sull'uso della virgola ai margini della scrittura letteraria e saggistica tra Sette e Ottocento. *Margini. Giornale della dedica e altro* 14. (https://www.margini.unibas.ch/web/rivista/numero_14/saggi/articolo1/ferrari.html) (Consultato il 02.02.2022.)

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.

Ferrari, Angela & De Cesare, Anna-Maria & Evangelista, Daria & Lala, Letizia & Marengo, Terry & Pecorari, Filippo & Piantanida, Giovanni & Rosi, Benedetta. 2022. Il corpus It-Ist_CH: un corpus rappresentativo dell'italiano istituzionale svizzero. In Baranzini, Laura & Casoni, Matteo & Christopher, Sabine (a cura di), *Linguisti in contatto 3. Ricerche di*

linguistica italiana in Svizzera e sulla Svizzera, 57–69. Bellinzona: Osservatorio linguistico della Svizzera italiana.

Ferrari, Angela & Lala, Letizia. 2011. Les emplois de la virgule en italien contemporain. De la perspective phono-syntaxique à la perspective textuelle. In Favriaud, Michel (a cura di), *Ponctuation(s) et architecturation du discours à l'écrit*, 53–88. Paris: Larousse/Armand Colin.

Ferrari, Angela & Lala, Letizia. 2013. La virgola nell'italiano contemporaneo. Per un approccio testuale (più) radicale. *Studi di Grammatica Italiana* XXIX-XXX. 479–501.

Ferrari, Angela & Lala, Letizia. 2021. *Interpunzioni creative. Esempi letterari degli anni Duemila*. Firenze: Cesati.

Ferrari, Angela & Lala, Letizia & Longo, Fiammetta & Pecorari, Filippo & Rosi, Benedetta & Stojmenova, Roska. 2018. *La punteggiatura italiana contemporanea. Un'analisi comunicativo-testuale*. Roma: Carocci.

Jakubíček, Miloš & Kilgarriff, Adam & Kovář, Vojtěch & Rychlý, Pavel & Suchomel, Vít. 2013. The TenTen corpus family. In *Proceeding of the 7th International Corpus Linguistics Conference CL*, 125–127.

Lala, Letizia. 2018. Il punto interrogativo. In Ferrari, Angela & Lala, Letizia & Longo, Fiammetta & Pecorari, Filippo & Rosi, Benedetta & Stojmenova, Roska, *La punteggiatura italiana contemporanea. Un'analisi comunicativo-testuale*, 183–199. Roma: Carocci.

Lala, Letizia. 2019. Sulle tendenze interpuntive nella narrativa italiana contemporanea. In Moretti, Bruno & Kunz, Aline & Natale, Silvia & Krakenberger, Etna (a cura di), *Le tendenze dell'italiano contemporaneo rivisitate. Atti del LII Congresso Internazionale di Studi della Società di Linguistica Italiana (Berna, 6-8 settembre 2018)*, 323–341. Milano: Officinaventuno.

Lenci, Alessandro & Montemagni, Simonetta & Pirrelli, Vito. 2005. *Testo e computer. Elementi di linguistica computazionale*. Roma: Carocci.

Longo, Fiammetta. 2018a. La lineetta singola. In Ferrari, Angela & Lala, Letizia & Longo, Fiammetta & Pecorari, Filippo & Rosi, Benedetta & Stojmenova, Roska, *La punteggiatura italiana contemporanea. Un'analisi comunicativo-testuale*, 141–153. Roma: Carocci.

Longo, Fiammetta. 2018b. Le lineette doppie. In Ferrari, Angela & Lala, Letizia & Longo, Fiammetta & Pecorari, Filippo & Rosi, Benedetta & Stojmenova, Roska, *La punteggiatura italiana contemporanea. Un'analisi comunicativo-testuale*, 127–140. Roma: Carocci.

Longo, Fiammetta. 2020. La lineetta nelle grammatiche dell'Ottocento. In Ferrari, Angela & Lala, Letizia & Pecorari, Filippo & Stojmenova Weber,

Roska (a cura di), *Capitoli di storia della punteggiatura italiana*, 231–246. Alessandria: Edizioni dell'Orso.

Onelli, Corinna & Proietti, Domenico & Seidenari, Corrado & Tamburini, Fabio. 2006. The DiaCORIS project: a diachronic corpus of written Italian. In *Proceedings of the 5th International Conference on Language Resources and Evaluation – LREC 2006*, Genova, 1212–1215.

Pecorari, Filippo. 2019. Punteggiatura in rete: i puntini di sospensione nella comunicazione mediata dal computer. *Linguistica e filologia* 39. 129–175.

Rossini Favretti, Rema. 2000. Progettazione e costruzione di un corpus di italiano scritto: CORIS/CODIS. In Rossini Favretti, Rema (a cura di), *Linguistica e informatica. Multimedialità, corpora e percorsi di apprendimento*, 39–56. Roma: Bulzoni.

Spina, Stefania. 2014. Il Perugia Corpus: una risorsa di riferimento per l'italiano. Composizione, annotazione e valutazione. In Basili, Roberto & Lenci, Alessandro & Magnini, Bernardo (a cura di), *The First Italian Conference on Computational Linguistics. Proceedings*, 354–359. Pisa: Pisa University Press.

Tonani, Elisa. 2011. Trattino. In Simone, Raffaele (a cura di), *Enciclopedia dell'italiano Treccani*, 1520–1522. Roma: Istituto della Enciclopedia Italiana. (https://www.treccani.it/enciclopedia/trattino_%28Enciclopedia-dell%27Italiano%29/) (Consultato il 02.02.2022.)

DOI: 10.17469/O2106SLI000021

PHILIPPE MARTIN

Intonation of telephone conversations in a Customer Care service

RATP-DECODA is a Customer Care Service corpus part of the ORFEO project (2020). It includes 1988 recordings of client requests made between 2009 and 2011 to the RATP (*Régie Autonome des Transports Parisiens*) call center. While some semantic and syntactic observations have already be made available in Brechet et al. (2012), no intonation analysis has so far been conducted to show the role of prosody in these specific discourse conditions. In this paper, prosodic annotations were performed according to the dependency prosodic structure model of Martin (2018b), where melodic contours located on stressed vowels, classified according to their glissando values (perceived as a melodic change vs. as a static tone), indicate dependency relations between stress accent phrases (i.e. stress groups) and ultimately determine the prosodic structure of the sentence.

Keywords: Intonation, prosodic structure, telephone conversation, customer service.

1. *Introduction*

Sentence intonation not only conveys emotions (joy, sadness, angry...), attitudes (arrogant, submissive, cheerful...) and socio-geographical origin (French from Paris, Toulouse, Lille...), but also something quite important for speech comprehension: the prosodic structure of the sentence, prerequisite for an efficient syntactic analysis of running speech by speakers. This prosodic structure is defined by dependency relations between the sentence accent phrases, relations indicated by melodic contours located on vowels of stressed syllables.

In French, a non-lexically stressed language, stressed syllables (excluding emphatic stress) are located of the final syllable of some word (not necessarily a content word, i.e., a verb, a noun, an adjective, or an adverb) with a rhythmic constrain limiting their interval between 250 ms and some 1250-1350 ms in running speech. They determine the

right boundary of the minimal prosodic units instantiated by accent phrases. It has been shown (Martin 2018) that the actual realization of stressed syllables depends on the speaker speech rate, leading to accent phrases containing 8 to 9 syllables for a fast speaker, and only 4 to 6 syllables for a slow speaker.

2. *The role of the prosodic structure in sentence comprehension*

Contrary to isolated sentences, whose acoustic image can be kept in memory for up to 20 or 30 seconds, sentences in running speech have a limited short-time memory of about 2 or 3 seconds. Since this limit does not give sufficient time for listeners to perform a syntactic analysis applied to the perceived string of words, the sentence prosodic structure is essential for language comprehension, as it provides a temporal and structural frame for syntactic decoding (cf. *syntactic bootstrapping*).

Besides, it has been experimentally shown that the perception of accent phrases is synchronized by delta brain oscillations, which explains why the duration of accent phrases in non-lexically stressed languages such as French and Korean is limited to a maximum of 1250-1350 ms (Martin 2018b), the minimal value of 250 ms referring to the time gap between consecutive stressed syllables. It explains as well why actual stress realizations (again excluding emphatic stress) is depending on speech rate (the average duration being 500-600 ms with accent phrases containing 4 to 5 syllables).

2.1 The independent prosodic structure

Contrary to the dominant Autosegmental-Metrical model which envisions prosodic events as percolating from syntax, the analysis proposed here assumes that the prosodic structure is planned and realized by the speaker before syntax in the sentence generation process at least for 3 to 4 accent phrases sequences. Hence, the phonological description of prosodic events should proceed independently from prosodic properties alone.

It is furthermore assumed here that pitch accent, i.e., the melodic movements located on stressed vowels, do interact with each other, as they indicate dependency relations between accent phrases. However, for French, a non-lexically stressed language, pitch accents

and boundary tones are in syncretism and are aligned on the same stressed vowels.

2.2 Description of melodic contours

Prosodic events are described efficiently:

1. From the localization on stressed syllables vowels, excluding emphatic stress.
2. By the categorization of stressed vowels melodic changes according to the following parameters:
   a. Sentence final (reaching the lowest or highest level in the sentence).
   b. Rising or falling.
   c. Above or below the glissando threshold

The glissando threshold determines the limit above which a melodic change is perceived and under which a static tone is perceived. The glissando and the glissando threshold can be approximated from the fundamantal frequency curve F0 in Herz by the formula (Rossi 1971):

Semitone = 12 * (log(F0/100.0)) / log(2.0)
Glissando = (Semitone2-Semitone1) / (t2-t1)
Glissando threshold: between 0,16 / t^2 and 0.32 / t^2

Applying the criteria above, the retained contour categories are:

Cneu → neutralized, rising or falling below the glissando threshold
Cfal ↘ falling, above the glissando threshold
Cris ↗ rising, above the glissando threshold
Cfal# ↘# falling above the glissando threshold, before a pause longer than 250 ms
Cdec ↓ final contour déclarative (lowest frequency)
C0n ← final contour déclarative postnucleus (same as Cneu)
Cint ↑ sentence final contour interrogative (highest frequency)
Cin ↑ sentence final contour interrogative postnucleus (same as Cint).

The sentence nucleus is defined as a segment of a sentence that can constitue a complete well-formed sentence by itself in isolation. It ends with either a terminal conclusive declarative or interrogative contour. The nucleus is eventually followed by one (or more) segment ended by a neutralized contour C0n (declarative case) or interroga-

tive Cin (interrogative case). C0n has the same realization that Cneu, whereas Cin is similar to Cint.

Using prosodic annotation adapted from F-ToBI (Delais et al. 2015) and integrating the glissando threshold, the retained prosodic events categories are:

H*/L* neutralized, under the glissando threshold
L*L- falling, above the glissando threshold
H*H- rising, above the glissando threshold
L*# falling > glissando threshold, before a pause > 250 ms
L*L% final conclusive declarative (lowest frequency)
H*/L* post final declarative
H*H% final conclusive interrogative (highest frequency)
H*H% post final interrogative

2.3 Dependency rules

The sentence prosodic structure is indicated by dependency relations between accent phrases (group of words with only one non-emphatic stressed syllable). These dependency relations are specified by pitch movements aligned on stressed vowels. They determine a hierarchical grouping of accent phrases which define the sentence prosodic structure. For French, dependency rules are as follows (⇨ and ⇦ indicate the direction of the dependency):

Cneu → ⇨ **{Cfal ↘, Cris ↗, Cfal# ↘#, Cdec ↓, Cint ↑}**
Cfal ↘ ⇨ **{Cris ↗, Cfal# ↘#}**
Cris ↗ ⇨ **Cdec ↓**
Cfal# ↘# ⇨ **Cdec ↓**
Cdec ↓ ⇦ **C0n** ← Declarative postnucleus
Cint ↑ ⇦ **Cin ↑** Interrogative postnucleus

For example, the dependency rule Cfal ⇨{Cris, Cfal#} indicates that the presence of a falling contour Cfal above the glissando threshold depends on the occurrence of either a rising Cris or falling contour before pause Cfal# later in the sentence (dependency "to the right"). It indicates the regrouping of all already existing groups of accent phrases with the last one containing Cfal, with all already existing groups of accent phrases where the last one contains either Cris or Cfal#. As there is no dependency of Cfal towards Cdec, Cfal cannot immediately be followed in a sentence by Cdec.

3. *The RATP-DECODA corpus*

RATP-DECODA is a Customer Care Service corpus part of the ORFEO project (2020). It includes 1988 recordings of client requests made between 2009 and 2011 to the RATP call center (Régie Autonome des Transports Parisiens).

The total duration of these calls is about 98 hours, with an average of 177 seconds per call. Recordings were made in a mp3 format, converted in the Orfeo corpus into stereo 16 bit 8000 Hz sampling rate, and for the present research in mono 16 bit 22050 Hz. The files are originally annotated in word and phone segments, in trs (Transcriber) and also into TextGrid (Praat) format.

Conversations usually involve 2 participants: a client, the customer agent and eventually a service voice. Call types as noted in Brechet et al. (2012) were: Info Traffic 22.5%; Route planning 17.2%; Lost and Found 15.9%; Registration card 11.4; Timetable 4.6%; Ticket 4.5%; Specialized calls 4.5%; empty 3.6%; New registration 3.4%; Price info 3.0%.

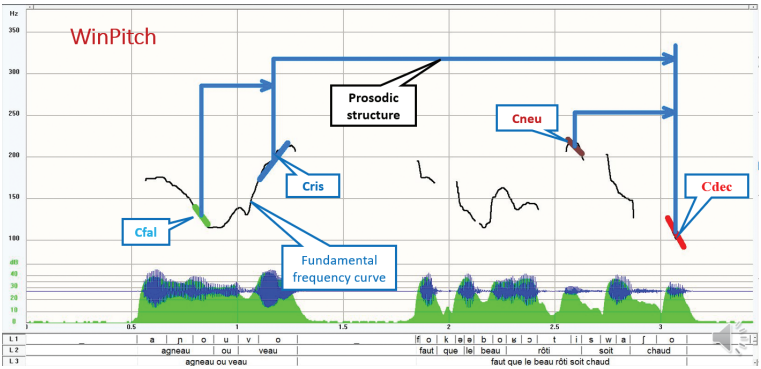
4. *Acoustic analysis*

The telephone landline also used by callers and the mp3 compressed recording format limit the actual sound bandwidth to about 120-3900 Hz. This may possibly affect the reliability of pitch trackers and automatic segmentation programs used for word and phone annotations, especially for male voices.

Nevertheless, using dedicated graphic functions of the speech analysis software WinPitch (2020), errors of segmentation were easily corrected manually, and pitch detection errors have been monitored thanks to the simultaneous display of a narrow band spectrogram allowing more reliable annotations (Martin 2018a).

An example of acoustic analysis using WinPitch is given Fig. 1, with the automatic display of the prosodic structure defined by pitch movements located on stressed vowels. The segmentation into words and API phones is done automatically by comparison with a TTS voice generating the text to analyze.

Figure 1



[*Agn**EAU***] Cfal↘ [*ou v**EAU***]↗ Cris [*faut que le beau rôt**I***] Cneu→ [*soit ch**AU**d*] Cdec↓ "Lamb or veal the beautiful roast must be hot". (stressed vowels are in bold capital), accent phrases are in brackets)

5. *Predicted configurations of the sentence prosodic structure*

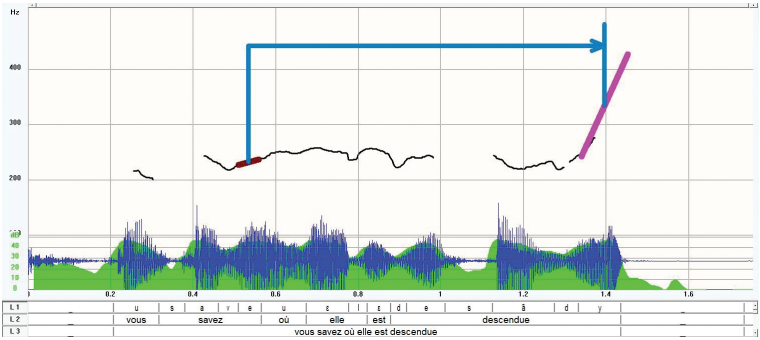
The four possible configurations of an interrogative sentence are:

a. Interrogative questions with or without a morphosyntactic marker ending with a rising terminal conclusive contour.
b. Declarative questions with a morphosyntactic marker falling terminal conclusive contour.
c. Interrogative postnucleus rising contour after the nucleus terminal conclusive interrogative.
d. Declarative postnucleus flat contour after the nucleus terminal conclusive declarative.

5.1 Interrogative questions rising terminal conclusive contour

Fig. 2 and 3 show an interrogative modality mode indicated by a rising terminal interrogative contour only. The first accent phrase of both examples is terminated by a neutralized contour Cneu necessary and sufficient in the configuration of the prosodic structure. The realization of an Cfal above the glissando threshold, while possibly attested, would be redundant.

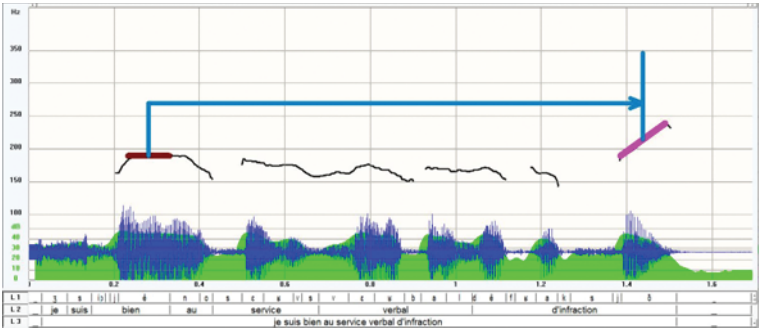
Figure 2



[*Vous sav**E**z*] Cneu → [*où elle est descend**U**e ?*] Cint ↑ "Do you know where she got off?" (20091112_RATP_SCD_0353 25.852 s 27.609 s).

An interrogative questions marked by a terminal rising interrogative contour Cint on the last accent phrase, the first [*Vous sav**E**z*] being ended by a neutralized contour Cneu, as any contour phonologically contrasting with Cdec or Cint would adequately indicate a simple prosodic structure with only two accent phrases.

Figure 3



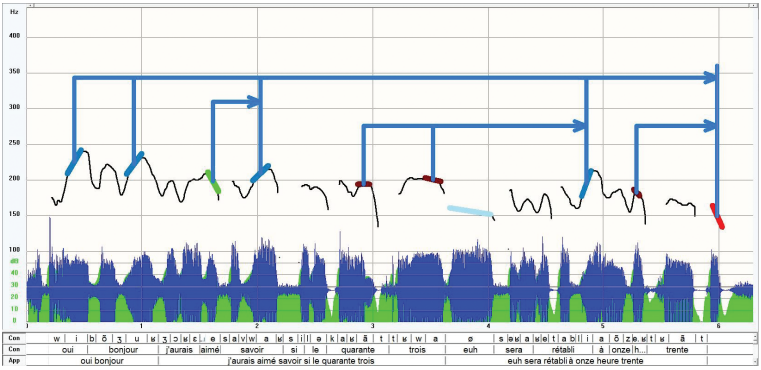
[*Je suis bi**EN***] Cneu → [*au service verbal d'infracti**ON** ?*] ↑Cint "Am I on verbal offense service?" (20091112_RATP_SCD_0215 8.057 s 9.648 s).

Another interrogative questions marked by a terminal rising interrogative contour Cint and a neutralized Cneu on the first accent phrase [*Je suis bi**EN***].

5.2 Declarative questions falling terminal conclusive contour

Examples of Fig. 4 and 5 are declarative sentences whose interrogative modality is only referred by the text (*j'aurais aimé savoir* "I would have liked to know" and *je voudrais savoir* "I would like to know".

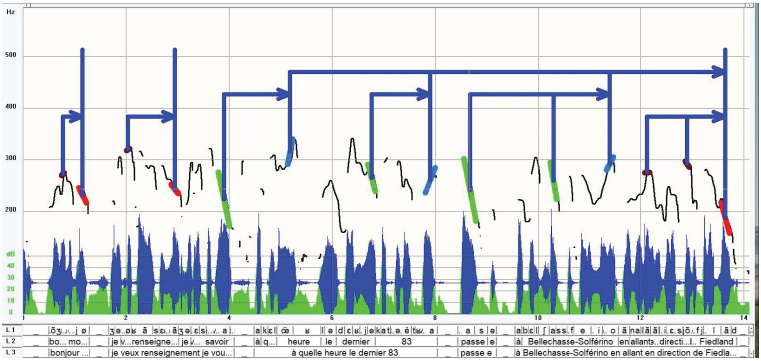
Figure 4



[*OuI*] Cris ↗ [*bonj**OU**r*] Cris ↗ [[*j'aurais aim**É***] Cfal ↘ *sav**OI**r*] Cris ↗ [*si le quar**ANt**e tr**OI**s euh sera rétabl**I***] Cris ↗ [*à onze h**EU**res tr**EN**te*] Cdec ↓ "Yes hello I would have liked to know if the forty three uh will be restored to eleven thirty" (20091112_RATP_SCD_0357 6.236 s 12.339 s).

This is an example of a declarative prosodic structure associated with a text with an indirect interrogation instantiated by *j'aurais aimé savoir*.

Figure 5



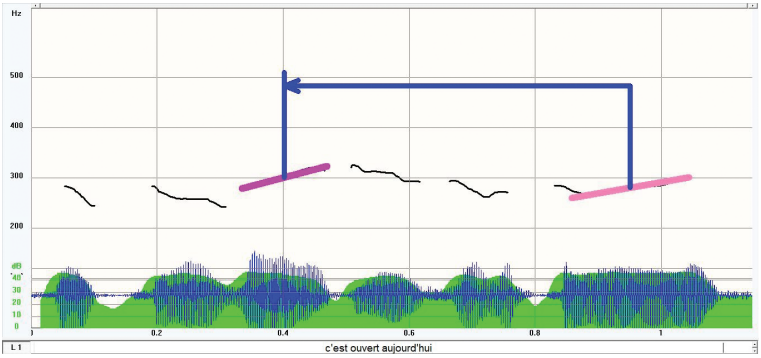
[[*Je voudrais sav**OI**r*] Cfal ↘ *à quelle h**EU**re*] Cris ↗ [[*le dern**iE**r*] Cfal ↘ *quatre vingt-tr**OI**s*] Cris ↗ [[*p**A**sse*] Cfal ↘ [*à Bellech**A**sse*] Cfal ↘ *Solférin**O***] Cris ↗ [*en all**AN**t en directi**ON** de Friedl**AN**d*] Cdec ↓ "I would like to know at what time the last 83 passes Bellechasse-Solférino heading towards Friedland". (20091112_RATP_SCD_0336 6.686 s 20.790 s).

Another example where both the prosodic structure and the associated text carry a declarative modality. The acoustic analysis shows the melodic slope contrast involving the dependency rule Cfal → Cris characteristic of French marking the dependency between accent phrases: [[*Je voudrais sav**OI**r*] *à quelle h**EU**re*], [[*le derni**E**r*] *quatre vingt-tr**OI**s*], [[*p**A**sse*] [*à Bellech**A**sse*] *Solférin**O***] Cris ↗.

5.3 Interrogative postnucleus rising contour after the nucleus terminal conclusive interrogative

Examples of Fig. 6 and 7 are interrogative sentences, deprived from morphosyntactic interrogative markers, and whose modality is indicated by a terminal conclusive interrogative contour Cint, followed by a postnucleus ending with a copy of Cint. This is an example of a *focus-topic* configuration.

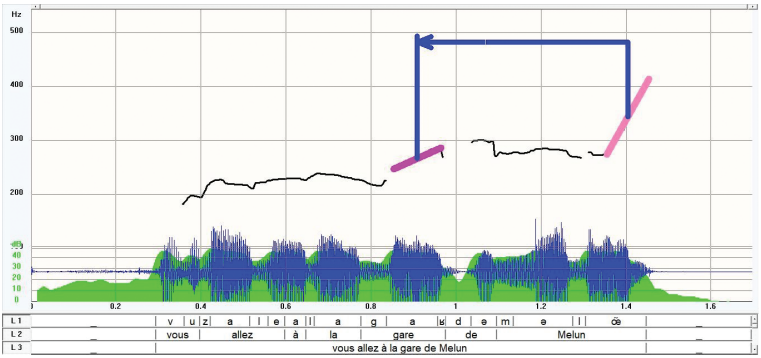
Figure 6



[*C'est ouv**E**rt*] Cint ↑[*aujourd'hu**I** ?*] Cin ↑ "Are you open today?"
(20091112_RATP_SCD_0060 18.045 s 21.860 s 2.560 s 3.705 s).

In this example, [*C'est ouv**E**rt*] is the nucleus, and [*aujourd'hu**I** ?*] the interrogative postnucleus, ended by a copy of the first interrogative Cint of the first accent phrase [*C'est ouv**E**rt*].

Figure 7



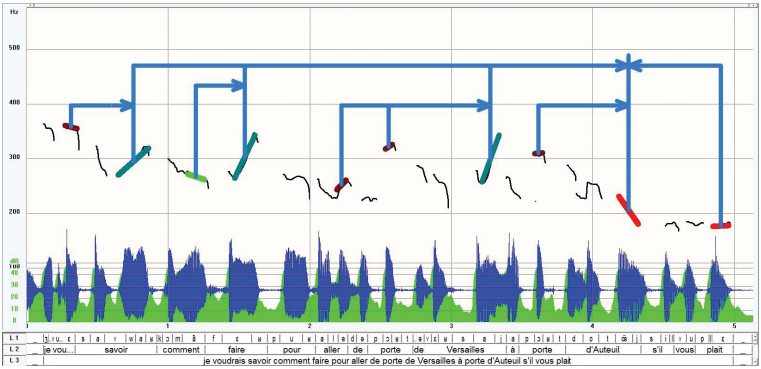
[*Vous allez à la g**A**re*] Cint ↑[*de Mel**UN** ?*] Cin ↑ "Are you going to Melun station?" (20091112_RATP_SCD_0023 16.420 s 18.118 s).

The nucleus is [*Vous allez à la g**A**re*] and the postnucleus [*de Mel**UN** ?*].

5.4 Declarative postnucleus flat contour after the nucleus terminal conclusive declarative

Fig. 8 and 9 show examples of the declarative counterpart of preceding cases of Fig. 6 and 7. The postnucleus *s'il vous plait* "please" are ended with a neutralized contour Cneu in a *focus-topic* declarative configuration.

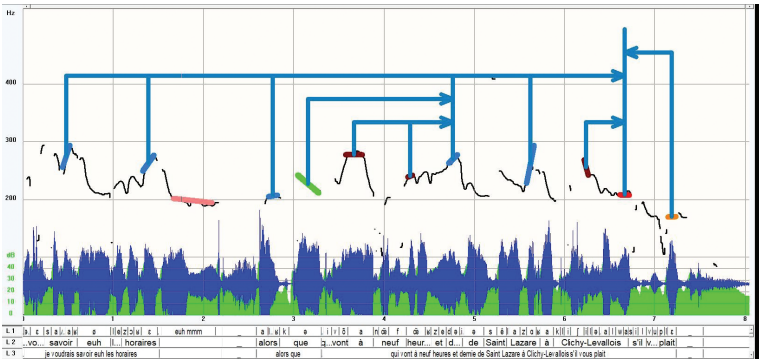
Figure 8



[*Je voudr**AI**s sav**OI**r*] Cris ↗ [[*comm**ENt***] Cfal↘ [*f**AI**re*] Cris ↗ [*pour all**ER** de P**O**rte de Ver**A**illes*] Cris ↗ [*à P**O**rte d'Aut**EU**il*] Cdec ↓ [*s'il vous pl**AI**t*] C0n ↓ "I would like to know how to get from Porte de Versailles to Porte d'Auteuil please" (20091112_RATP_SCD_0344 8.067 s 13.197 s).

Here the nucleus is [*Je voudr**AI**s sav**OI**r comm**EN**t f**AI**re pour all**ER** de P**O**rte de Ver**A**illes à P**O**rte d'Aut**EU**il*] and the postnucleus [*s'il vous pl**AI**t*].

Figure 9



[*Je voudrais sav**OI**r*] Cris ↗ euh [*les hor**AI**res*] Cris ↗ *euh* [*al**O**rs*] Cris ↗ [[*que*] Cfal ↘ *qui v**ON**t à neuf h**EU**res et dem**I**e*] Cris ↗ [*de Saint Lazarre*] Cris ↗ [*à Clich**Y**-Levallois*] Cdec ↓ [*s'il vous pl**AI**t*]
C0n ← "I would like to know uh the hours uh mmm whereas which go at half past nine from Saint-Lazare to Clichy-Levallois please" (20091112_RATP_SCD_0070 8.794 s 16.839 s).

In this example, [*Je voudrais sav**OI**r* euh *les hor**AI**res euh al**O**rs que qui v**ON**t à neuf h**EU**res et dem**I**e de Saint Lazarre à Clich**Y**-Levallois*] is the nucleus and [*s'il vous pl**AI**t*] the postnucleus.

6. *In summary*

These few examples aim to show that sentence intonation is not "the cherry on the syntactic tree", but on the contrary shapes speech production both on the paratactic and syntactic axis. Its importance in actual speech perception stems essentially from the fact that listeners must quickly analyze the content of a sentence, given the allowed short time of speech memory given in continuous speech (2 to 3 seconds).

References

Bechet, Frederic & Maza, Benjamin & Bigouroux, Nicolas & Bazillon, Thierry & El-Bèze, Marc & De Mori, Renato & Arbillot, Eric. 2012. DECODA: a call-centre human-human spoken conversation corpus. In

Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12), pp. 1343-1347.

Delais-Roussarie, Elisabeth & Post, Brechtje & Avanzi, Mathieu & Buthke, Carolin & Di Cristo, Albert & Feldhausen, Ingo & Jun, Sun-Ah & Martin, Philippe & Meisenburg, Trudel & Rialland, Annie & Sichel-Bazin, Rafèu & Yoo, Hi-Yon. 2015. Intonational Phonology of French: Developing a ToBI system for French. In Sónia Frota & Pilar Prieto (eds.), *Intonation in Romance*, 63-100. Oxford: Oxford University Press.

Martin, Philippe. 2018a. Prosodic annotation in adverse recording conditions. In De Dominicis, Amedeo (a cura di), *International Workshop. Speech audio archives: preservation, restoration, annotation, aimed at supporting the linguistic analysis*, Accademia Nazionale dei Lincei.

Martin, Philippe. 2018b. *Intonation, structure prosodique et ondes cérébrales*, London: ISTE.

ORFEO. 2020. Corpus d'étude pour le français contemporain https://www.projet-orfeo.fr/

Rossi, Mario. 1971. Le seuil de glissando ou seuil de perception des variations tonales pour les sons de la parole. *Phonetica* 23, 1–33.

WinPitch. 2020. Computer program, www.winpitch.com

DOI: 10.17469/O2106SLI000022

ANNA-MARIA DE CESARE

La concezione delle congiunzioni e degli avverbi negli schemi di annotazione dei corpora d'italiano scritto: breve ricognizione e alcune proposte

Il presente studio vuole fare il punto sulle PoS usate nell'annotazione dei più recenti corpora d'italiano scritto. Per motivi di spazio, ci soffermiamo su due PoS: quelle legate alle congiunzioni e agli avverbi. Gli obiettivi perseguiti sono i seguenti: ricostruire la concezione teorico-descrittiva delle congiunzioni e degli avverbi nella linguistica dei corpora legata all'italiano; valutare quanto questa concezione sia vicina a quella della grammatica tradizionale e fare emergere gli elementi innovativi; infine, alla luce dei risultati ottenuti, formulare una serie di desiderata e proposte per rivedere le PoS associate alle congiunzioni e agli avverbi, tenendo conto delle recenti messe a punto della ricerca teorica e computazionale.

Parole chiave: schemi di annotazione, PoS-tag, congiunzioni, avverbi, corpora d'italiano scritto.

1. *Introduzione*

Tra i molteplici livelli ai quali si può annotare un corpus, il più basico, e da decenni ormai ritenuto standard (cfr. le linee guida EAGLES discusse in Monachini 1995), è senza dubbio quello relativo al markup delle parti del discorso (nome, verbo, aggettivo, pronome, articolo, avverbio, preposizione, congiunzione e interiezione) – nella linguistica dei corpora più comunemente chiamate *parts of speech* (PoS). Si tratta di un livello di annotazione linguistica del testo che interessa la morfosintassi, e la cui unità di riferimento è la *parola*.[1]

[1] Come è noto, il termine 'parola' non ha nessuna concretezza a livello teorico: una sua definizione più puntuale si basa su criteri di natura grafica, morfologica e seman-

Tenuto conto del loro carattere ontologico, il presente studio si propone di fare il punto sulle PoS usate nell'annotazione dei corpora d'italiano scritto (contemporaneo). Per motivi di spazio, ci concentreremo qui solo sulle PoS legate alle parti invariabili del discorso, in particolare sulle congiunzioni e sugli avverbi. Gli obiettivi perseguiti in questa sede sono tre:[2]

1. Ricostruire la concezione teorico-descrittiva di congiunzioni e avverbi nella linguistica dei corpora dell'italiano scritto (§ 2).
2. Valutare quanto questa concezione sia vicina a quella della grammatica tradizionale e fare emergere gli elementi innovativi, in sintonia con proposte teoriche recenti (§ 3).
3. Chiarire la necessità di rivedere le PoS associate alle congiunzioni e agli avverbi; sarebbe a nostro avviso utile tenere conto in modo più sistematico delle recenti messe a punto della ricerca – sia teorica sia computazionale – sulle PoS (§ 4).

2. *La concezione teorico-descrittiva delle congiunzioni e degli avverbi nella linguistica dei corpora dell'italiano scritto: una ricognizione*

La nostra ricognizione si basa su due aspetti: la natura delle etichette (E. *tags*) relative alle congiunzioni e agli avverbi nei *tagset* di dieci recenti corpora d'italiano scritto; e la natura dei lemmi etichettati come congiunzione e avverbio nelle liste di tre corpora a confronto. Il primo aspetto permette soprattutto di capire le proprietà definitorie delle due PoS studiate in questa sede; il secondo, invece, fa luce sull'estensione di queste due classi di parole.

2.1 Analisi delle etichette nei tagset di dieci corpora d'italiano scritto

Gli schemi di annotazione relativi al livello morfosintattico, in particolare alle parti del discorso, sono stati analizzati in modo attento per l'italiano da ultimo nel rapporto tecnico di Venturi 2009 (cfr. anche Barbera 2007b). Il rapporto si sofferma su 12 sistemi diversi per an-

tica. Per una riflessione teorica in merito, rimandiamo alle pagine di Graffi (1994: 36-37) e Chiari (2007: 49).

[2] Tralasceremo qui tutte le questioni relative all'applicazione automatica di PoS-*tagger*, ampiamente discusse nell'ambito della linguistica computazionale (cfr., tra altri, Bernardi *et al*. 2006 e Tamburini 2016 sull'italiano; Schmid 1994 per altre lingue).

notare le PoS, di cui otto sono concepiti per annotare testi scritti (tra questi, uno è pensato per i testi antichi: cfr. Barbera 2007a), tre per annotare discorsi parlati e un ultimo per annotare testi appartenenti a entrambe le tipologie (cfr. Venturi 2009: 5).

Nel presente studio ci soffermiamo su tre (degli otto) schemi di annotazione dell'italiano scritto considerati da Venturi 2009; si tratta di schemi applicati a dieci recenti corpora (cfr. i dati della Tabella 1 in Appendice):

a. il PoS-tagset concepito seguendo le linee guida EAGLES (Monachini 1995), applicato al CORIS-CODIS (cfr. Tamburini 2000);
b. lo schema progettato da Marco Baroni, che si configura oggi come quello più fortunato (è applicato a ben 8 dei 10 corpora da noi considerati);
c. lo schema di Achim Stein, applicato al corpus Araneum Italicum Maius (per una descrizione più puntuale, cfr. § 2.2).

L'ordine in cui abbiamo menzionato i tre schemi di annotazione riflette la loro ideazione cronologica: il PoS-tagset descritto al primo punto è stato sviluppato prima degli altri due; inoltre, come già indicato, lo schema di Stein si basa su quello di Baroni. Non per questo, e lo vedremo nei prossimi paragrafi, le PoS relative alle congiunzioni e agli avverbi sono più articolate nell'ultimo schema.

2.2 Valutazione delle *tag* relative alle congiunzioni e agli avverbi

Le etichette concepite per le congiunzioni e gli avverbi nei tre schemi di annotazione considerati sono molto semplici, soprattutto se paragonate a quelle di PoS variabili, *in primis* del verbo e del nome. Nei tre schemi considerati, le congiunzioni e gli avverbi sono associati a una o (al massimo) due *tag* (cfr. Tabella 1 in Appendice).

Nel PoS-tagset sviluppato in base agli standard EAGLES (applicato al CORIS-CODIS), c'è una sola *tag* associata agli avverbi (ADV), mentre sono due le *tag* applicate alle congiunzioni: CONJ_C (che sta per Congiunzioni Coordinanti) e CONJ_S (per congiunzioni subordinanti). Nello schema di annotazione previsto da Baroni si trova invece una sola *tag* per le congiunzioni (CON), mentre sono due le *tag* concepite per gli avverbi: ADV_mente (per gli avverbi in *-mente*) e ADV (per tutti gli altri avverbi). Infine, nello schema di Stein tro-

viamo una sola *tag* per gli avverbi (ADV) e una sola *tag* anche per le congiunzioni (CON).

Un confronto tra i tre schemi di annotazione permette di rilevare alcune differenze nella scelta delle *tag* associate alle congiunzioni e agli avverbi. Nel caso delle congiunzioni, o si usa una sola etichetta generale, chiamata 'Conjunction' (Baroni, Stein), oppure si usano due etichette più specifiche, chiamate 'Coordinating conjunction' e 'Subordinating conjunction', senza che vi sia una *tag* generale per la classe (EAGLES). Per quanto riguarda gli avverbi, o si usa l'etichetta generale 'Adverb' (EAGLES, Stein), o si opta per due etichette (Baroni): 'Adverb' (che va in questo caso intesa alla luce della seconda: sono tutti gli avverbi non derivati con il suffisso *-mente*) e 'Adverb in -mente'. Nei tre schemi di annotazione considerati, l'interpretazione dell'etichetta ADV non è dunque uniforme.

A quanto detto si aggiunge una differenza relativa ai criteri in base ai quali le due PoS si articolano al loro interno: le congiunzioni sono suddivise in base a un criterio sintattico (coordinazione vs. subordinazione), mentre gli avverbi sono distinti in base a un criterio morfologico (± derivazione con il suffisso *-mente*).

2.3 Analisi dei lemmi etichettati CON e ADV: dati quantitativi

Passiamo ora all'analisi dei lemmi etichettati come 'congiunzione' (CON) e 'avverbio' (ADV) in tre corpora d'italiano presi a campione tra i dieci considerati:[3]

1. Timestamped JSI web corpus Italian 2014-2021 (TIt)
2. itTenTen16 (itTT)
3. Araneum Italicum Maius (AItM)

I tre corpora sono stati selezionati in base al fatto che presentano importanti punti in comune: la loro modalità di creazione (tramite web-crawling) e la tipologia alla quale appartengono. Si tratta di 'corpora di ambito generale' (così secondo SkE), con un'importante area di sovrapposizione relativa alla prosa giornalistica. Allo stesso tempo i tre corpora si differenziano per almeno due aspetti: la loro dimensione e il loro schema di annotazione (cfr. Tabella 2).

[3] I tre corpora sono disponibili sulla piattaforma Sketch Engine (SkE), cfr. Kilgarriff *et al*. 2014.

Tabella 2 - *Proprietà di tre corpora a confronto*

Corpus	*Dimensione (n. di parole)*	*File di parametri*
TIt	ca. 8,7 miliardi	TT Baroni
itTT	ca. 5 miliardi	TT Baroni
AItM	ca. 900 milioni	TT Stein

La frequenza assoluta dei lemmi etichettati CON e ADV nei tre corpora è proposta nella Tabella 3.[4]

Tabella 3 - *Frequenza assoluta dei lemmi CON e ADV*

Corpus	*CON*	*ADV*
TIt	50	16.154
itTT	47	38.124
AItM	783	1.912

Da un punto di vista quantitativo generale va dapprima rilevato che la categoria degli avverbi (che include anche quelli in *-mente*) ha un'estensione molto maggiore di quella delle congiunzioni. Tale differenza riflette proprietà semantiche e categoriali ben note delle due classi (cfr. De Cesare 2019: 23-25): gli avverbi formano una classe lessicale aperta, i cui elementi costitutivi hanno un significato perlopiù denotativo; le congiunzioni entrano invece in una classe grammaticale chiusa, i cui membri veicolano un significato di tipo istruzionale.

Se osserviamo poi i risultati ottenuti per le due PoS in modo separato, spicca soprattutto la differenza tra i dati dei tre corpora. Il dato relativo agli avverbi è senza dubbio più difficile da spiegare: non vi è infatti nessuna correlazione tra la dimensione dei tre corpora e la quantità di avverbi presenti in ognuno di essi. Il numero di ADV presenti in TIt e itTT (due corpora annotati con lo stesso schema), è inversamente proporzionale alla dimensione delle due risorse: sono due volte meno numerosi nel primo corpus (16.154 vs. 38.124), che è però

[4] Pur essendo disponibili sulla stessa piattaforma (SkE), i tre corpora non sono sempre interrogabili allo stesso modo perché presentano interfacce in parte diverse. Per ottenere informazioni sui lemmi associati alle due PoS che ci interessano bisogna seguire due cammini distinti: nel caso di TIt/itTT si possono cercare i lemmi taggati CON e ADV direttamente nella "lista di frequenza". Nel caso di AItM, invece, si deve per forza partire da una ricerca relativa a tutte le *tag*, per poi selezionare i lemmi relativi alle CON e agli ADV.

due volte più ampio del secondo (8,7 vs 5 miliardi di parole). Vi è poi il fatto che il numero di avverbi in AItM è sorprendentemente basso (1912). La differenza tra i corpora (almeno tra i due primi) sembra da ricondurre alla natura dei testi di cui si compongono. Per capire meglio i risultati ottenuti, bisognerebbe proporre un'analisi più approfondita, che in questa sede non siamo tuttavia in grado di offrire.

I dati relativi alle congiunzioni sono in parte più facili da interpretare: i due primi corpora (TIt e itTT) presentano praticamente lo stesso numero di lemmi etichettati CON (rispettivamente 50 e 47). Questa cifra riflette bene le proprietà intrinseche della classe, che non è solo chiusa, ma anche composta da un numero relativamente ridotto di membri. In altre parole, diversamente dagli avverbi, il numero di lemmi taggati CON non dipende dalla dimensione del corpus. Alla luce di queste considerazioni, sorprende dunque ancora una volta il dato ottenuto per il corpus AItM, che è però questa volta inaspettatamente elevato (783). Una verifica più puntuale dei lemmi associati alla *tag* CON in questo corpus permette di fare luce sull'ampio numero di forme riportate: AItM presenta un'etichettatura 'difettosa' almeno a partire dai lemmi che occupano il rango d'uso 100. Se scorriamo i lemmi etichettati CON dopo il rango 100, troviamo per esempio forme come 'O_o', 'intr-o', 'e-o', 'Ukiyo-e', 'WEB-MA', 'saxdax-e', che non hanno chiaramente nulla a che fare con le CON, e non sono neanche veri e propri lemmi. Si tratta in molti casi semplicemente di forme che includono una congiunzione prototipica (*e, ma, o*).

2.4 Analisi dei lemmi etichettati CON: valutazione qualitativa

Soffermiamoci ora sulle forme di CON che entrano a far parte delle tre liste di frequenza (nel caso del corpus AItM limitiamo l'analisi ai primi 50 lemmi).

Nei tre corpora le teste di lista sono assolutamente identiche: troviamo (nello stesso ordine) *e, ma, o, ed* e *se*. Sono poi presenti, ma non più allo stesso rango d'uso, la variante grafica *od* e le varianti grafico-semantiche *ovvero, oppure* e *ossia*. Inoltre, sempre nei tre corpora, sono inclusi *né* e la variante *nè* (la prima occupa sempre un rango d'uso più elevato), *sebbene* e *seppure*. Altre CON presenti in tutte e tre le liste sono *mentre* e *sia* (che occupano ranghi elevati), *nonostante* e *bensì* (che occupano ranghi intermedi).

Comune alle tre liste è poi la presenza di *che*. Il lemma non occupa però gli stessi ranghi d'uso, né è associato alla stessa frequenza. In TIt e itTT, *che* ha una frequenza d'impiego molto bassa (esso compare, rispettivamente, 1.388 e 1.732 volte) e occupa uno degli ultimi ranghi. In AItM, invece, *che* compare 110269 volte e occupa il rango 10. La differenza tra TIt e itTT da una parte e AItM dall'altra si spiega facilmente alla luce del fatto che i primi due corpora prevedono una *tag* ad hoc per la parola *che*, ovvero CHE (torneremo a parlare di questa scelta nel § 3.2).

Tutte e tre le liste di CON includono anche un numero piuttosto consistente di lemmi composti da *che*, come per esempio (citando solo le cinque voci più frequenti) *nonché, affinché, poiché, finché, anziché* (TIt); *nonché, poiché, affinché, finché, anziché* (itTT); *perché, perchè, nonché, poiché, affinché* (AItM). I lemmi che contengono la forma *che* sono numerosi soprattutto nei due primi corpora, dove formano circa la metà delle due liste. In AItM, i lemmi formati con *che* sono invece meno numerosi (ci sono ca. 10 entrate in meno). Il dato si spiega con il fatto che nel terzo corpus altri lemmi taggati CON hanno una frequenza d'uso più elevata e occupano i primi 50 ranghi (declassando a ranghi più bassi forme marginali di CON basate su *che*: è il caso di *sicché, giacché, allorché*). Si tratta in particolare di *come, quando, visto, oltre, qualora, così, salvo, ebbene.* Queste forme sono CON solo in AItM. In TIt e itTT sono quasi tutte etichettate WH. Lo stesso vale del resto anche per *perché*.

Va infine rilevato che nelle tre liste di CON compaiono forme come *dopo, tra / fra* ecc., che possiamo descrivere in modo unitario: sono parole polifunzionali, principalmente taggate PRE. Per queste forme si osservano però differenze importanti tra i tre corpora. Prima di tutto in merito al loro numero: ve ne sono due in itTT (*senza* e *dopo*), quattro in TIt (*dopo, senza, tra* e *fra*) e sei in AItM (*fino, prima, per, senza, dopo, a*). Un'altra differenza riguarda la loro frequenza d'uso: *dopo*/CON è per esempio molto più frequente nei due primi corpora (903.400 occ. in TIt; 298.125 occ. in itTT e 728 occ. in AItM). In questo caso, di nuovo, le differenze non si spiegano unicamente in base alla diversa dimensione dei tre corpora (vi sono tre volte più occorrenze in TIt che in itTT).

3. *Valutazione delle PoS associate alle congiunzioni e agli avverbi: tra concezione tradizionale e proposte innovative*

3.1 Cenni sulla concezione tradizionale delle congiunzioni e degli avverbi

Nella lezione tradizionale sulle parti del discorso (nell'ambito delle grammatiche italiane rappresentata per esempio da Serianni 2000), le parole sono raggruppate in classi (o paradigmi) in base alla loro flessione: a un primo livello di analisi, si distinguono le parole variabili da quelle invariabili. Le parti variabili possono essere ulteriormente distinte (in nome, verbo, aggettivo, articolo, pronome) osservando più da vicino i tratti flessivi di cui si compongono (per es. il genere, numero, la persona, il tempo ecc.). Questo *modus operandi* non è invece possibile per le parole invariabili che – per definizione – non possono essere flesse. La distinzione tra le quattro classi invariabili del discorso (avverbio, congiunzione, preposizione e interiezione) verte dunque su criteri di natura semantica e/o sintattica (De Cesare 2019: 26-27).

In parte a seguito della natura dei criteri in base ai quali sono suddivise, le categorie invariabili del discorso non hanno confini molto netti. Basta considerare l'esempio paradigmatico della parola invariabile *dopo*, che secondo la grammatica tradizionale può fungere da avverbio, congiunzione e preposizione. La sua specifica classe di appartenenza dipende dalle proprietà sintattiche che intrattiene con il resto della frase. Quando la parola *dopo* non entra in relazione con nessun altro costituente seguente, cioè non regge un complemento, la si assegna alla categoria degli avverbi (cfr. 1); quando, invece, *dopo* regge una frase (anche di modo finito, come *aver sostenuto l'esame*), abbiamo una congiunzione (2); e quando regge un sintagma nominale (come *l'esame*), una preposizione (3):

(1) Ti sentirai meglio dopo. (avverbio)
(2) Ti sentirai meglio dopo *aver sostenuto l'esame*. (congiunzione)
(3) Ti sentirai meglio dopo *l'esame*. (preposizione)

Nella grammatica tradizionale ognuna delle quattro categorie invariabili è ulteriormente suddivisa in base a criteri morfosintattici. Nel caso delle due classi che ci interessano, possiamo osservare la suddivisione riportata in Tabella 4: le congiunzioni sono suddivise in tre gruppi (De Cesare 2019: 41); gli avverbi, invece, in quattro gruppi (De Cesare 2019: 52).

Tabella 4 - *Suddivisione delle congiunzioni e degli avverbi (ed esempi)*

Congiunzioni	*Avverbi*
semplici (*e, né, ma, quindi, anche*)	semplici (*ora*)
composte (*oppure*)	composti (*soprattutto*)
locuzioni (*dal momento che, visto che*)	locuzioni (*d'ora in poi*)
	derivati (*ovviamente, bocconi*)

3.2 CON e ADV: tra tradizione e innovazione

Complessivamente, la nostra indagine – che andrebbe naturalmente allargata e approfondita – permette di osservare che la concezione teorico-descrittiva delle congiunzioni e degli avverbi associata ai corpora d'italiano scritto più recenti è relativamente tradizionale.

Questo è vero prima di tutto se osserviamo le *tag* previste nei tre schemi usati per annotare dieci recenti corpora d'italiano scritto: in questi schemi troviamo infatti due etichette di primo livello – CON per le congiunzioni e ADV per gli avverbi. Inoltre, alla stregua delle scelte della grammatica tradizionale, le articolazioni interne alle due *tag* si basano su criteri sintattici (CONJ_C e CONJ_S) e morfologici (cfr. ADV_mente vs. ADV, che copre tutte le altre forme): le *tag* più specifiche sono dunque concepite in base a criteri di natura eterogenea. A questo punto bisogna però notare una cosa importante: le scelte adottate per etichettare i corpora analizzati sono più semplici di quelle della grammatica tradizionale. In effetti, come abbiamo visto (cfr. Tabella 4) la grammatica tradizionale individua ben quattro categorie morfologiche diverse dell'avverbio (cfr. De Cesare 2019: 52): gli avverbi semplici, composti, derivati e le locuzioni avverbiali.

Un confronto tra i tre schemi di annotazione permette anche di osservare delle differenze tra i *tagset*. Lo schema relativo agli standard EAGLES è quello più tradizionale. Diversamente dagli altri due (quello di Baroni e di Stein), esso non prevede una serie di *tag* di primo livello legate alle congiunzioni e agli avverbi. Pensiamo in particolare alle etichette CHE e WH-WORD, ispirate alle proposte della linguistica teorica: la prima *tag* si rifà ad una nuova concezione della congiunzione *che*, concepita come 'complementatore'; la seconda *tag* accoglie invece la proposta di raggruppare parole come *chi, dove, perché* ecc. – nella grammatica tradizionale concepite come trasversali agli avverbi, pronomi e aggettivi (per una discussione, cfr. Salvi 2013: 37-39) – in un'unica nuova categoria (le *wh-word* della grammatica generativa).

Se spostiamo la nostra attenzione sulle forme che entrano a far parte del lemmario delle CON, possiamo osservare che nelle liste di frequenza dei tre corpora analizzati (annotati con i due schemi più recenti, di Baroni e Stein) sono presenti molte parole tradizionalmente concepite come congiunzioni, a cominciare dalle forme prototipiche di congiunzioni coordinanti (*e, ma, o*) – con le rispettive varianti grafiche (*ed, od*) e grafico-semantiche (*oppure, ossia, ovvero*) – e subordinanti (*se, che*).

Detto questo, è doveroso mette in luce che le tre liste analizzate presentano almeno un aspetto innovativo importante. Forme come *quindi, inoltre* e *anche*, che la grammatica tradizionale annovera tra le congiunzioni, non sono etichettate CON, ma ADV – si tratta dunque di avverbi (tali forme non compaiono nella lista delle ca. 700 CON del corpus AItM).

Per il resto, sia a livello terminologico che a quello concettuale, i corpora esaminati tengono poco conto delle proposte mosse negli ultimi anni nell'ambito della linguistica generale (cfr. Prandi 2007; Salvi 2013, 2014; De Cesare 2019) e nel campo della linguistica computazionale (Bernardi *et al*. 2006; D'Errico *et al*. 2016). Una di queste riguarda la categoria delle congiunzioni. Entrambi gli approcci (quello teorico della linguistica generale e quello a-teorico di quella computazionale) argomentano a favore dell'eliminazione di questa parte invariabile del discorso e propongono di ridistribuire le forme tradizionalmente concepite come congiunzioni nella classe degli avverbi (*inoltre, anche*), delle preposizioni (*dopo, fra, perché*) e degli operatori detti 'logici' o 'sintattici' (*e, o, ma; che, se*). Accogliere questa proposta nei file di parametri da applicare ai corpora d'italiano significherebbe anche elaborare maggiormente le *tag* associate alla categoria degli avverbi, proponendo un'articolazione più fine delle etichette di secondo livello.

4. *Conclusioni: tra desiderata e proposte*

I corpora rappresentativi dell'italiano scritto dell'ultima generazione, di cui abbiamo analizzato vari esempi paradigmatici (TIt, itTT e AiTM), si distinguono per la loro modalità di creazione e dimensione: sono il risultato di sistemi di webcrawling e contengono miliardi di parole. A fronte della rapida evoluzione registrata negli ultimi anni

nell'ambito della creazione di nuovi corpora, i progressi fatti nel campo della loro annotazione grammaticale sono piuttosto contenuti.[5]

La nostra analisi delle etichette concepite per le congiunzioni e gli avverbi negli schemi di annotazione impiegati in dieci recenti corpora d'italiano scritto (contemporaneo) permette di affermare che il metalinguaggio della linguistica dei corpora italiana non ha ancora recepito e integrato le proposte della linguistica teorica e computazionale: per l'annotazione delle due PoS impiega un tagging basato su una concezione tradizionale delle congiunzioni e degli avverbi e le *tag* sono relativamente fisse (non si registrano grandi cambiamenti dallo schema basato sugli standard EAGLE agli schemi successivi).

Alla luce dei risultati ottenuti in questo studio ci sembra dunque doveroso esprimere un desiderio: quello di tornare a riflettere sulle *tag* relative alle congiunzioni e agli avverbi negli schemi di annotazione dei testi d'italiano scritto (ma non solo). Converrebbe a nostro parere concepire nuove *tag* (di primo, ma anche di secondo livello), basate non più tanto sulle proprietà morfologiche delle parole quanto su quelle sintattiche (una riflessione importante in questo senso è in Schütze 1995 e Tamburini 2016).

Un buon punto di partenza per la revisione delle PoS associate agli avverbi è lo schema di annotazione morfosintattico del Turin University Treebank/TUT (Lesmo *et al*. 2002). In sintonia con la ricerca teorica sulle parti del discorso, in TUT parole come *sì* e *no* non sono più etichettate ADV (seguendo la concezione della grammatica tradizionale) ma PHRAS (che sta per *phrasals* 'frasali'), una nuova *tag* di primo livello concepita per la classe delle profrasi (sulle proprietà definitorie della categoria, cfr. Bernini 1995; per una discussione relativa al riassestamento del quadro relativo alle parti invariabili del discorso, cfr. De Cesare 2019: 59-62).

Lo schema TUT prevede poi varie *tag* di secondo livello, che sono tuttavia ancora largamente basate su criteri semantici tradizionali (basta considerare le classi degli avverbi di maniera, di quelli locativi, temporali, negativi e quantificativi). Da ciò consegue che nello schema TUT mancano alcune sottoclassi di avverbi ormai ben consolidate a livello teorico-descrittivo, tra cui quelle degli avverbi connettivi (in TUT, *inoltre* è per esempio taggato congiunzione coordinativa), frasa-

[5] La riflessione teorico-descrittiva sulle PoS non manca però in rapporto ai corpora d'italiano antico (cfr. Barbera 2007a; Iacobini *et al*. 2017).

li (sempre in TUT, *forse* è avverbio di dubbio) e focalizzanti. Questi ultimi includono parole come *anche, persino/perfino* (in TUT concepiti come 'avverbi di intensità'), *solo-soltanto* (etichettati 'avverbi di limitazione') e *nemmeno-neanche* (concepiti come 'avverbi di negazione').[6]

Una revisione delle PoS associate alle congiunzioni e agli avverbi permetterebbe di migliorare l'etichettatura automatica dei corpora. I risultati ottenuti per la parola *dopo* bastano per mostrare che una concezione tradizionale delle PoS è molto problematica. In itTT, per esempio, *dopo* è spesso etichettata CON laddove – almeno così secondo il punto di vista della grammatica tradizionale – si tratta chiaramente di un avverbio: cfr. i sintagmi *due mesi dopo* e *otto giorni dopo* nella schermata dell'Appendice 2. La ricerca teorica recente risolve questo problema raggruppando parole come *dopo* in un'unica classe: quella della preposizione (per dettagli, cfr. Salvi 2013: 119-121).

Problemi di etichettatura come quello appena illustrato andrebbero risolti al più presto, vuoi perché molte congiunzioni e molti avverbi (con significato non denotativo) sono associati a una frequenza d'uso elevata, vuoi perché hanno ricadute importanti sui risultati di altre ricerche, come la creazione di Word Sketches. Una revisione delle etichette relative alle congiunzioni e agli avverbi aumenterebbe il grado di accuratezza del tagging automatico e, di conseguenza, la descrizione dei dati linguistici.

Riferimenti bibliografici

Barbera, Manuel. 2007a. Un tagset per il corpus Taurinense. Italiano antico e linguistica dei corpora. In Barbera, Manuel & Corino, Elisa & Onesti, Cristina (a cura di), *Corpora e linguistica in rete*, 135-168. Perugia: Guerra Edizioni.

Barbera, Manuel. 2007b. Mapping dei tagset in bmanuel.org / corpora.unito.it. Tra *guidelines* e prolegomeni. In Barbera, Manuel & Corino, Elisa & Onesti, Cristina (a cura di), *Corpora e linguistica in rete*, 135-168. Perugia: Guerra Edizioni.

Baroni, Marco & Bernardini, Silvia & Comastri, Federico & Piccioni, Lorenzo & Volpi, Alessandra & Aston, Guy & Mazzoleni, Marco. 2004. Introducing the *la Repubblica* Corpus: A large, annotated, TEI(XML)-

[6] Per dettagli sugli avverbi focalizzanti, cfr. De Cesare (2019: 90-95).

compliant corpus of newspaper Italian. *Proceedings of LREC 2004*, 1771-1774. Lisbon: ELDA.

Bernardi, Raffaella & Bolognesi, Andrea & Seidenari, Corrado & Tamburini, Fabio. 2006. POS tagset design for Italian. In *Proceedings of the 5th International Conference on Language Resources and Evaluation*. Genova.

Bernini, Giuliano. 1995. Le profrasi. In Renzi, Lorenzo & Salvi, Giampaolo & Cardinaletti, Anna (a cura di), *Grande grammatica italiana di consultazione* 3, 175-222. Bologna: Il Mulino.

Chiari, Isabella. 2007. *Introduzione alla linguistica computazionale*. Bari: Laterza.

De Cesare, Anna-Maria. 2019. *Le parti invariabili del discorso*. Roma: Carocci.

D'Errico, Marianna & Grandi, Nicola & Paternesi Meloni, Serena & Tamburini, Fabio. 2016. Induzione di categorie grammaticali e lessicali. In Dedè, Francesco (a cura di), *Categorie grammaticali e classi di parole. Statuto e riflessi metalinguistici*, 115-137. Roma: Il Calamo.

Graffi, Giorgio. 1994. *Sintassi*. Bologna: Il Mulino.

Iacobini, Claudio & De Rosa, Aurelio & Schirato, Giovanna. 2017. Criteri e strategie di classificazione morfo-sintattica dei testi del corpus MIDIA. In D'Achille, Paolo & Grossmann, Maria (a cura di), *Per la storia della formazione delle parole in italiano. Un nuovo corpus in rete (MIDIA) e nuove prospettive di studio*, 33-51. Firenze: Cesati.

Kilgarriff, Adam & Baisa, Vít & Bušta, Jan & Jakubíček, Miloš & Kovář, Vojtěch & Michelfeit, Jan & Rychlý, Pavel & Vít Suchomel. 2014. The Sketch Engine: ten years on. *Lexicography* 1(1). 7-36.

Lesmo, Leonardo & Lombardo, Vincenzo & Bosco, Cristina. 2002. Treeback development. The TUT approach. In *Proceedings of ICON02*. Mumbay, India.

Monachini, Monica. 1995. ELM-IT: An Italian Incarnation of the EAGLES-TS. Definition of Lexicon Specification and Classification Guidelines. In *Technical report*. Pisa.

Prandi, Michele. 2007. Avverbi di collegamento e congiunzioni. In San Vicente, Félix (a cura di), *Partículas. Particelle. Estudios de lingüística contrastiva español e italiano*, 89-103. Bologna: CLUEB.

Salvi, Giampaolo. 2013. *Le parti del discorso*. Roma: Carocci.

Salvi, Giampaolo. 2014. La classificazione delle parti del discorso. *Italogramma* 8. 55-74.

Schmid, Helmut. 1994. Probabilistic Part-of-Speech Tagging Using Decision Trees. In *Proceedings of International Conference on New Methods in Language Processing*. Manchester, UK.

Schütze, Hinrich. 1995. Distributional Part-of-speech Tagging. In *Proceedings of 7th EACL*, 141-148. Dublin, Ireland.

Serianni, Luca. 2000. *Italiano*, con la collaborazione di A. Castelvecchi. Torino: Garzanti.

Tamburini, Fabio. 2000. Annotazione grammaticale e lemmatizzazione di corpora in italiano. In Rossini Favretti, Rema (a cura di), *Linguistica e informatica: multimedialità, corpora e percorsi di apprendimento*, 57-73. Roma: Bulzoni.

Tamburini, Fabio. 2016. A BiLSTM-CRF PoS-tagger for Italian tweets using morphological information. In *CLiC-it/EVALITA 2016*. Napoli.

Venturi, Giulia. 2009. Rassegna comparativa degli schemi di annotazione morfosintattica per la lingua italiana. In *Rapporto Tecnico TRIPLE*.

Appendice

Tabella 1 - *Corpora e schemi di annotazione a confronto*

	Nome corpus	*Annotazione*	*Tag/Code*	*Parte del discorso*
1.	CORIS-CODIS	PoS-tagset concepito seguendo le linee guida EAGLES (Monachini 1995)	• ADV • CONJ_C • CONJ_S	• Avverbi • Cong. Coord. • Cong. Subord.
2.	la Repubblica	Tree Tagger basato sul file di parametri di Marco Baroni (Baroni *et al*. 2004)	• ADV	• Avverbio
3.	Timestamped JSI web corpus Italian		• ADV:mente	• Avverbio in *-mente*
4.	COMPARE-IT		• CON	• Congiunzione
5.	Paisà			
6.	OPUS2 Italian			
7.	itTenTen16			
8.	itWAC (reduced)			
9.	MIDIA			
10.	Araneum Italicum Maius (2014)	Tree Tagger basato sul file di parametri di Achim Stein	• ADV • CON	• Avverbio • Congiunzione

Appendice 2 - *La parola* dopo/*CON in itTT*

1	lanazione.it	e mi sbatte a terra . </s><s> Duecento metri	dopo	si ferma per prestarmi soccorso . </s><s> Arriva u
		in CON CLI VER:fin PRE NOUN SENT DET:num NOUN	CON	CLI VER:fin PRE VER:infi:cli NOUN SENT VER:fin AI
2	virgilio.it	roblemi sono molto differenti . </s><s> Il maggiore	dopo	vari cambiamenti di scuola (diversi sistemi) è
		NOUN VER:fin ADV ADJ SENT ART ADJ	CON	ADJ NOUN PRE NOUN PUN ADJ NOUN PUN AUX:s
3	innomedimaria.i...	ili di nuovo al cielo </s><s> - otto giorni	dopo	ai discepoli con Tommaso </s><s> - alcuni gi
		t:fin PRE NOUN ARTPRE NOUN NOCAT DET:num NOUN	CON	ARTPRE NOUN PRE NPR NOCAT DET:indef N
4	innomedimaria.i...	i discepoli con Tommaso </s><s> - alcuni giorni	dopo	sul Lago di Tiberiade , miracolo della pesca (
		PRE NOUN PRE NPR NOCAT DET:indef NOUN	CON	ARTPRE NPR PRE NPR PUN NOUN ARTPRE NOUN P
5	amazingcomics.i...	arrello . </s><s> Nato in provincia di Venezia ,	dopo	essermi diplomato all ' Istituto Statale d' Arte di
		NOUN SENT NPR PRE NOUN PRE NPR PUN	CON	AUX:infi:cli VER:ppast NOUN PUN NPR NPR PRE NPR ART
6	chiesacattolica...	rato in attesa di processo , rilasciato sei mesi	dopo	per non aver commesso il fatto ; e quella
		past PRE NOUN PRE NOUN PUN VER:ppast DET:num NOUN	CON	PRE NEG AUX:infi VER:ppast ART NOUN PUN CON PRO:demo
7	chiesacattolica...	he non ha avuto nessun ' altra imputazione	dopo	di allora e non sta godendo di alcun beneficio
		CHE NEG AUX:fin VER:ppast DET:indef PUN DET:indef NOUN	CON	PRE ADV CON NEG VER:fin VER:geru PRE DET:indef NOUN
8	quipo.it	altro la scatola sul n ° 2 , quadrato	dopo	quadrato , fino ad arrivare al dieci . </s>
		PRO:indef ART NOUN ARTPRE ADJ NOCAT NUM PUN ADJ	CON	ADJ PUN PRE PRE VER:infi ARTPRE DET:num SENT
9	indire.it	e i suoi fratelli si recano al tempio e ,	dopo	averlo purificato , demoliscono l' altare dei sa
		ON ART DET:poss NOUN CLI VER:fin ARTPRE NOUN CON PUN	CON	AUX:infi:cli VER:ppast PUN VER:fin ART NOUN ARTPRE N
10	innomedimaria.i...	ifiuto di Nazaret viene narrato da Marco molto tempo	dopo	che l' attività pubblica di Gesù è iniziata .
		NOUN PRE NPR AUX:fin VER:ppast PRE NPR ADV NOUN	CON	CHE ART NOUN ADJ PRE NPR AUX:fin VER:ppast SENT
11	repubblica.it	mo paese che ha riconosciuto l' Ucraina il giorno	dopo	che questa s ' è proclamata indipendente d
		DJ NOUN CHE AUX:fin VER:ppast ART NPR ART NOUN	CON	CHE DET:demo ADJ PUN AUX:fin VER:ppast ADJ ART
12	dnet.it	no del Mali . </s><s> Ciononostante due anni	dopo	rprenderanno gli scontri . </s><s> Ad aggravare l
		L ARTPRE NPR SENT NPR DET:num NOUN	CON	VER:fin ART NOUN SENT PRE VER:infi AI
13	amyresource.it	a che non dipenderà da un Pc o altro computer	dopo	che abbiamo visto AInc ritrattare costantemente le su
		CHE NEG VER:fin PRE ART NPR CON DET:indef NOUN	CON	CHE VER:fin NOUN NPR VER:infi ADV:mente ART DET
14	sangiorgiohotel...	stando sul lato sinistro della strada . </s><s>	Dopo	circa 800mt si scorgerà la statua intitolata a Garibal
		VER:geru ARTPRE NOUN ADJ ARTPRE NOUN SENT	CON	ADV NPR CLI VER:fin ART NOUN VER:ppast PRE NPR
15	lacasadellezucc...	Giganti " Is Concias " svoltare a destra :	dopo	circa 100m al bivio prendere la via Monte Nied
		E NPR PUN NOUN NPR PUN VER:infi PRE NOUN PUN	CON	ADV ADJ ARTPRE NOUN VER:infi ART NOUN NPR NPR
16	niederwieser.it	er la macinazione della carne . </s><s> Ora ,	dopo	otre trent ' anni di attività nel settore , Tippe
		RE ART NOUN ARTPRE NOUN SENT ADV PUN	CON	ADV ADJ PUN NOUN PRE NOUN ARTPRE NOUN PUN NP

DOI: 10.17469/O2106SLI000023

IØRN KORZEN

Cosa ci rivelano i corpora sulla complessità testuale dell'italiano?

In questo lavoro discuto la nozione di complessità linguistica con particolare riguardo alla strutturazione testuale. Paragono l'italiano e il danese e come base empirica mi servo di tre corpora di testi paralleli di tipologie diverse: il *corpus Europarl* (discorsi argomentativi tenutisi al Parlamento europeo), il *corpus di Mr. Bean* (90 esposizioni narrative, scritte e orali, di due episodi di Mr. Bean prodotte da studenti universitari di Torino e di Copenaghen) e il *corpus SugarTexts* (testi espositivi tecnici sulla produzione di zucchero da barbabietola). In questi testi analizzo due fenomeni che si differenziano cross-linguisticamente in modo particolare: il numero di proposizioni per periodo e la loro testualizzazione finita vs. non-finita, ossia il grado della loro deverbalizzazione. Con riferimento a questi due fenomeni non vi è dubbio che l'italiano si manifesti come lingua assai più complessa del danese.

Parole chiave: complessità linguistica, strutturazione testuale, corpora paralleli, densità testuale, deverbalizzazione.

1. *Introduzione*

"L'italiano è una lingua complessa e difficile!". Sono più di 40 anni che sento regolarmente, in qualità di docente di italiano in Danimarca, tale lamentela da parte dei miei studenti, e colleghi docenti delle altre lingue romanze sentono lagnanze simili. Ma è davvero così? Alcune lingue o gruppi linguistici sono veramente più complessi di altri? E se sì, secondo quali parametri, a quali livelli linguistici e, inoltre, come verificare tale ipotesi?

La nozione di "complessità" è approdata alla linguistica da altre scienze quali fisica, antropologia e filosofia, ed è generalmente definita come la quantità di informazioni necessarie per comprendere e spiegare adeguatamente un sistema o un insieme di elementi (Merlini Barbaresi 2003: 24, 2005: 302). In linguistica il concetto di complessità è stato ampiamente indagato e discusso negli ultimi decenni, spes-

so con definizioni simili. Alcuni studiosi distinguono tra complessità "assoluta", ossia intrinseca al sistema linguistico stesso, e complessità "relativa", ossia vissuta soggettivamente dai parlanti e basata su specifici aspetti problematici per parlanti e/o apprendenti[1].

Nel suo approccio sociolinguistico, Moretti (2018: 40-42) distingue tra complessità "spontanea o primaria", legata alla creazione della lingua in questione, e complessità "secondaria", causata ad esempio dalla normativizzazione di varietà particolarmente formali, caratteristica molto evidente dell'italiano e confermata anche da Berruto (2012: 176) che – in accordo con i miei studenti – conclude: "Non converrebbe, in conclusione, sostenere che, più che essere l'italiano popolare una varietà linguistica semplificata, sia l'italiano standard e colto una varietà linguistica particolarmente complessa, elaborata, in un certo senso 'innaturale', grazie appunto alla sua precoce standardizzazione letteraria, aulica e elitaria...?".

Le pagine seguenti sono strutturate in questo modo: nella sezione 2 citerò altri punti di vista e considerazioni degli studiosi sul concetto di complessità, dopodiché, nelle sezioni 3-4, presenterò i corpora di testi paralleli sui quali baserò le mie analisi e osservazioni: il *corpus Europarl*, il *corpus di Mr. Bean* e il *corpus SugarTexts*. Paragonerò la lunghezza dei periodi italiani e danesi misurata come numero di parole e di proposizioni per periodo, e nella sezione 5 confronterò la testualizzazione (esplicita, implicita o nominalizzata) delle singole proposizioni dei tre corpora. Sulla base dei miei risultati tirerò alcune conclusioni nella sezione 6.

2. *Definizioni e prime indicazioni di diversa complessità italiana-danese*

Probabilmente è impossibile definire la complessità complessiva di una lingua in modo oggettivo e razionale (Deutscher 2009: 247); nella migliore delle ipotesi, essa può essere intesa come l'accumulo di diversi valori che insieme accrescono o diminuiscono la complessità di particolari dimensioni linguistiche. Va aggiunto che la complessità "relativa" e "learner-related" non è mai costante dato che dipende lar-

[1] Cfr. per esempio Masi (2003), Merlini Barbaresi (2004: sez. 1), Miestamo (2009: 81ff), Dahl (2009: 50-52) e Fiorentino (2009: 282ff).

gamente dalla distanza tra la L1 dell'apprendente e la L2 in questione[2]. Anche la nozione di complessità testuale risulta alquanto complessa: "More precisely, textual complexity turns out to be the result of the cumulative effects of the interaction among different variables that belong to all the levels of texture." (Masi 2003: 142).

Molti studiosi che propongono definizioni più specifiche adottano come punto di partenza nell'approccio generale alla complessità di un oggetto "the amount of information needed to recreate or specify it" (Dahl 2009: 50). McWhorter (2001: 125) definisce la complessità linguistica come "the degree of overt signalling of various phonetic, morphological, syntactic, and semantic distinctions beyond communicative necessity", e come esempi egli menziona marcatura di genere, multipli tempi passati, il congiuntivo e verbi nominalizzati. Nello stesso spirito, Fiorentino (2009: 282-286) cita, tra i fattori che contribuiscono alla complessità linguistica, un alto numero di sottotipi o di varianti alternative di un dato elemento o di una data funzione (per esempio morfologica), un alto numero di regole sintattiche e poca trasparenza nella relazione tra forma e funzione. Per una definizione simile, vedi Nichols (2009: 111-112).

In un'analisi delle differenze tra lingua parlata e scritta, Chafe (1985) introduce il concetto di "idea units", ossia "unità di pensiero", che contengono "all the information a speaker can handle in a single focus of consciousness" (Chafe 1985: 106). Nella lingua scritta tali unità di pensiero sono testualizzate come periodi, e chiare indicazioni di differenze di testualizzazione e di complessità, per esempio tra l'italiano e il danese, possono essere ottenute da corpora multilingui di testi paralleli, cioè testi autentici prodotti indipendentemente nelle lingue in questione, ma in situazioni equivalenti e per target e con contenuti equivalenti.

[2] Cfr. Deutscher (loc.cit), Bertuccelli (2003: 139), Maas (2009: 177), Nichols (2009: 120ff), Moretti (2018: 37), Korzen (2021). Sul tema dell'acquisizione della seconda lingua (SLA), cfr. anche Ellis (2016).

3. *Il corpus "Europarl"*

Un corpus frequentemente adoperato per paragoni linguistici è il *corpus Europarl,* e un paragone per esempio dei periodi di tutti i testi italiani e danesi L1 degli anni 1996-2010 porta ai seguenti risultati:

Tabella 1 - *Corpus Europarl: lunghezza dei periodi*

Corpus Europarl (anni 1996-2010)	*1. Parole*[3]	*2. Periodi*	*3. Parole per periodo*	
			Numeri medi	*Differenza*
Testi italiani	1.657.592	47.405	35,0	45,2 %
Testi danesi	546.425	22.668	24,1	

Come risulta dalla Tabella 1, misurati come numero medio di parole, i periodi italiani sono notevolmente più lunghi di quelli danesi, la differenza arrivando al 45,2 %[4]. Però di per sé, il semplice numero di parole per frase non è necessariamente indicativo di una complessità alta o bassa (Masi 2003: 141; Merlini Barbaresi 2004: sez. 2; Fiorentino 2009: 309; Korzen 2021: 22): bisogna capire a che cosa servono le molte "parole in più" nei testi italiani, se semplicemente a descrizioni più specificate e dettagliate oppure ad un numero più elevato di fatti e avvenimenti.

A tal fine, ho calcolato e paragonato il numero di proposizioni di tutti i periodi, ossia il numero di unità consistenti di un predicato e relativi argomenti, unità che possono essere testualizzate come frasi principali o subordinate, e se subordinate: esplicite, implicite o nominalizzate. Dato che tutti questi calcoli e paragoni sono stati condotti manualmente, qui mi sono servito di un sottocorpus *Europarl* di dimensioni più modeste, più precisamente di 70 testi in ogni lingua, scel-

[3] Le differenze di rappresentanza italiana e danese nel Parlamento europeo rendono i testi italiani degli anni in questione circa tre volte più numerosi dei testi danesi.

[4] Naturalmente, varie riserve vanno fatte nell'operare con calcoli basati sull'unità "parola (grafica)". Alcune differenze tipologiche portano ad un numero di parole più basso in danese (ad esempio l'articolo definito che in danese è spesso enclitico e i molti nomi composti danesi che corrispondono a costrutti nome + preposizione + nome in italiano). Altre differenze comportano invece un numero più basso in italiano (ad esempio verbo + pronome enclitico in italiano, inesistente in danese, e il fenomeno del pro-drop). Tuttavia, la maggior parte delle differenze interlinguistiche menzionate in questa tabella e in quelle seguenti sono di dimensioni che, a mio parere, consentono di utilizzarle come indicazioni di fondamentali differenze testuali.

ti in modo da rappresentare una varietà di riunioni parlamentari, una varietà di argomenti trattati e una varietà di autori; cfr. anche Korzen & Gylling (2017), dove ci eravamo serviti di 50 testi in ogni lingua.

Tabella 2 - *Sottocorpus Europarl: proposizioni e periodi*

Sottocorpus Europarl	*1. Testi*	*2. Parole*	*3. Proposizioni*	*4. Periodi*	*5. Proposizioni per periodo*	
					Numeri medi	*Differenza*
italiano	70	22.707	2.827	685	4,13	57,0 %
danese	70	22.705	2.758	1.049	2,63	

Questa tabella rivela un'altra cosa interessante: i due sottocorpora sono di dimensioni equivalenti quanto al numero di testi e di parole (colonne 1-2) e di dimensioni piuttosto simili quanto al numero di proposizioni (colonna 3). Invece, per quanto riguarda il numero di periodi (colonna 4) si osserva una grande differenza, differenza riscontrabile, di conseguenza, anche nel numero di proposizioni per periodo (colonna 5), dove essa supera perfino la differenza dei numeri di parole per periodo che avevamo visto in Tabella 1, colonna 3. Nei sottocorpora indagati, le parole "in più" nei testi italiani servono quindi a testualizzare nuovi eventi, avvenimenti o situazioni e in quel modo a rendere più concettualmente complessi i singoli periodi.

4. *Altri due corpora*

Come è noto, il *corpus Europarl* consiste di testi argomentativi, più precisamente dei discorsi politici tenutisi al Parlamento europeo (Koehn 2005, https://statmt.org/europarl/). Per assicurarmi che le differenze illustrate nelle Tabelle 1-2 non fossero limitate a tale particolare tipo e genere testuale, ho aggiunto alle mie analisi altri due corpora, anch'essi di dimensioni modeste, ma di tipi e generi diversi:

– The *Mr. Bean corpus*, testi narrativi: 90 esposizioni, scritte e orali, di due episodi di Mr. Bean ("The Library" e parte di "Merry Christmas Mr Bean") prodotte da 27 studenti dell'Università di Torino e da 18 studenti dell'Università di Copenaghen. Il corpus fu creato nel 1995 da un gruppo di docenti dell'Università di Copenaghen e della Copenhagen Business School, incluso il sottoscritto (Skytte et al. 1999, http://blog.cbs.dk/mrbean-korpus/), con la collaborazione

dell'Università di Torino e di Carla Bazzanella. Facemmo vedere i due episodi agli studenti chiedendo loro di riferire poi gli episodi oralmente e per iscritto nella propria madrelingua.

– Il corpus *SugarTexts* – Telling the *SugarStory* in diverse languages, (Smith 2009, http://www.sugartexts.dk/), che consiste di testi espositivi tecnici sulla produzione di zucchero da barbabietola in sei lingue diverse romanze e germaniche e nel russo. Il corpus focalizza particolarmente i processi effettuati in seguito all'arrivo delle barbabietole agli zuccherifici. I testi sono stati raccolti da dottorandi e laureandi della Copenhagen Business School in collaborazione con i loro tutor, fra cui il sottoscritto, e appartengono a generi svariati come siti web e dépliant di aziende produttrici di zucchero, enciclopedie, manuali, guide di vario tipo e libri informativi per diversi target. Per le mie analisi mi sono servito di 15 testi italiani e di 15 testi danesi.

Le dimensioni di questi due corpora sono citate nella Tabella 3.

Tabella 3 - *Corpus di Mr. Bean e SugarTexts: numero di parole e di proposizioni*

Corpus	*Testi*	*1. Parole*	*2. Proposizioni*	*3. Periodi*	*4. Parole per periodo*	*5. Proposizioni per periodo*	
						Num. medi	*Differ.*
Mr. Bean	*ital.*	7.278	1.388	319	22,8	4,35	32,6 %
	dan.	7.262	1.189	363	20,0	3,28	
Sugar Texts	*ital.*	4.819	898	194	24,8	4,63	82,3 %
	dan.	4.851	832	327	14,8	2,54	

Anche qui, i testi italiani e quelli danesi sono, in entrambi i corpora, grosso modo di dimensioni equivalenti quanto al numero di parole e di proposizioni (colonne 1-2). Invece i periodi italiani sono più lunghi quanto al numero di parole per periodo (colonna 4), e soprattutto quanto al numero di proposizioni per periodo (colonna 5), le differenze essendo particolarmente evidenti nei testi espositivi.

5. *La testualizzazione delle singole proposizioni*

Non vi è dubbio che un alto numero di proposizioni nello stesso periodo aumenti la complessità e la compattezza del periodo, ma un al-

tro fattore importante è la testualizzazione delle stesse proposizioni. Fra gli elementi che accrescono la complessità delle "unità di pensiero" (cioè dei periodi nella lingua scritta), Chafe (1985: 108-117) elenca frasi subordinate e apposizioni, participi presenti e passati e nominalizzazioni. Di tali testualizzazioni, i verbi impliciti e nominalizzati – ossia le testualizzazioni "deverbalizzate" (Korzen 2009, 2018; Korzen & Gylling 2017) – sono particolarmente interessanti perché aumentano anche la cosiddetta "densità testuale" e rendono i testi meno "vincolanti" nella terminologia di Sabatini (1999).

La densità testuale è definita come la relazione tra la quantità di materiale linguistico di una sequenza testuale e le informazioni che tale sequenza intende esprimere (Fabricius-Hansen 1996, 1998, 1999; Jansen 2003; Hansen-Schirra *et al.* 2007; Korzen & Gylling 2017), e le forme verbali non finite trasmettono un contenuto testuale con meno materiale linguistico rispetto ad una struttura con verbo finito, la quale in genere contiene una congiunzione e possibilmente un verbo ausiliario finito e un soggetto (Korzen 2014, 2015):

(1) a. *Arrivato tardi*, Luca ha perso l'inizio del film.
b. *Dato che Luca era arrivato tardi*, ha perso l'inizio del film.

(2) a. *Arrivando tardi*, perderai l'inizio del film.
b. *Se tu arrivi tardi*, perderai l'inizio del film.

Allo stesso tempo, le strutture non finite sono meno "vincolanti" perché lasciano un'interpretazione semantica precisa al ricevente (Sabatini 1999), ad esempio causa o condizione come illustrato negli esempi (1a)-(2a). In (3a) la nominalizzazione esprime un contenuto temporale:

(3) a. *All'arrivo di Luca*, siamo andati al cinema.
b. *Quando Luca è arrivato*, siamo andati al cinema.

Per queste ragioni tali strutture sono fortemente sconsigliate per esempio nei manuali e guide alla redazione dei documenti amministrativi e legislativi di tutti i livelli: comunale, provinciale, regionale, statale ed europeo (Korzen 2015). La maggiore densità linguistica e i meno vincoli interpretativi richiedono un maggiore impegno interpretativo da parte del ricevente (Korzen 2021), insomma comportano una maggiore complessità[5].

[5] Cfr. anche Merlini Barbaresi (2003: 40ff, 2004: sezione 6) per il ruolo delle forme verbali non finite per la complessità testuale nelle ricette di cucina.

In un paragone cross-linguistico basato sui tre corpora indagati le testualizzazioni implicite e nominalizzate puntano su differenze notevoli:

Tabella 4 - *Corpus Europarl, Mr. Bean, SugarTexts: testualizzazione delle proposizioni*

Corpus	*Testi*	*1. Frasi principali*	*2. Frasi subordinate esplicite*	*3. Frasi subordinate implicite*	*4. Nominalizzazioni*
Europarl	*italiani*	21,3 %	37,6 %	22,5 %	18,6 %
	danesi	28,1 %	49,2 %	12,4 %	10,3 %
Mr. Bean	*italiani*	40,9 %	19,5 %	33,7 %	5,9 %
	danesi	54,6 %	27,2 %	15,7 %	2,4 %
SugarTexts	*italiani*	29,8 %	17,8 %	26,7 %	25,6 %
	danesi	50,1 %	29,4 %	8,9 %	11,5 %

Nelle colonne 3-4 si osserva come, in tutti e tre i corpora, le testualizzazioni implicite e nominalizzate siano assai più frequenti nei testi italiani che nei testi danesi. Invece in tutti i testi danesi le percentuali di testualizzazioni esplicite, principali o subordinate (colonne 1-2), superano quelle italiane. Buoni esempi di testualizzazioni dense e – almeno per un danese – complesse perché consistono di periodi lunghi, compatti e pieni di verbi impliciti e nominalizzati, sono citati in (4)-(5) che provengono da due testi italiani del *corpus Europarl*.

(4) Accanto al deciso sostegno ad un approccio microeconomico, destinato ad incoraggiare i paesi più poveri ad investire nel loro stesso avvenire lo sviluppo del microcredito, l'Unione auspica il mantenimento delle preferenze commerciali con i paesi più poveri e più economicamente vulnerabili. (ep-98-04-01.txt:39, deputato autore: Amadeo Amadeo)[6].

(5) [Q]uanto più noi riconosciamo l'autorevolezza degli organi monetari in questione – ed io mi associo a questo riconoscimento, non senza rilevare con convinzione l'esigenza di un governo politico dell'economia, anche a livello europeo – tanto più occorre pretendere che vengano dati esempi limpidi di efficienza, di moderazione salariale e di trasparenza. (ep-01-01-16.txt:39, deputato autore: Gianni Pittella).

[6] Le sigle riferentisi ai discorsi *Europarl* ("ep") indicano anno-mese-giorno, seguiti dal numero del discorso del giorno in questione ("txt").

Invece gli esempi danesi in (6)-(7), dello stesso corpus, illustrano testualizzazioni completamente diverse, molto più verbali e semplici: composte di periodi brevi con frasi esplicite e un'unica subordinazione del primo livello.

(6) Der er mange lobbyister i Europa-Parlamentet. De indgår som en naturlig del af vores arbejde og bidrager med oplysninger og synspunkter. De kan nok ikke undværes, men vi skal have regler for deres aktiviteter. Ford-betænkningen er et godt bud på nogle regler, der kan gennemføres. (ep-96-07-16.txt:258, deputato autore: Freddy Blak).
'Ci sono molti lobbisti al Parlamento europeo. Costituiscono una parte naturale del nostro lavoro e contribuiscono con informazioni e opinioni. Probabilmente non se ne può fare a meno, ma dobbiamo avere regole per le loro attività. La relazione Ford è una buona proposta per alcune regole che possono essere implementate.'

(7) I øvrigt løser direktivforslagene ikke forbrugernes problemer med handel over grænserne. Der er store afstande mellem forbrugere og sælgere i forskellige lande, både geografisk og sprogligt. Derfor kan man spørge sig selv, om harmonisering overhovedet har noget formål og nogen virkning på dette område. Jeg ønsker nemlig ikke harmonisering for harmoniseringens skyld. (ep-98-03-09.txt:60, deputata autrice: Ulla Sandbæk).
'Inoltre le direttive proposte non risolvono i problemi dei consumatori con il commercio attraverso le frontiere. Ci sono grandi distanze tra consumatori e venditori in paesi diversi, sia geograficamente che linguisticamente. Perciò ci si può chiedere se l'armonizzazione abbia affatto uno scopo e un effetto in questo settore. Non desidero l'armonizzazione per il bene dell'armonizzazione.'[7]

[7] Queste due traduzioni di (6)-(7) sono mie e molto fedeli ai testi fonte. Le traduzioni ufficiali del Parlamento Europeo, scaricabili dall'Europarl Parallel Corpus 1996-2011, sono le seguenti:
(6') 'Ci sono molti lobbisti al Parlamento Europeo. Costituiscono ormai un elemento naturale del nostro lavoro al quale contribuiscono con informazioni e punti di vista. Non se ne può fare a meno, ma occorre avere regole che disciplinino le loro attività. La relazione Ford rappresenta un ottimo spunto per regole possibili.'
(7') 'Inoltre la proposta di direttiva non risolve il problema del consumatore rispetto agli acquisti effettuati all'estero. La distanza tra venditore e consumatore di paesi diversi è enorme, sia geograficamente che linguisticamente. Pertanto è lecito chiedersi

6. *Conclusioni*

La complessità linguistica è un fenomeno estremamente intricato, e come molti studiosi hanno affermato, probabilmente è impossibile determinare e definire una nozione quale la "complessità complessiva" di una lingua; nella migliore delle ipotesi, essa può essere considerata come l'accumulo di una serie di valori diversi. Nelle pagine precedenti ho messo a fuoco due tali valori, cioè il numero di proposizioni per periodo e il grado della loro deverbalizzazione, due fenomeni che si dimostrano particolarmente differenti nelle due lingue qui indagate, l'italiano e il danese, e particolarmente problematici per un danese che voglia perfezionare il suo italiano.

Anche se è possibile trovare testualizzazioni italiane e danesi strutturalmente diverse da quelle dimostrate rispettivamente in (4)-(5) e in (6)-(7), questi esempi costituiscono buone illustrazioni delle differenze cross-linguistiche citate nelle Tabelle 1-4 sopra, e gli esempi (4)-(5) confermano la particolare complessità strutturale che un apprendente danese deve affrontare nel suo incontro con l'italiano. Viceversa, la forma della testualizzazione danese vista in (6)-(7) può apparire banale e semplicistica ad un italiano.

Naturalmente l'indagine dei due fenomeni trattati in queste pagine non può per niente esaurire la problematica della complessità linguistica, ma non vi è dubbio che le citate differenze di compattezza e di densità testuale, documentate in modo palese dai corpora, siano una buona ragione per cui un danese possa trovare l'italiano "una lingua complessa".

Riferimenti bibliografici

Berruto, Gaetano. 2012. L'italiano popolare e la semplificazione linguistica. In Berruto, Gaetano, *Saggi di sociolinguistica e linguistica* a cura di Giuliano Bernini & Bruno Moretti & Stephan Schmid & Tullio Telmon, con la collaborazione di Gloria Scarano, 141–181. Alessandria: Edizioni dell'Orso. [Prima pubblicato in *Vox Romanica* 42 (1983). 38–79].

se l'armonizzazione in questo campo possa avere un senso e produrre degli effetti. Personalmente non sono a favore dell'armonizzazione come valore assoluto.'

Si noti come entrambe queste traduzioni siano più strutturalmente "eleganti" dei testi fonte e più sintatticamente complessa: in (6') con una subordinazione in più e in (7') con un verbo implicito (*effettuati*) in più.

Bertuccelli, Marcella. 2003. Cognitive complexity and the lexicon. In Merlini Barbaresi, Lavinia (a cura di), *Complexity in Language and Text,* 67–115. Pisa: Edizioni Plus.

Chafe, Wallace L. 1985. Linguistic differences produced by differences between speaking and writing. In Olson, David R. & Torrance, Nancy & Hildyard, Angela (eds.), *Literacy, Language and Learning. The Nature and Consequences of Reading and Writing,* 105–123. Cambridge: Cambridge University Press.

Dahl, Östen. 2009. Testing the assumption of complexity invariance: the case of Elfdalian and Swedish. In Sampson, Geoffrey & Gil, David & Trudgill, Peter (eds.), *Language Complexity as an Evolving Variable,* 50–63. Oxford: Oxford University Press.

Deutscher, Guy. 2009. "Overall complexity": a wild goose chase? In Sampson, Geoffrey & Gil, David & Trudgill, Peter (eds.), *Language Complexity as an Evolving Variable,* 243–251. Oxford: Oxford University Press.

Ellis, Nick C. 2016. Salience, cognition, language complexity, and complex adaptive systems. *Studies in Second Language Acquisition* 38. 341–351. doi:10.1017/S027226311600005X

Fabricius-Hansen, Cathrine. 1996. Informational density – a problem for translation and translation theory. *Linguistics* 34. 521–565.

Fabricius-Hansen, Cathrine. 1998. Information density and translation, with special reference to German – Norwegian – English. In Johansson, Stig & Oksefjell, Signe (eds.), *Corpora and Cross-linguistic Research: Theory, Method, and Case Studies,* 197–234. Amsterdam: Rodopi.

Fabricius-Hansen, Cathrine. 1999. Information packaging and translation. Aspects of translational sentence splitting (German – English/ Norwegian). In Doherty, Monika (ed.), *Sprachspezifische Aspekte der Informationsverteilung,* 175–213. Berlin: Akademie-Verlag.

Fiorentino, Giuliana. 2009. Complessità linguistica e variazione sintattica. *Studi Italiani di Linguistica Teorica e Applicata* XXXVIII(2). 281–312.

Hansen-Schirra, Silvia & Neumann, Stella & Steiner, Erich. 2007. Cohesive explicitness and explicitation in an English-German translation corpus. *Languages in Contrast* 7(2). 241–265.

Jansen, Hanne. 2003. *Densità informativa: Tre parametri linguistico-testuali. Uno studio contrastivo inter- ed intralinguistico*. Copenhagen: Museum Tusculanum Press.

Koehn, Philipp. 2005. Europarl: A parallel corpus for statistical machine translation. *Conference Proceedings: The Tenth Machine Translation Summit,* 79–86. Thailand: Phuket.

Korzen, Iørn. 2009. Struttura testuale e anafora evolutiva: tipologia romanza e tipologia germanica. In Korzen, Iørn & Lavinio, Cristina (a cura di), *Lingue, Culture e Testi Istituzionali*, 33–60. Firenze: Franco Cesati.

Korzen, Iørn. 2014. Struttura testuale e anafora nella traduzione del discorso politico: un'indagine tipologico-comparativa. In Garavelli, Enrico & Suomela-Härmä, Elina (a cura di), *Dal manoscritto al web: canali e modalità di trasmissione dell'Italiano. Tecniche, materiali e usi nella storia della lingua: Atti del XII Congresso della Società Internazionale di Linguistica e Filologia Italiana (SILFI)*, 391–400. Firenze: Franco Cesati.

Korzen, Iørn. 2015. Frasi complesse e complessità frasale: il discorso politico in un'ottica tipologico-comparativa. In Bruno, Carla & Casini, Simone & Gallina, Francesca & Raymond Siebetcheu (a cura di), *Plurilinguismo/ Sintassi. Atti del XLVI Congresso Internazionale della Società di Linguistica Italiana (SLI)*, 625–642. Roma: Bulzoni.

Korzen, Iørn. 2018. L'italiano: una lingua esocentrica. Osservazioni lessicali e testuali in un'ottica tipologico-comparativa. In Korzen, Iørn (a cura di), *La Linguistica Italiana nei Paesi Nordici. Studi Italiani di Linguistica Teorica e Applicata*, XLVII(1). 15–36.

Korzen, Iørn. 2021. Are some languages more complex than others? On text complexity and how to measure it. In Gargiulo, Marco & Haukås, Åsta & Korzen, Iørn (eds.), *When language typology meets multilingualism. From languages to uses and people. Globe. A Journal of Language, Culture and Communication* 12. 18–31.

Korzen, Iørn & Gylling, Morten. 2017. Text structure in a contrastive and translational perspective: On information density and clause linkage in Italian and Danish. In Czulo, Oliver & Hansen-Schirra, Silvia (eds.), *Crossroads between Contrastive Linguistics, Translation Studies and Machine Translation*, 31–64. Berlin: Language Science Press. https://zenodo.org/record/1019687.

Maas, Utz. 2009. Orality versus literacy as a dimension of complexity. In Sampson, Geoffrey & Gil, David & Trudgill, Peter (eds.), *Language Complexity as an Evolving Variable,* 164–177. Oxford: Oxford University Press.

Masi, Silvia. 2003. The literature on complexity. In Merlini Barbaresi, Lavinia (a cura di), *Complexity in Language and Text,* 117–145. Pisa: Edizioni Plus.

McWhorter, John H. 2001. The world's simplest grammars are creole grammars. *Linguistic Typology* 5. 125–166.

Merlini Barbaresi, Lavinia. 2003. Towards a theory of text complexity. In Merlini Barbaresi, Lavinia (a cura di), *Complexity in Language and Text,* 23–66. Pisa: Edizioni Plus.

Merlini Barbaresi, Lavinia. 2004. Levels of text complexity. In van Sterkenburg, Piet (ed.), *Linguistics Today – Facing a Greater Challenge*. Amsterdam/Philadelphia: John Benjamins, CD-Rom.

Merlini Barbaresi, Lavinia. 2005. Il discorso economico/argomentativo: marcatezza e complessità della previsione. In Leandro Schena, Chiara Preite & Vecchiato, Sara (a cura di), *Gli insegnamenti linguistici nel nuovo ordinamento: Lauree triennali e specialistiche dell'area economico-giuridica,* 301–324. Milano: Egea.

Miestamo, Matti. 2009. Implicational hierarchies and grammatical complexity. In Sampson, Geoffrey & Gil, David & Trudgill, Peter (eds.), *Language Complexity as an Evolving Variable,* 80–97. Oxford: Oxford University Press.

Moretti, Bruno. 2018. Che cosa ha da dire la sociolinguistica sul tema della complessità delle lingue. *Rivista Italiana di Dialettologia* 42. 35–52.

Nichols, Johanna. 2009. Linguistic complexity: a comprehensive definition and survey. In Sampson, Geoffrey & Gil, David & Trudgill, Peter (eds.), *Language Complexity as an Evolving Variable,* 110–125. Oxford: Oxford University Press.

Sabatini, Francesco. 1999. *"Rigidità-esplicitezza"* vs *"elasticità-implicitezza":* possibili parametri massimi per una tipologia dei testi. In Skytte, Gunver & Sabatini, Francesco (a cura di), *Linguistica testuale comparativa,* 141–172. Copenhagen: Museum Tusculanum Press.

Skytte, Gunver & Korzen, Iørn & Polito Paola & Strudsholm, Erling (a cura di). 1999. *Tekststrukturering på italiensk og dansk. Resultater af en komparativ undersøgelse / Strutturazione testuale in italiano e danese. Risultati di una indagine comparativa*. Copenhagen: Museum Tusculanum Press.

Smith, Viktor. 2009. Telling the SugarStory in seven Indo-European languages. What may and what must be conveyed? In Korzen, Iørn & Lavinio, Cristina (a cura di), *Lingue, culture e testi istituzionali,* 61–76. Firenze: Franco Cesati.

DOI: 10.17469/O2106SLI000024

MARIAFRANCESCA GIULIANI

Sulla diatopicità del repertorio lessicale degli antichi testi italiani

È possibile utilizzare un corpus diacronico tipologicamente eterogeneo e linguisticamente sfaccettato (anche se fatalmente sproporzionato nella rappresentazione del multilinguismo soggiacente) come il *Corpus TLIO* per verificare il ruolo della variazione diatopica nell'assetto complessivo del repertorio lessicale o si deve assumere come dato strutturale la tendenziale omogeneità linguistica di tale risorsa? L'articolo discute del peso della diatopicità nel lessico degli antichi testi di area italiana presenti nel corpus utilizzando valutazioni di ordine qualitativo e quantitativo.

Parole chiave: Linguistica dei corpora, Variazione linguistica e lessicale, Storia dell'italiano scritto, Lessicografia, Lessicologia.

1. *Introduzione*

Oggetto di questo articolo è una riflessione sulle condizioni di variazione che interessano il lessico nel corpus storico che per eccellenza rappresenta le antiche varietà italiane: mi riferisco naturalmente al *Corpus TLIO*: il corpus di riferimento del *Tesoro della Lingua Italiana delle Origini* (TLIO).

Si tratta, com'è noto, di un corpus "plurilingue", virtualmente rappresentativo della situazione linguistica a monte dei più antichi testi di area italiana elaborati entro la fine del sec. XIV, in assenza, dunque, di una norma linguistica unitaria. La raccolta è ben più ampia rispetto alla selezione testuale dell'aureo Trecento fiorentino delimitata dagli accademici della Crusca per legittimare una tradizione lessicale italiana destinata a essere accolta dalla comunità dei letterati (cfr. Giuliani 2019: 103-19): quale carattere differenziale può emergere dunque dall'esplorazione lessicale del *Corpus TLIO?*

1.1 Un repertorio lessicale omogeneo?

Converrà discutere da subito di convergenze con il corpus di testi su cui si fonda il *Vocabolario della Crusca*: nel quadro del *Corpus TLIO* la componente dei testi toscani è significativa e preponderante dal punto di vista quantitativo: a gennaio 2022, su un totale di 2991 testi di diversa area, 1589 testi, ovvero più della metà, sono toscani.

Il peso quantitativo di questa componente resta rilevante ad ogni aggiornamento del corpus nonostante una delle linee programmatiche seguita dall'OVI per le nuove inclusioni valorizzi l'appartenenza dei nuovi testi ad aree linguistiche scarsamente documentate[1].

Ma prescindendo dalla sproporzione quantitativa tra documentazione toscana e non toscana – un dato che ostacola di per sé una rappresentazione linguistica congruente della pluralità originaria –, è doveroso notare che la possibilità di accedere a tradizioni di lingua locale sembra disattesa dalla *facies* stessa della documentazione. Tomasin (2019: 158) evidenzia la tendenziale omogeneità della lingua dei testi antichi, soprattutto nel versante sintattico e lessicale, un'omogeneità che emerge nel contrasto tra lingua antica e moderna, legittimando l'attribuzione dell'etichetta "italiano antico" al complesso degli antichi volgari italiani.

La lingua degli antichi testi italiani in questi termini fa *sistema*: Burgassi & Guadagnini (2017) sperimentano un punto di vista complessivo e "supra-testuale" per ricostruire la struttura del vocabolario italiano antico rappresentato dal *Corpus OVI* (un corpus che nel 2017 raccoglieva le premesse dell'attuale *Corpus TLIO*) secondo un modello centro / periferia (cfr. pp. 13-17). La posizione di ogni elemento lessicale partecipe del repertorio lessicale degli antichi testi italiani sarebbe definibile in termini relativi, ovvero determinandone la collocazione in serie con altri elementi sulla base dell'uso e delle caratteristiche di attestazione. In questo quadro «la 'periferia' comprende [...] i vocaboli rari e fortemente caratterizzati dal punto di vista diatopico, diastratico o stilistico». (ib.: 14), tuttavia, nel complesso, anche in virtù della sproporzione quantitativa prima menzionata, la variazione diatopica non avrebbe un'incidenza significativa nel determinare l'assetto del lessico antico.

Muovendo, diversamente, da una valorizzazione della pluralità dei sistemi linguistici riflessi dal *Corpus TLIO* (cfr. Barbato 2019: 244),

[1] Rimando in proposito ai *Criteri per l'aggiornamento del Corpus TLIO e del Corpus OVI* < http://www.ovi.cnr.it/files/Criteriaggiornamentocorpus-1.pdf>.

in questo articolo cercherò di indagare quale diatopicità[2] possa essere colta, sul fronte del lessico, per il tramite del corpus e delle risorse di gestione allestite dall'OVI, discutendo parimenti di modelli descrittivi e di strategie di accertamento utilizzabili per sondare la rappresentazione di *differenze* legate a circuiti locali.

2. *Un'"architettura" per descrivere la variazione*

Una lingua storica esposta a *differenze* diatopiche, diafasiche e diastratiche è, secondo Coseriu (1967), un sistema di sistemi, un complesso di tradizioni descrivibile ricorrendo al paradigma dell'architettura variazionale: nell'architettura di una lingua storica la differenza è diversità non incasellata entro opposizioni funzionali; queste ultime appartengono alla struttura di una lingua omogenea e dunque funzionale, ovvero a una "tecnica del discorso".

Un'architettura che contempera opzioni alternative e diverse è un ottimo modello descrittivo per la pluralità linguistica convergente rappresentata dal *Corpus TLIO*. Il modello è ulteriormente raffinato dall'idea che le dimensioni della variazione non siano mai totalmente indipendenti, ma al contrario reciprocamente legate. Nella formulazione teorica di Coseriu (1981: 16), la relazione tra dialetto, livello e stile di lingua è orientata e precisabile in termini gerarchici e inclusivi:

Diatopia → Diastratia → Diafasia

Il fattore diatopico, agisce all'interno del fattore diastratico che, a sua volta, può essere mediato dal fattore diafasico: dunque la dimensione diafasica filtra e veicola elementi diastraticamente e diatopicamente marcati. Dovremo precisare che la dimensione diafasica è totalmente riassorbita nel *Corpus TLIO* dalla dimensione diamesica: la scrittura (il sistema dei generi scritturali) filtra e convoglia ogni evidenza di variazione. Ne consegue che ogni opzione diatopicamente marcata lungi dall'essere solo partecipe della variazione primaria, sia legata a una consuetudine comunicativa, a un progetto testuale e anche a un

[2] Termini come *diatopicità* e *diatopismo* sono poco usati nel quadro della ricerca italiana, ma ricorrono frequentemente nell'approccio francese e iberico allo studio della variazione linguistica nello spazio. Uso entrambi i termini con l'intento di valorizzare una variazione che prescinde dalla opposizione a una norma linguistica di riferimento.

intento di caratterizzazione: in tal senso può aver spazio anche nell'uso di scriventi e autori non legati a uno specifico contesto geografico.

2.1 Una varietà e una comunità di scriventi

«Messer l'amiraglio, come ti piace, da parte del tuo Comune da Sorrenti ilocati *quissi* palombola, e stipati *quissi* agostari per uno taglio di calze...»: nel riportare il dialogo tra i sorrentini ribelli e Carlo d'Angiò (confuso per l'Ammiraglio Ruggeri di Lauria), il cronista toscano Giovanni Villani (*Nuova Cronica*, libro VIII, cap. XCIII) dà spazio a una caratterizzazione linguistica sostenuta soprattutto dal dimostrativo centro-merid. *quisso* 'codesto' (cfr. TLIO s.v.); l'editore annota, peraltro, che in prima redazione il dialogo aveva una caratterizzazione dialettale ancora più spiccata, valorizzata, ad esempio, da *chiace* in luogo di *piace* (cfr. *Nuova Cronica*, ed. Porta 1990-91, vol. I: 554).

Il diatopismo occasionale, volto evidentemente a dettagliare l'intento realistico perseguito nella rappresentazione di un ambiente, si distingue evidentemente dal prestito di per sé volatile ma capace di far serie con l'ibridismo che può caratterizzare stabilmente, per un certo periodo, la comunicazione di un gruppo di scriventi. È indicativo il caso dei gallicismi effimeri riscontrabili in maniera massiccia nei libri di conto e nella produzione epistolografica dei mercanti e banchieri senesi attivi nelle piazze francesi, fiamminghe ed anche inglesi a partire dal primo trentennio del sec. XIII. La casistica è stata esaminata in tempi recenti da Cella (2010): il TLIO dispone della possibilità di isolare questa serie di prestiti occasionali nell'insieme delle voci documentate solo nei testi di una varietà specifica, nel nostro caso il senese, marcate da un'annotazione contenuta nel modulo della voce che schematizza la distribuzione geolinguistica di un lemma[3]. Sono "att. solo in testi sen." alcuni termini votati all'univocità referenziale come *busciello* 'unità di misura per aridi' (fr. ant. *boissiel*, *boissel*), *entrea* 'tassa d'ingresso alle fiere', 'apertura di una fiera' (fr. ant. *entrèe*), *famma* 'moglie' (fr. *femme*), *talametiere* 'fornaio' (fr. ant. *talametier*), destinati al decadimento col venir meno delle condizioni comunicative che ne

[3] La ricerca selettiva delle voci documentate solo nei testi di una varietà specifica sarà accessibile agli utenti del TLIO quando sarà pienamente operativa una versione evoluta del sistema di gestione e consultazione del database lessicografico.

sostennero l'adozione nei testi della comunità di individui operativa tra Siena e i mercati esteri.

In più di un caso, dunque, la geografia dei contatti dovrà essere considerata prioritaria rispetto alla geografia dei contesti linguistici in cui trovano elaborazione i testi.

2.2 Contatti e trafile: aree, testi e testimoni

L'incidenza di fattori variazionali di natura diversa va considerata anche vagliando il lessico di testi provenienti da aree che sono state esposte a contatti linguistici e culturali ben documentati.

Se è noto che la stratigrafia del lessico dei testi siciliani trecenteschi include significative tracce dell'interferenza di varietà gallo-italiche interagenti con le varietà isolane e il dato è per certi aspetti focalizzato dall'attestazione ristretta di alcuni lemmi del *Corpus TLIO* tra testi siciliani e testi settentrionali (v. ad es. le voci TLIO *badagliare, bresca 1, gaida*), è vero anche che il lessico di alcune testimonianze chiama in causa convergenze con il versante romanzo nord-occidentale che sollecitano interpretazioni singolari.

Riccardo Ambrosini (1977) che ha approntato un esame stratigrafico (ovvero orientato alla classificazione dei componenti storico-genetici) del lessico dei testi siciliani in prosa del '300, con l'intento «di dare un quadro, per così dire dialettico delle condizioni linguistiche della Sicilia e dei vari strati che le hanno caratterizzate, dalla latinizzazione all'età dei prestiti spagnoli» (ib.: 17), aveva già ben chiara la *facies* linguistica peculiare del *Valeriu Maximu* di Accursio di Cremona, volgarizzamento siciliano dei *Factorum et dictorum memorabilium libri* consultato nell'edizione di Ugolini (1967). Musso (2013) rileva nel testo, edito da Ugolini sulla base del testimone trecentesco Madrid, Biblioteca Nacional de España 8883, molteplici tratti linguistici (grafico-fonetici e morfologici oltre che lessicali) che sembrerebbero rimandare a un modello traduttivo catalano.

In questa sede ci limitiamo a segnalare la presenza nel testimone di lemmi che hanno riscontro solo in testi e varietà del versante nord-occidentale della Romania, lemmi che possono rinviare a una microstoria di contatti linguistici e culturali che merita probabilmente un supplemento d'indagine (cfr. anche Vaccaro 2019: 72): v. tra i molteplici esempi l'aggettivo *subreru* 'superiore' (privo di riscontri nel repertorio dei dialetti siciliani) che affianca le numerose attestazioni di *sobrer*

nelle poesie dell'Anonimo Genovese[4] richiamando, nel contempo, il prov.a. *sobrier* e il cat. *sobrer* 'più alto, superiore in forza o potenza' (DELCat VII,974, s.v. *sobre*).

3. *Differenze diatopiche: percorsi valutativi*

La casistica fin qui introdotta non intende escludere che sia possibile accertare il legame di una forma lessicale registrata dai testi (e dunque presente nel corpus) con uno specifico contesto storico-geografico.

Il riscontro dialettico tra la documentazione storica di un punto o di un'area e la documentazione dialettale corrispondente è naturalmente un efficace strumento diagnostico per validare la diatopicità di una forma testuale, e può sostenere, in alcuni casi, anche la formulazione di inferenze di ordine storico-linguistico. La documentazione TLIO per *badagliare* 'spalancare la bocca emettendo un sospiro, sbadigliare', tuttora rappresentativa della totalità dei testi che nel corpus documentano la voce, contiene esempi tratti da due diversi testi siciliani (Accurso di Cremona, 1321/37 e Senisio, *Declarus*, 1348) e da un testo ligure (*Sam Gregorio in vorgà*, XIV sm.): la documentazione storica è ben integrata nell'articolo BATACULARE del LEI che, pur in assenza dell'antico riferimento ligure, rimarca in maniera eloquente, con i riferimenti dialettali, la distribuzione del tipo lessicale in un'area settentrionale che corre dalla Liguria alla Lombardia (con sconfinamenti ticinesi) fino all'Emilia e trova un'appendice nella varietà gallo-italica di Trecchina, in Basilicata, e nel siciliano: nel commento si osserva che «le voci siciliane (innovative rispetto alle forme meridionali di HALĀRE) paiono essere gallo-italiche, irradiate poi nell'isola» (ib.: 229).

Non abbiamo testimonianze antiche per il tipo *alàre*, ampiamente attestato nel meridione continentale dalla Campania e dall'Abruzzo fino alla Calabria centro-meridionale (cfr. AIS c. 170), ma disponiamo, per il tramite del TLIO e del *Corpus TLIO*, di una documentazione toscana abbastanza ricca per *sbadigliare*, adoperato, ad esempio, nel volgarizzamento pisano del Cavalca dei *Dialogi* di Gregorio

[4] Qui e di seguito per le abbreviazioni relativi ai testi del corpus non chiarite nella bibliografia si rimanda alla *Bibliografia dei Testi Volgari* citati dal TLIO: <http://pluto.ovi.cnr.it/btv>.

Magno, da cui dipende anche una versione ligure (*Dialogo de Sam Gregorio composito in vorgà*); è nel corpus, inoltre, anche il più antico volgarizzamento del testo tardo-latino redatto in siciliano dal messinese Giovanni Campulu[5].

Allineo di seguito il dettato dell'originale tardo-latino alle tre traduzioni:

(1) *oscitavit, oculos aperuit*
Sbadigliò gli occhi aprì
'sbadigliò e aprì gli occhi'
(Greg., *Dial.*, l. 3, cap. 17)

(2) *sbadigliò e aperse gli occhi*
(Cavalca, *Dialogo S. Greg.*, a. 1330 (pis.): 167.15)

(3) *baglià e averse li ogli*
(*Sam Gregorio in vorgà*, XIV sm. (lig.): 186.31)

(4) *acconmenczau a flatare et aperire li ochi*
(GiovanniCampulu, 1302/37 (mess.): 101.28)

Notiamo tre distinte opzioni lessicali (*sbadigliare, badagliare, fiatare*) presumibilmente condizionate da fattori diatopici e diafasici: notiamo in particolare che il volgarizzatore siciliano varia utilizzando, nel quadro di una perifrasi incoativa, un iperonimo polisemico patrimoniale e di base latina (*flatare* per 'respirare', cfr. nel TLIO *fiatare* 'respirare; soffiare, ecc.' e cfr. il sic. *çiatari* 'respirare, fiatare; prendere respiro, ecc.' in VS), in luogo dell'atteso *badagliare*[6].

Andando oltre il caso di studio specifico evidenziamo qui l'importanza e l'interesse delle serie sinonimiche composte dalle testimonian-

[5] Per un quadro delle antiche traduzioni dei *Dialogi* prodotte in area italiana si veda la rassegna di Cerullo (2016: 17-22) e la bibliografia ivi riportata.

[6] Si noti che nel *Declarus* (D) di Angelo Senisio, primo glossario latino-siciliano (consultabile nel *Corpus TLIO*), *badaglare / badaglari* glossa più di un verbo latino (nel dettaglio *halo* (D *alo*), *oscito* (D *osito*), *hio* (D *hyo*) e *exippito* (D *exiptito*)), ma vale fondamentalmente 'aprire la bocca ed emettere un sospiro': ha dunque una semantica più ristretta rispetto a *flatare* e una più spiccata connotazione espressiva. Non possiamo escludere peraltro che il dettato del testo latino tradotto fosse parzialmente diverso o che il prestito non fosse noto al volgarizzatore o al copista del testo restituito dall'edizione. Sulla complessa tradizione del *Libru de lu Dialagu* e sulla collocazione linguistica (Calabria settentrionale) del suo più importante testimone (il ms Roma, Biblioteca Nazionale Centrale Vitt. Em 20) si veda la bibliografia critica ripercorsa recentemente da Cerullo (2016: 19-20 e n. 18).

ze del corpus, ulteriore strumento diagnostico che consente di formulare valutazioni sull'incidenza del fattore diatopico nell'articolazione del repertorio. È evidente l'interesse delle traduzioni e delle glosse di diversa area, utilizzabili come risposte omogenee a una medesima inchiesta[7]. In assenza del contrasto tra varietà elicitate dalle traduzioni parallele di un medesimo testo originale, l'individuazione di sinonimi e geosinonimi può contare sulle ricerche trasversali praticabili nel vocabolario: in particolare è funzionale a un'indagine onomasiologica, pur embrionale, la ricerca di parole presenti nelle definizioni. Un problema non banale nell'impostare la ricerca sarà delimitare la porzione semantico-concettuale del lessico incline a favorire la risalita nei testi di opzioni lessicali alternative o differenziali che abbiano una connotazione diatopica o piuttosto diastratica o diafasica.

3.1 Esperimenti sul lemmario esclusivo di un testo

In che misura un testo didascalico e dottrinale come il *Commento* alla Commedia del bolognese Iacopo della Lana arricchisce il lemmario complessivo del TLIO con parole marcate in diatopia? Quale porzione del vocabolario del testo dà spazio al dominio delle *differenze* dotate di un radicamento locale?

Il testo è attualmente consultabile nel *Corpus TLIO* e nel *Corpus OVI* nell'edizione di Volpi (2009). Mirko Volpi ha optato per l'edizione sinottica di due degli oltre cento codici che tramandano il *Commento*: il Riccardiano-1005 Braidense AG XII2 (Rb), trascritto a Bologna dal maestro Galvano da Bologna tra la metà degli anni trenta (non prima del 1334) e il decennio successivo, e il codice Trivulziano 2263 (M2), realizzato nel 1405 dal pisano Paolo di Duccio Tosi, rappresentano al meglio – nell'esame condotto dallo studioso – i due principali rami linguistici (rispettivamente emiliano-veneto e toscano) che caratterizzano la tradizione del *Commento* (cfr. Volpi 2010: 13-26).

L'inclusione del testo nei due corpora consente di illustrare la finalità che attualmente contraddistingue le due raccolte. Il *Corpus TLIO*, proiettato a una rappresentazione il più possibile esaustiva delle antiche varietà italoromanze, accoglie l'edizione del codice bolognese Rb che tramanda un testo giudicato molto vicino all'originale, scritto dal Lana tra il 1323-24 e il 1328. Il *Corpus OVI*, calibrato su un nucleo

[7] Ho fornito alcuni esempi di questa metodica in Giuliani (2016: 111sgg.).

testuale più ampio, accoglie l'edizione del toscano M2, «miglior rappresentante della linea linguistica toscana» (Volpi 2010: 26).

Poiché il *Corpus OVI* contiene per intero il *Corpus TLIO*, integrandolo con altri testi, entrambe le versioni sono consultabili in maniera sinottica nel *Corpus OVI*[8].

La diatopicità del commento del Lana riportato da Rb è valutabile secondo una duplice linea orizzontale: innanzitutto focalizzando l'attenzione sul confronto linguistico apparentemente paritario col lessico dantesco – ben sostenuto e valorizzato da molteplici chiarificazioni geosinonimiche –, in secondo luogo impostando un confronto con la versione toscanizzata tramandata da M2, un adattamento che tocca talora la sola *facies* grafica e fonetica ma che ricorre non di rado a sostituzioni e traduzioni, naturalmente guidate da un intento interpretativo.

Ma quante e quali *differenze* lessicali appaiono marcate sul piano diatopico?

Ho provato a sondare questo dato sul "lemmario esclusivo" del *Commento* che può essere isolato (selezionando le quattro unità testuali che rappresentano l'edizione di Rb nel *Corpus TLIO*) utilizzando una specifica funzione dal software GATTO (http://www.ovi.cnr.it/Il-Software.html#gatto): trattasi dell'insieme dei lemmi non associati a forme testuali presenti in altri testi del corpus. Si tratta di una quota lessicale che contribuisce a incrementare il lemmario del TLIO.

Ho ottenuto così una lista di 1572 lemmi che ho provveduto a scremare, eliminando lemmi latini o comunque non italoromanzi e trascurando le forme lemmatizzate come ambiguità grammaticali (a.g.) e, ulteriormente, i lemmi che duplicano lemmi già attestati integrandone la sola categorizzazione grammaticale (ad es. sostantivi attestati nel *Commento* come aggettivi o verbi attivi attestati come pronominali); si tratta, in quest'ultimo caso, di lemmi contemplati da un'unica entrata del TLIO. Ho ottenuto in questa maniera una lista di 382 hapax o lessemi ad attestazione monotestuale dei quali circa 61 (ovvero ca. 1/6 del totale) sono accertabili come localismi d'ambiente settentrionale e più precisamente emiliano-bolognese o veneto-veneziano. Preciso a tal proposito che il testo fu scritto dal Lana probabilmente a Venezia e accoglie un significativo numero di

[8] Preciso che il *Corpus OVI* non è lemmatizzato: può essere consultato per forme e per "lemmi muti", ovvero per coppie forma-lemma ricavate dal dizionario di macchina del *Corpus TLIO*.

venetismi, rispondenti anche alla tensione verso una *scripta* sovramunicipale, leggibile idealmente entro uno spazio culturale compreso tra Venezia e Bologna (cfr. Volpi 2010: 13-14; 35sgg.).

I numeri sono indicativi ma la natura del contributo lessicale legato a un circuito locale è meglio chiarito da una classificazione semantica macroscopica. Si veda il seguente schema tripartito accompagnato da una selezione di esempi (le forme di citazione corrispondono a lemmi presenti attualmente nel corpus):

– Lessico che designa enti, elementi ed eventi osservati nell'ambiente naturale e sociale (persone, animali, vegetali, oggetti, conformazioni del terreno, fenomeni atmosferici):
  Cfr. *bugame* 'buco, cavità', *cacciafusto* 'macchina bellica da lancio', *cazola* 'susina vuota', *cocale* 'gabbiano', *fanticina* 'fanciulla di tenera età', *frassare* 'rompere', *intoppedo* 'intoppo', *lucinero* 'lampo, baleno', *lucire* 'risplendere', *ludria* 'lontra', *magarasso* 'ramarro', *mambretta* 'piastra che orna la cintura', *masegnola* 'pietra di copertura', *medazolo* 'piccola capanna di paglia', *modiglione* 'mensola', *moscare* 'circondarsi di mosche', *muttilare* 'muggire', *orello* 'orlo', *raffrezzare* 'aumentare la velocità (rif. a un volatile)', *riolo* 'rigagnolo', *rusco* 'spazzatura', *stancarolo* 'puntello', *stellata* 'palizzata'.
– Lessico legato all'esperienza fisica e corporea:
  Cfr. *asogar* 'tirare con una fune', *brancioni* 'carponi, con le mani e le ginocchia a terra', *bruscular* 'perdere l'equilibrio e cadere', *culatta* 'deretano', *gargarozzo* 'gola', *lisegar* 'scivolare', *mormorero* 'mormorio', *pedicare* 'camminare', *peteggiare* 'emettere peti', *poggiatello* 'parte superiore della gota', *rigrappare* 'raccogliere le gambe' *scalpeggiare* 'calcare con i piedi, calpestare', *tremolazzo* 'brivido provocato dal freddo'.
– Parole che caratterizzano, attitudini, stati emotivi e atteggiamenti (con riferimento a persone o animali):
  Cfr. *aggattigliare* 'lasciarsi blandire', *arabido* 'dominato dalla rabbia', *asivo* 'umile, modesto' e *asivamente* 'in maniera modesta', *grepolo* 'rabbioso (rif. a un cane)', *radegheza* 'errore', *remesedada* 'mescolanza senza ordine', *scalmaccio* 'inquietudine'.

È doveroso precisare che l'accertamento del rapporto precipuo con un circuito locale si basa in grossa misura sul riscontro fornito dallo

studio lessicale di Volpi (2010), dai repertori dialettali o dalle raccolte di latino medievale emiliano e veneto di Sella (1937 e 1944). Il contrasto con il testo dantesco e con l'adattamento toscano proposto dal M2 fornisce ulteriori elementi a sostegno dell'identificazione della diatopicità di alcuni lemmi esclusivi di Rb.

Così, ad esempio, la voce dantesca *vivagno* 'margine' glossata con *orello* («*Vivagno* [si è] li extremi *orelli* del panno» ed.: 1980) ha riscontro anche nel parm. e moden. *orel* 'orlo' (Malaspina 1856-59 e Maranesi 1867-68 s.v.), *frassare* 'rompere' di Rb («rompendo e *frassando* l'arbore» ed.: 1624), sostituito in M2 da *stracciare* («rompendo e *stracciando* l'arbore» ed.: 1625) ha riscontro anche nel romagn. *frassè* 'sbadire, rompere o disfare la baditura' (Morri 1840 s.v.), ma *poggiatello* 'parte superiore della gota' («*poçadelli* delle gote») sostituito in M2 e da altri codici dal tipo *pomello* («*pomelli* delle gote» ed.: 1413) non ha riscontri al di fuori del testo (cfr. Volpi 2010: 158).

Il fattore diatopico coinvolge una porzione di lessico che nell'ottica sistemica del GraDIt ha i caratteri psico-mentali del "vocabolario d'alta disponibilità" (cfr. De Renzo 2005: 216), vincolato al quotidiano, al locale, al vicino, al noto e al temporaneo, un vocabolario che nel GraDIt è considerato parte integrante del lessico comune.

Oltre questo gruppo di localismi (emiliani, veneti o più ampiamente settentrionali) la lista dei lemmi esclusivi del *Commento* lanèo include un repertorio cospicuo di lessico tecnico, filosofico, scientifico e retorico ricco di latinismi crudi e neoformazioni. Rientrano in questa quota anche lemmi divenuti rilevanti e centrali nel vocabolario contemporaneo: v. ad es. *eccellere* 'essere superiore', e *probabilità* 'condizione di ciò che si crede verosimile'. Si deve osservare che una netta distinzione tra voci settoriali / filosofico-scientifiche e cultismi per un verso e localismi per altro verso potrebbe essere unilaterale se non fallace: il tecnicismo *lunazione* 'periodo di rivoluzione della luna intorno alla terra, mese' – presente solo nel *Commento* lanèo (forma di Rb *lunationi*) e nel *Fiore di virtù* (1313/23, bologn.) (forma *lunaxoni*) – si distingue dai lemmi prima menzionati per la maggiore astrattezza del contenuto, pur legato all'esperienza dei cicli del mondo naturale, ma è verosimilmente un localismo emiliano, richiamato infatti dal romagn. *lunazion* 'il tempo del corso della luna dal principio del novilunio fino al termine dell'ultimo quarto' (Morri 1840 s.v.). Non possiamo escludere d'altronde che appartengano storicamente a un

lessico legato a una fruizione locale alcuni cultismi e semicultismi rappresentativi di tassonomie concettuali che hanno permeato la cultura di specifiche comunità. Colpisce la veste latineggiante del *lunationi* lanèo trattato certamente alla stregua di latinismi e neoformazioni di attestazione monotestuale come *accertatione*, *apostrofatione, conciliatione, discrepatione, equiparatione, manitione* ('trattenimento?', cfr. Volpi 2010: 127), *radiatione / radiacione* e *ventilatione*, e certamente partecipe della progettualità di un testo legato alla cultura universitaria bolognese e destinato ad alimentare quella stessa cultura entro un più ampio ambiente settentrionale (cfr. Volpi 2010: 36-9).

«I lessemi di alta disponibilità raramente affiorano nei testi scritti e nel parlato esofasico, ma sono altamente presenti nell'endofasia e si individuano intervistando campioni di locutori e accertando dalle risposte lo scarto tra frequenza percepita, ritenuta alta [...] e frequenza reale spesso prossima a zero anche in un corpus di molti milioni di occorrenze» (De Mauro 2016: 50). È evidente che in un corpus diacronico la bassa frequenza non distingue di per sé le *differenze* correlate a condizionamenti diatopici isolandole da quelle dovute ad altro tipo di fattori. Nessun carattere lessicale intrinseco, di ordine qualitativo o quantitativo, distingue *frassare*, da *lunazione* e da *probabilità*, voci rese differenti e rappresentative di più di una dimensione variazionale solo nel confronto con altre sincronie linguistiche interne e esterne rispetto al corpus qui esaminato.

Riferimenti bibliografici

AIS: Jaberg, Karl & Jud, Jakob. 1928-40. *Sprach- und Sachatlas Italiens und der Südschweiz, Atlante Italo-Svizzero.* Rongier & Co.: Zofingen.

Ambrosini, Riccardo. 1977. *Stratigrafia lessicale dei testi siciliani dei secoli XIV e XV*. Palermo: Centro di Studi filologici e linguistici siciliani.

Barbato, Marcello. 2019. Recensione a "Cosimo Burgassi, Elisa Guadagnini, La tradizione delle parole. Sondaggi di lessicologia storica (Strasbourg, EliPhi, 2017)". *Medioevo Romanzo* 43: 242-245.

Burgassi, Cosimo & Guadagnini, Elisa. 2017. *La tradizione delle parole. Sondaggi di lessicologia storica*. TraLiRo – Lexicologie, onomastique, lexicographie. Strasbourg: ÉLiPhi.

Cella, Roberta. 2010. Prestiti nei testi mercantili toscani redatti di là dalle Alpi. Saggio di glossario fino al 1350. *La lingua italiana* VI. 57-99.

Cerullo, Speranza. 2016. Un volgarizzamento inedito dei *Dialogi* di Gregorio Magno in un codice senese. *Critica del testo* 19(2). 9-76.

Corpus OVI: *Corpus OVI dell'italiano antico* (ultimo aggiornamento: 10 gennaio 2022), diretto da Pär Larson, Elena Artale & Diego Dotto, Istituto Opera del Vocabolario Italiano: http://gattoweb.ovi.cnr.it/.

Corpus TLIO: *Corpus TLIO per il vocabolario* (ultimo aggiornamento: 10 gennaio 2022), diretto da Pär Larson, Elena Artale & Diego Dotto, Istituto Opera del Vocabolario Italiano: http://tlioweb.ovi.cnr.it/.

Coseriu, Eugenio. 1967. Structure lexicale et enseignement du vocabulaire. In *Les Théories linguistiques et leurs applications*, 9-51. Strasbourg: Conseil e l'Europe, Aidela.

Coseriu, Eugenio. 1981. Los conceptos de dialecto, nivel y estilo de lengua y el sentido propio de la dialectología. *Lingüística Española Actual* 3(1). 1-32.

DELCat: Coromines, Joan. 1980-91. *Diccionari etimològic i complementari de la llengua catalana, amb la collaboració de Joseph Gulsoy & Max Cahner*. 9 voll. Barcelona: Curial edicions catalanes.

De Mauro, Tullio. 2016. La stratificazione diacronica del vocabolario di base italiano. In Leonardi, Lino & Maggiore, Marco (a cura di), *Attorno a Dante, Petrarca, Boccaccio: la lingua italiana. I primi trent'anni dell'Istituto CNR Opera del Vocabolario Italiano: 1985-2015*. Atti del convegno internazionale, sotto l'Alto Patronato del Presidente della Repubblica (Firenze, 16-17 dicembre 2015). BOVI. Supplementi 5, 45-52. Alessandria: Edizioni dell'Orso.

De Renzo, Francesco. 2005. Nuove rilevazioni sul vocabolario di base e di alta disponibilità. In Tullio De Mauro & Isabella Chiari, *Parole e numeri. Analisi quantitative dei fatti di lingua*, 215-32. Roma: Aracne.

Giuliani, Mariafrancesca. 2016. Tra lessicografia e geolinguistica (rileggendo Folena). In Buchi, Éva & Chauveau, Jean-Paul & Pierrel, Jean-Mari (a cura di), *Actes du XXVIIe Congrès international de linguistique et de philologie romanes (Nancy, 15-20 juillet 2013), Section 8: Linguistique variationnelle, dialectologie et sociolinguistique*, a cura di Jean-Paul Chaveau, Marcello Barbato, Ines Fernández-Ordóñez: 105-120. Strasbourg: Société de linguistique romane / ÉLiPhi. 2 voll.

Giuliani, Mariafrancesca. 2019. Le antiche voci non toscane nella tradizione lessicografica italiana: l'approccio della Crusca e del Tommaseo Bellini. In Leonardi, Lino & Squillacioti, Paolo (a cura di), *Italiano antico, italiano plurale*, Atti del Convegno OVI, 13-14 settembre 2018, BOVI. Supplementi 7, 103-26. Alessandria: Edizioni dell'Orso.

GraDIt: *Grande dizionario italiano dell'uso*, ideato e diretto da Tullio De Mauro. Torino: UTET, 2000.

LEI: *Lessico Etimologico Italiano*, fondato da Max Pfister, diretto da Elton Prifti & Wolfgang Schweickard. Reichert: Wiesbaden 1979-.

Malaspina, Carlo. 1856-59. *Vocabolario parmigiano-italiano: accresciuto di più che cinquanta mila voci*. Parma: Carmignani.

Maranesi, Ernesto. 1867-68. *Vocabolarietto domestico del dialetto modenese colla voce corrispondente italiana*, Modena: Tip. dell'Imm. Concezione.

Morri, Antonio. 1840. *Vocabolario romagnolo-italiano*. Faenza: dai tipi di P. Conti all' Apollo.

Musso, Pasquale. 2013. Interferenze catalane in un volgarizzamento siciliano del XIV secolo. In Krefeld, Thomas & Oesterreicher, Wulf & Schwägerl-Melchior, Verena (a cura di), *Reperti di plurilinguismo nell'Italia spagnola (sec. XVI-XVII)*, 29-50. Berlin-Boston: De Gruyter.

Porta, Giuseppe (a cura di). 1990-91. *Giovanni Villani, Nuova Cronica*. 3 voll. Parma: Fondazione Pietro Bembo / Ugo Guanda Editore.

Sella, Pietro. 1937. *Glossario latino-emiliano*. Città del Vaticano: BAV.

Sella, Pietro. 1944. *Glossario latino-italiano (Stato della Chiesa, Veneto, Abruzzi)*. Città del Vaticano: BAV.

TLIO: *Tesoro della Lingua Italiana delle Origini*, diretto da Paolo Squillacioti [fondato da Pietro G. Beltrami, poi diretto da Lino Leonardi], Firenze, Istituto Opera del Vocabolario Italiano <http://tlio.ovi.cnr.it/> (ultimo aggiornamento: 15.02.2022).

Tomasin, Lorenzo. 2019. *Il caos e l'ordine. Le lingue romanze nella storia della cultura europea*. Torino: Piccola Biblioteca Einaudi.

Ugolini, Francesco A. (a cura di). 1967. *Valeriu Maximu translatatu in vulgar messinisi per Accursu di Cremona*. Palermo: Mori. Centro di studi filologici e linguistici siciliani.

Vaccaro, Giulio. 2019. «Seminavano grano nelle carreras della città». Parole e saperi dalla Spagna all'Italia nel Trecento. In Cadeddu, Maria Eugenia & Marras, Cristina (a cura di), *Linguaggi, ricerca, comunicazione. Focus CNR*. Plurilinguismo e migrazioni I, 67-84. Roma: CNR, Edizioni.

Volpi, Mirko (a cura di). 2009. *Iacomo della Lana, Commento alla 'Commedia'*, con la collaborazione di Arianna Terzi. 4 voll. Roma: Salerno Ed.

Volpi, Mirko. 2010. *«Per manifestare polida parladura». La lingua del Commento Lanèo alla Commedia nel ms. Riccardiano-Braidense*. Roma: Salerno Ed.

VS: *Vocabolario siciliano* fondato da Giorgio Piccitto, Catania-Palermo: Centro di Studi filologici e linguistici siciliani, 1977-2002.

DOI: 10.17469/O2106SLI000025

VITTORIO GANFI, VALENTINA PIUNNO

Diacronia e sincronia delle polirematiche con struttura preposizionale: un'analisi su corpora

Questo lavoro intende chiarire le dinamiche di lessicalizzazione e grammaticalizzazione che hanno condotto alla formazione di polirematiche con la struttura di sintagmi preposizionali in italiano contemporaneo, a partire dall'analisi delle configurazioni in latino e nelle fasi più antiche di lingua. Nell'analisi si intende (a) fornire un quadro rappresentativo delle configurazioni in latino, in italiano antico e in italiano contemporaneo, (b) chiarire il rapporto diacronico tra configurazioni che impiegano gli stessi lessemi, (c) mettere in luce le divergenze generali interne al sistema delle polirematiche con struttura di sintagma preposizionale. Nell'analisi della trafila diacronica che ha condotto alle configurazioni contemporanee, viene infine analizzata l'evoluzione delle diverse funzioni, individuando un percorso storico che ha portato alla diversificazione funzionale del sistema delle polirematiche preposizionali in italiano.

Parole chiave: sintagmi preposizionali, unità lessicali complesse, grammaticalizzazione, lessicalizzazione, italiano antico e contemporaneo.

1. *Introduzione*

Con il termine *polirematiche* ci si riferisce a uno dei poli della scala di aggregazione delle parole (Simone 2006a, 2007), che mira a rappresentare le diverse unità combinatorie sintagmatiche in un *continuum* dalla sintassi al lessico (Figura 1). La scala organizza le diverse combinazioni in base al grado di coesione tra costituenti e il livello di lessicalizzazione del sintagma. Nel modello, le polirematiche sono rappresentate verso il polo che tende al lessico.

Figura 1 - *Scala di aggregazione (Simone 2006a, con modifiche)*

In accordo con Voghera (2004: 56) le *polirematiche* possono essere definite come "sequenze che non superano di norma l'estensione di un sintagma e che presentano una coesione interna maggiore di quella prevedibile sulla base della loro struttura sintattica". Le polirematiche hanno ricevuto denominazioni, definizioni e classificazioni di diverso tipo (Masini 2009). La terminologia si differenzia in base al livello di analisi considerato e nella scelta dei 'confini combinatori'. Nella letteratura italiana, tra le denominazioni più comuni troviamo: a) *unità lessicali superiori* (Dardano 1978), ovvero sequenze derivate da frasi soggiacenti al sintagma; b) *lessemi complessi* (De Mauro & Voghera 1996), ovvero sequenze dalla cristallizzazione sintattica variabile, portatrici di un significato che non è ricostruibile a partire dalla somma dei significati dei costituenti e "fattore d'ordine nella norma della lingua" (De Mauro & Voghera 1996: 106); c) *parole costruzionali o sintagmatiche* (Simone 1996, 2006a, 2006b; Masini 2009), vale a dire sequenze costituite da due o più lessemi, che tendono ad agire come se fossero unità lessicali autonome, costituendo una nuova designazione e formando "una classe a sé anche nella competenza del parlante" (Simone 1996: 48).

A seconda della distribuzione sintattica e del valore funzionale, le polirematiche possono essere inglobate in profili combinatori diversi (per es. nominale, aggettivale, verbale, ecc). Inoltre, in base alla configurazione sintagmatica, possono essere raggruppate in diversi *formati* o *schemi* sintagmatici. Tra quelli individuati in letteratura (Voghera 1994), in questo lavoro ci si sofferma sulle polirematiche con forma di sintagma preposizionale, molto diffuse in ambito romanzo (Piunno 2018), ed eterogenee sia in termini strutturali sia in termini funzionali (Giacalone Ramat 1994; Casadei 2001; Voghera 1994; 2004; Ganfi & Piunno 2017; Piunno & Ganfi 2019, 2021). Per chiarezza, ne riportiamo due esempi, il primo (*a portata di mano*) con funzione simile a quella di un aggettivo, il secondo (*dal momento che*) con funzione di congiunzione:

(1) un'enciclopedia ancora più vicina alle tue esigenze. Agile, utile, dinamica, sempre *a portata di mano*

(2) Sono quasi le 4 e *dal momento che* non riuscirò a fare merenda con un piatto di spaghetti, gentilmente declino

Questo lavoro intende (a) fornire un quadro rappresentativo delle configurazioni in latino, in italiano antico e in italiano contemporaneo, (b) individuare le possibili correlazioni tra la frequenza di occorrenza nei corpora e il valore grammaticale delle strutture, (c) chiarire il rapporto diacronico tra le configurazioni nelle diverse fasi di lingua, (d) mettere in luce le divergenze tra le diverse fasi indagate.

Si propone, pertanto, un'analisi sincronica e diacronica, basata su corpora e opere lessicografiche. In particolare, in linea con il tema del volume, viene dato maggiore rilievo ai dati quantitativi, e a come questi possono sostenere e guidare l'analisi qualitativa. Per l'italiano contemporaneo si fa riferimento a dati estratti da due *corpora* di italiano scritto (il *corpus la Repubblica* e il *corpus ITTenTen16*), per mezzo di ricerche basate sulla selezione delle più frequenti sequenze di parti del discorso tra loro adiacenti[1]. A queste unità sono state aggiunte le sequenze polirematiche presenti nel *GRADIT* (De Mauro 1999), che a causa della bassa frequenza non emergono dal corpus. Per il latino è stato consultato il corpus *PHI Latin Texts*[2]. La ricerca qualitativa sull'italiano antico si basa sulle entrate lessicografiche del dizionario *TLIO*, mentre quella quantitativa fa riferimento a dati estratti dal *corpus OVI*[3]. 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. A seconda del tipo di funzione sono prese in considerazione la coesione sintattica, il grado di apertura alla variabilità paradigmatica, la schematicità, la semantica e la frequenza di occorrenza nei corpora selezionati. I dati impiegati derivano da diverse analisi svolte dagli autori negli ultimi anni[4]. I tipi di polirematiche presi in considerazione per le varie fasi dell'italiano sono quelli con funzione avverbiale, agget-

[1] In particolare, i dati relativi alle polirematiche SP avverbiali, aggettivali, e multifunzione, sono stati ricavati dallo studio condotto in Piunno (2018) per mezzo del corpus di italiano scritto giornalistico *la Repubblica* (http://sslmit.unibo.it/repubblica); i dati sulle polirematiche SP preposizionali derivano dallo studio di Ganfi e Piunno (2017), basato sul corpus *ITTenTen16* (https://www.sketchengine.eu/corpora-and-languages/corpus-list/). Da quest'ultimo corpus sono state ricavate anche le polirematiche congiunzionali.

[2] Sito web: https://latin.packhum.org/index.

[3] Siti web: *Corpus Opera del Vocabolario Italiano* (http://www.ovi.cnr.it), *Tesoro della Lingua Italiana delle Origini* (http://tlio.ovi.cnr.it/TLIO).

[4] Cfr. Piunno (2015, 2016, 2018, 2020), Ganfi e Piunno (2017), Piunno e Ganfi (2019, 2021), Ganfi (2021).

tivale, preposizionale e congiunzionale. L'ordine in cui vengono presentati richiama una gerarchia implicazionale individuata e discussa in Piunno e Ganfi (2021, in stampa), che non può essere approfondita in questa sede.

2. *Analisi sincronica*

Le polirematiche con funzione avverbiale raccolte dal corpus di italiano contemporaneo sono complessivamente 2413, per un totale di 69 schemi sintagmatici diversi. Si tratta di sequenze ormai largamente presenti nei dizionari e nelle grammatiche, e del gruppo più numeroso sia in termini di esemplari sia dal punto di vista delle strutture. Queste sequenze ricoprono la tipica funzione avverbiale di modificazione del verbo (3), avverbio (4), aggettivo (5), o di un'intera frase (6):

(3) usciva *di rado* e mai senza qualcuno che lo accompagnasse

(4) sarà pubblicato in Gazzetta ufficiale stasera o *al più tardi* domani

(5) Adempimenti burocratici *a dir poco* barocchi affaticano i docenti

(6) *Sul serio*, non ce la faccio più con queste vacanze!

Tra le 69 configurazioni raccolte, emerge per frequenza il pattern [Prep Nome] (per es. *a volte*), che si attesta nel 47% degli esempi. In questo caso, le strutture combinatorie si formano essenzialmente grazie a due schemi, che coprono insieme l'80% dei casi raccolti: [*a* Nome] (per es. *a stento, a forza, a capo, a mano, a piacimento*) e [*in* Nome] (per es. *in amicizia, in anticipo, in branco, in pratica, in coppia*). Questi due pattern (o alcuni loro sottotipi, cf. Piunno 2018) sono più regolari (Haspelmath 2002), in termini di forme. Diversi gradi di astrattezza o schematicità sono riconducibili ai pattern più frequenti, tra cui in particolare la configurazione [Prep Nome Prep Nome], che mostra sia sequenze completamente lessicalizzate (per es. *a scanso di equivoci*, *in odore di santità*, all'estremo destro della scala di aggregazione in Figura 1) sia formati riempiti solo parzialmente da unità lessicali[5].

[5] Per es. la struttura [*di* NOME *in* NOME], che può attirare a sé nomi che indicano un'unità temporale (*di giorno in giorno, di minuto in minuto*), con un range chiuso) o un'unità spaziale (*di città in città, di luogo in luogo*), con possibilità combinatorie più ampie (Piunno 2020).

Per la funzione aggettivale sono state raccolte 2046 sequenze, per un totale di 33 configurazioni sintagmatiche (circa la metà rispetto alle avverbiali). Si tratta, pertanto, di un gruppo meno popolato, sia in termini di esemplari sia sul piano strutturale. Tali sequenze sono spesso ancora ignorate nei dizionari e nelle grammatiche. Le polirematiche aggettivali ricoprono la tipica funzione di modificazione del nome (7)-(8), e come gli aggettivi possono essere modificate da avverbi (9)-(10):

(7) Egli era un giurista, un uomo talentuoso e *di buon cuore*

(8) Ricamatrice esperta in tutti i tipi di ricamo *a mano* e *a macchina*

(9) Un forno con la F maiuscola, [...] gestito da fornai e commesse veramente *alla mano* e simpatici

(10) In questa lista non troverete le pizzerie solo *da asporto*

Come gli aggettivi propriamente detti, oltre alla funzione attributiva, si trovano in posizione predicativa:

(11) I prodotti sono di qualità (oltre che *a buon prezzo*)

Anche in questo caso, i più frequenti sono quelli con struttura [Prep Nome] (per es. *a pieghe, da incasso*), che coprono il 66% degli esempi raccolti. Rispetto alle sequenze avverbiali, le polirematiche aggettivali con struttura [Prep Nome] si distribuiscono più o meno equamente tra i diversi pattern: per es. il pattern [*a* Nome] ricorre nel 30% dei casi, seguito da [*in* Nome] (29%), [*di* Nome] (22%) e [*da* Nome] (18%).

Tra i due tipi descritti fino ad ora, vale la pena menzionare gli schemi *multifunzione*, un gruppo di sequenze dal valore sia avverbiale sia aggettivale, in base al contesto. Nei dati raccolti, ammontano a 585 sequenze, per un totale di 13 configurazioni sintagmatiche. Di seguito alcuni esempi in funzione aggettivale (12a-13a) e avverbiale (12b-13b):

(12) a. La confettura extra di Ribes Nero è prodotta utilizzando le bacche tonde del frutto *a grappolo*
b. Nodi semplici o più complessi per fermare, allineare, raggruppare o legare *a grappolo*

(13) a. Edizione ridotta e *a reti unificate*
b. Lo Speciale Coppa Italia andrà in onda, eccezionalmente, *a reti unificate*

In questo caso il numero di formati sintagmatici possibili è piuttosto ristretto (sono solo 13 i pattern raccolti). La struttura che occorre più frequentemente anche negli esempi multifunzione è [Prep Nome] (per es. *in ferie*), che copre ben il 71% dei casi.

Ulteriore funzione associabile alle poliremiatiche in oggetto è quella preposizionale. L'analisi, basata in questo caso sul *corpus ITTenTen16*, ha permesso di estrarre 307 sequenze, per un totale di 20 configurazioni che risalgono allo stesso unico pattern [Prep Nome Prep]. Le strutture con funzione preposizionale possono avere valori di diverso tipo, come ad esempio locativo (14), causativo (15), comitativo (16), agentivo (17), beneficiario (18):

(14) le persone hanno protestato *di fronte a* un ospedale

(15) *A causa di* una magia sbagliata la strega finisce prigioniera

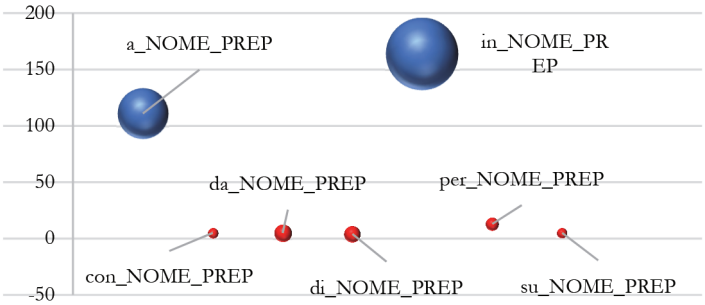
(16) Tante proposte per allenarvi, pedalare *in compagnia di* nuovi amici

(17) Le adesioni proseguono, sia *da parte di* allevatori che di operatori

(18) Banca Etica privilegerà l'erogazione del credito *a favore di* organizzazioni appartenenti al terzo settore

Gli esempi raccolti sono riconducibili essenzialmente a due configurazioni, che coprono quasi l'intera totalità dei casi: [*in* Nome Prep] (per es. *in caso di*; attestato nel 53% dei casi) e [*a* Nome Prep] (per es. *a cura di*; 36%). In questo caso si è scelto di mettere in relazione la numerosità dei pattern (il numero di *types*) con la frequenza di occorrenza di ciascuna poliremiatica (il numero di *tokens*). La Figura 2 rappresenta tale relazione:

Figura 2 - *Rapporto tra type e token nelle polirematiche con funzione preposizionale*



A ciascuna sfera del grafico è associata una specifica configurazione. La posizione della sfera (in altezza) indica il numero di esemplari riconducibili ad uno stesso schema; il diametro indica la frequenza di occorrenza nel *corpus* di tutte le polirematiche di quel gruppo. Le sfere rosse in basso individuano pattern preposizionali meno produttivi, ai quali tuttavia risalgono alcune preposizioni complesse pienamente lessicalizzate. In alcuni casi, la frequenza assoluta del pattern preposizionale corrisponde alla frequenza assoluta della singola preposizione complessa (per es. *da parte di* e *di fronte a*): tali sequenze sono spesso altamente grammaticalizzate (cf. Ganfi & Piunno 2017). Le sfere blu invece rappresentano schemi con i) un'altissima frequenza assoluta, ii) una più ampia flessibilità lessicale e iii) una rilevante produttività in termini di nuove forme. I formati [*a* Nome *di*] e [*in* Nome *di*] ricorrono infatti nel 90% delle strutture considerate. In questo caso è più difficile associare un valore funzionale definito alla configurazione sintagmatica, data l'alta variabilità dei lessemi con cui lo schema si costruisce.

Le polirematiche congiunzionali estratte dal *corpus* sono 110, per un totale di 11 configurazioni. Di seguito si registrano alcuni esempi, in cui la congiunzione polirematica introduce subordinate esplicite o implicite.

(19) Una buona dieta aiuta, *dal momento che* il corpo ha bisogno di agenti detossificanti

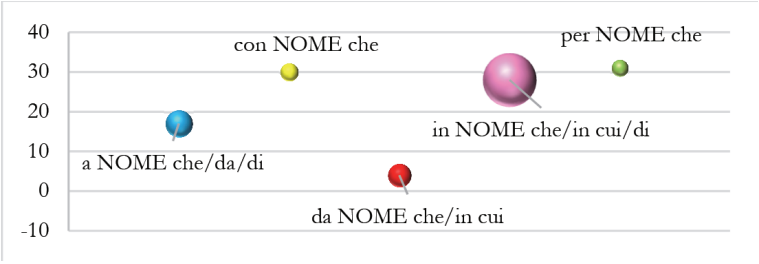
(20) È possibile pubblicare e redistribuire testi e immagini *a patto che* se ne citi la fonte

(21) ferme eventuali responsabilità dell'avvocato, *nell'ipotesi in cui* abbia agito senza consultare gli eredi

(22) Questa congregazione nacque sotto Gregorio XV *con il fine di* diffondere la fede cattolica

Riconducendo gli schemi alle preposizioni che li caratterizzano, si possono evidenziare cinque diversi tipi di formati, dalla numerosità simile[6], fatta eccezione per il gruppo [*da* Nome Cong/Prep], a cui può essere ricondotto un minor numero di esempi. Se però consideriamo la frequenza d'uso oltre che i tipi sintagmatici, questo gruppo si mostra quantitativamente rilevante. Usando il grafico a sfere, notiamo che la sfera relativa a questo formato (la sfera rossa), presenta un diametro piuttosto ampio (ha una somma di frequenze relativamente alta), ma si trova lungo la parte bassa del grafico (presenta quindi un numero esiguo di esemplari). Anche in questo caso l'alta frequenza dello schema si associa quasi interamente a un'unica polirematica congiunzionale: *dal momento che*.

Figura 3 - *Rapporto tra type e token nelle polirematiche con funzione congiunzionale*



Al contrario, la sfera rosa, che presenta il diametro maggiore (e quindi la frequenza di occorrenza più alta), ha anche un elevato numero di esemplari (si trova in alto nel grafico). Questi criteri consentono di organizzare le configurazioni sulla base del grado di lessicalizzazione e produttività. Le strutture più frequenti tendono ad assumere tratti

[6] Gli schemi congiunzionali identificati sono i seguenti: [*per* Nome Cong/Prep] (per es. *per il timore che*; 28%), [*con* Nome Cong/Prep] (per es. *con la speranza che*; 27%), [*in* Nome Cong/Prep] (per es. *in modo da*; 25%), [*a* Nome Cong/Prep] (per es. *al punto che*; 16%) e [*da* Nome Cong/Prep] (per es. *dal momento che*; 4%).

di schematicità e ad essere impiegate nelle forme più vicine alla grammatica (cf. Ganfi & Piunno 2017).

3. *Analisi diacronica*

Per l'analisi diacronica, il primo confronto è stato quello fatto con il latino. L'analisi del *corpus* ha permesso di mostrare che in latino le polirematiche con schema SP possono presentare esclusivamente la funzione avverbiale. L'inventario dei tipi costruzionali e delle forme del latino è più limitato rispetto al sistema dei modificatori avverbiali romanzi. Sul piano della variabilità strutturale, possono essere individuate delle configurazioni pienamente lessicalizzate, che non presentano variabilità paradigmatica, come si può vedere dai seguenti esempi[7]:

(23) *Ad* *ultimum* dolori succubuit (Latino)
PREP ultimo.ACC dolore.ABL cedette
'Alla fine cedette al dolore'
(Q. Rufus, *Historiae Alexandri Magni* 10.5.24.1)

(24) *ad* *hunc* *modum* coepit (Latino)
PREP questo.ACC modo.ACC cominciò
'in questo modo cominciò'
(Cornelius Tacitus, *Annales* 2.37.9)

(25) Rem publicam suam *in perpetuum* adflixerunt (Latino)
Stato.ACC suo.ACC PREP perpetuo vessarono
'vessarono sempre il proprio stato'
(Titus Livius, *Ab Urbe Condita* 28.41.1.1)

(26) haec vobis dixi *per* *iocum* (Latino)
queste.cose.ACC voi.DAT dissi PREP gioco
'Vi ho detto queste cose per scherzo'
(Titus Maccius Plautus, *Poenulus* 541, 542, 573)

[7] Vale la pena notare che il latino presenta anche strutture con un limitato grado di variazione paradigmatica, in cui la variabilità lessicale appare, comunque, molto limitata. Si osservino i seguenti esempi:
(1) [*summo/a cum* Nome$_{\{studio,audacia\}}$] *summo cum studio* 'con grande diligenza'
(2) [Agg $_{\{mirum,\ maiorem\}}$ *in modum*] *maiorem in modum* 'maggiormente'
(3) [*cum* Nome $_{\{cura,\ clade\}}$ *magno/a*] *cum cura magna* 'con grande cura'

Passando all'analisi dell'italiano antico, si può affermare, primariamente, che tutti i formati dell'italiano contemporaneo sono attestati anche nel *corpus* OVI.

Per quanto riguarda le configurazioni preposizionali con funzione avverbiale, sono state collezionate 1503 sequenze, suddividibili in 52 schemi diversi. L'analisi delle configurazioni mostra analogie con le strutture contemporanee, giacché in entrambe le fasi di lingua i tre schemi più frequenti, ovvero i tipi sintagmatici [Prep Nome] (per es. *a saetta* 'con grande velocità'; attestato nel 49% dei casi), [Prep Agg Nome] (per es. *con breve parola* 'brevemente'; 16% degli esempi), [Prep Art Nome] (per es. *all'alba*; 12% dei casi) coprono circa i tre quarti di tutte le polirematiche preposizionali con funzione avverbiale. In italiano antico si apprezzano configurazioni pienamente lessicalizzate, come gli esempi (27)-(28), ma anche strutture caratterizzate da un certo grado di variabilità[8]:

(27) perdere *in ciocca* (A. Pucci, *Centiloquio*, a. 1388 9, 42 1, 103, 15)
'perdere del tutto'

(28) venire *in gesta* (Sacchetti, *Rime*, XIV, 11, 7, 16, 3)
'arrivare in gruppo'

Per la funzione aggettivale in italiano antico sono state raccolte 281 sequenze, distinte in 5 schemi sintagmatici. Dallo studio del dato quantitativo, si ricava che il formato [Prep Nome] (per es. (canzone) *di cortesia*) è quello di gran lunga più produttivo, giacché viene impiegato nel 66% delle configurazioni raccolte, contro il 28% delle strutture [Prep Agg Nome] (per es. (uomo) *di dubbio padre*) e il 4% delle sequenze [Prep Art Nome] (per es. (luogo) *alla campestra*). Sul piano della variabilità paradigmatica delle sequenze aggettivali, sono state

[8] Particolarmente significativi sono i tipi costruzionali con schema astratto [[Prep$_1$ Nome$_1$] [Prep$_1$ Nome$_1$]], come negli esempi seguenti:
(1) [[*a* Nome$_{1<\text{concreto}>}$][*a* Nome$_{1<\text{concreto}>}$]]= 'continuità temporale/gradualità'
(es. *a brano a brano* 'a pezzo a pezzo', *a ciocca a ciocca* 'uno per uno')
(2) [[*a* Nome$_{1<\text{parte corpo}>}$][*a* Nome$_{1<\text{parte corpo}>}$]]= 'contatto/opposizione spaziale'
(es. *a cuore a cuore, a corpo a corpo, a costa a costa* 'a contatto ravvicinato')
(es. *a fronte a fronte* 'l'uno di fronte all'altro')
Si noti che negli esempi permane il medesimo valore semantico, malgrado la variazione lessicale.

individuate sia configurazioni altamente coese sia formati sintagmatici che presentano un certo grado di variabilità[9].

Per l'italiano antico si rivela significativo lo studio delle polirematiche multifunzione, ovvero delle combinazioni che possono modificare nomi e verbi. Sono state individuate 52 sequenze di questo tipo, suddivise in 5 schemi sintagmatici. La configurazione [Prep Nome] risulta essere quella più comune, visto che la si riscontra nell'83% delle combinazioni[10]. Negli esempi che seguono si mostra una stessa sequenza usata in funzione sia avverbiale (29a) sia aggettivale (29b):

(29) a. Lo signore dee amare suoi sudditi *di gran cuore*
(Tesoro volg. (ed. Gaiter), XIII ex. (fior.), L. 9, cap. 2, vol. 4, pag. 284.5)

b. e questi sono chiamati gente *di gran cuore* e di grande animo
(Egidio Romano volg., 1288 (sen.), L. 1, pt. 2, cap. 22, pag. 64.27)

Comparando le due funzioni (avverbiale e aggettivale), è possibile supporre l'esistenza della gerarchia implicazionale *Aggettivali* > *Multifunzione* > *Avverbiali* (Piunno & Ganfi 2021), che abbia anche una pertinenza diacronica, e che sarebbe avvalorata dalla presenza degli avverbiali in latino.

Per la funzione preposizionale, sono state raccolte 293 preposizioni complesse distinte in 25 schemi, di cui [*a* Nome *di*] copre il 17% e [*in* Nome *di*] il 15%. Si osservino i seguenti esempi:

(30) adonqua se move ella *a contrario de* li altri planeti.
(Restoro d'Arezzo, 1282 (aret.), L. I, cap. 12, pag. 19.17)

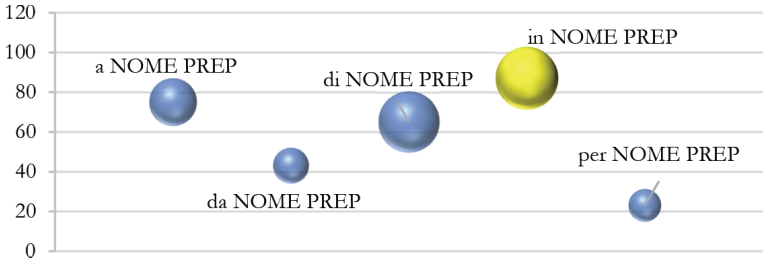
[9] Tra questi ultimi, vale la pena notare il formato che seleziona nomi di oggetti concreti e designa proprietà temporali del nome che va a modificare:
(1) [Nome_1 [*in* $\text{Nome}_{2<\text{concreto}>}$]] = 'proprietà temporale di Nome_1'
a. *fanciullo* in culla 'bambino'
b. *puledro* in dentatura 'giovane puledro'
c. *pianta* in fiore 'pianta fiorita'
Vale la pena notare che la medesima configurazione appare in italiano contemporaneo in alcune polirematiche aggettivali pienamente lessicalizzate, per es. i*n erba*.

[10] Le altre strutture identificate sono: [Prep Art Nome] (per es. *al colmo* 'al massimo'; 7%), [Prep Agg Nome] (per es. *di gran cuore* 'generoso/ generosamente'; 6%), [Prep Art Nome Prep Nome] (per es. *al peso di fiera*; 2%), [Prep Nome Cong Prep Nome] (per es. *a mazze e a frusti* 'armati/con armi'; 2%).

(31) Et così s'aviò la donna honesta *in compagnia del* franco cavalieri
(*Poes. an. pis.*, XIII ex. (3), 90, pag. 1350)

A differenza di altre classi, le preposizioni presentano una distribuzione meno polarizzata dei formati sintagmatici. Tuttavia, anche per questo gruppo alcuni schemi risultano essere più produttivi, mentre altri tendono a essere rappresentati da un numero esiguo di polirematiche. A questo proposito, risulta significativo il confronto tra *type* e *token*. Il grafico in Figura 4, che impiega le stesse modalità analitiche viste per l'italiano contemporaneo, mette in mostra i formati sintagmatici più frequenti, che vengono usati per costruire molte preposizioni complesse (per es. [*in* Nome Prep], rappresentato dalla sfera gialla):

Figura 4 - *Rapporto tra type e token nelle polirematiche con funzione preposizionale in italiano antico*



Rispetto all'italiano contemporaneo, non vi è la suddivisione così marcata tra formati produttivi e quelli che si associano a strutture grammaticalizzate. Le diverse sfere, infatti, si posizionano al centro del grafico, dimostrando di avere alta variabilità.

Sono state, infine, raccolte 50 polirematiche con funzione congiunzionale, raggruppate in 19 configurazioni sintagmatiche. Si osservino i seguenti esempi:

(32) Et *in caso che* non andasseno
(Stat. pis., 1322-51, [1330] Agg., cap. 2, pag. 596.19)

(33) *per cagione d'*insegnare
(Brunetto Latini, *Rettorica*, c. 1260-61 pag. 132, riga 21)

L'analisi quantitativa mostra che i tipi più frequenti sono quelli che presentano una struttura [Prep Nome Cong/Prep], che realizzano oltre il 60% dei casi. In particolare, a differenza dell'italiano contemporaneo, le sequenze congiunzionali tendono a raggrupparsi in uno

schema preminente, che copre quasi il 60% dei casi, vale a dire il pattern [*a* Nome Cong] (per es. *a cagione che*).

4. *Conclusioni*

Concludendo, abbiamo cercato di dimostrare che un approccio che sia al contempo qualitativo e quantitativo può essere utile allo studio delle unità polirematiche di tipo preposizionale. Sul piano sincronico, tali strutture possono essere analizzate prendendo in considerazione da un lato il pattern sintagmatico che caratterizza diversi esemplari, e dall'altro lato, la relazione quantitativa tra uno schema e la sua frequenza di uso. Nel primo caso il rapporto tra sequenze diverse può rivelare il consolidamento di una specifica struttura (lessicalizzazione) o di uno schema sintattico che può avere acquisito maggiore regolarità. L'elevata popolosità di uno schema può essere infatti indice del suo radicamento o della sua costruzionalizzazione (vale a dire, di una nuova associazione tra la struttura di uno schema sintagmatico e il relativo significato). Infine, il rapporto quantitativo tra tipi di strutture e frequenza d'uso può mettere in luce la rigida associazione tra un formato sintagmatico e un lessema, e lo specifico valore grammaticale che ne deriva. Sul piano diacronico, il confronto delle diverse fasi di lingua permette di identificare i percorsi che portano alla cristallizzazione e alla generalizzazione di alcuni pattern combinatori più astratti e produttivi, e di verificarne i possibili cambiamenti nel tempo (per es. aumento o perdita di produttività degli schemi). Il confronto tra i sistemi preposizionali mostra un incremento dello schematismo delle configurazioni nei dati contemporanei. Infine, l'analisi suggerisce l'esistenza di aree di sovrapposizione tra le diverse classi (per es. aggettivale/avverbiale, o preposizionale/congiunzionale) (cf. Piunno & Ganfi in stampa). L'approfondimento della questione, che meritebbe ulteriore spazio, permette di chiarire i rapporti diacronici tra le diverse funzioni.

Ringraziamenti

Per le preziose osservazioni, i nostri ringraziamenti vanno ai curatori del volume e agli organizzatori del convegno, ad Anna-Maria De Cesare, a Mariafrancesca Giuliani, e a Chiara Celata. Le responsabi-

lità per ogni eventuale errore presente nel testo è nostra. Per i soli fini accademici, l'articolo è così suddiviso: i paragrafi 1 e 3 sono da attribuire a Vittorio Ganfi, i paragrafi 2 e 4 a Valentina Piunno.

Riferimenti bibliografici

Casadei, Federica. 2001. Le locuzioni preposizionali. Struttura lessicale e gradi di lessicalizzazione. *Lingua e Stile* 36(1). 21–36.

Dardano, Maurizio. 1978. *La formazione delle parole nell'italiano di oggi*. Roma: Bulzoni.

De Mauro, Tullio. 1999. *Il grande dizionario italiano dell'uso*. Torino: Utet.

De Mauro, Tullio & Voghera, Miriam. 1996. Scala mobile. Un punto di vista sui lessemi complessi. In Benincà, Paola & Cinque, Guglielmo, & De Mauro, Tullio (a cura di), *Italiano e dialetti nel tempo. Studi dedicati a Giulio Lepschy*, 99–128. Roma: Bulzoni.

Ganfi, Vittorio & Piunno, Valentina. 2017. Preposizioni complesse in italiano antico e contemporaneo. Grammaticalizzazione, schematismo e produttività. *Archivio Glottologico Italiano*. CII(2). 184–204.

Ganfi, Vittorio. 2021. Diacronia della preposizione multiparola *fino a*. *L'analisi Linguistica e Letteraria* 29(2). 69–96.

Giacalone Ramat, Anna. 1994. Fonti di grammaticalizzazione. Sulla ricategorizzazione di verbi e nomi come preposizioni. In Cipriano, Palmira & Di Giovine, Paolo & Mancini, Marco (a cura di), *Miscellanea di studi linguistici in onore di Walter Belardi*, 877-896. Roma: Il Calamo.

Haspelmath, Martin. 2002. *Understanding Morphology*. London: Arnold.

Masini, Francesca 2009. *Parole sintagmatiche in italiano*. Cesena/Roma: Caissa Italia.

Piunno, Valentina 2015. Sintagmi Preposizionali come Costruzioni Aggettivali. *Studi e Saggi Linguistici*. 53(1). 65–98.

Piunno, Valentina. 2016. Multiword Modifiers in Romance languages. Semantic formats and syntactic templates. *Yearbook of Phraseology*. 7. 3–34.

Piunno, Valentina. 2018. *Sintagmi preposizionali con funzione aggettivale e avverbiale*. München: LINCOM Studies in Romance Linguistics.

Piunno, Valentina. 2020. Le combinazioni di parole parzialmente riempite in alcune lingue romanze. Schematismo e predicibilità semantica. *Romanica Olomucensia*. 32(1). 143–171.

Piunno, Valentina & Ganfi, Vittorio. 2019. Usage-based account of Italian Complex Prepositions denoting the Agent. *Revue Romane.* 54(1).141–175.

Piunno, Valentina & Ganfi, Vittorio. 2021. Synchronic and diachronic analysis of prepositional multiword modifiers across Romance languages. *Lingvisticae Investigationes*. 43(2). 352–379.

Piunno, Valentina & Ganfi, Vittorio. in stampa. Diachrony and synchrony of multiword prepositional phrases in (some) Romance languages. A corpus-based analysis. In Dejan, Stosic & Bras, Myriam & Abrard, Océane & Minoccheri, Chiara (Eds.), *Langage et Discours en Débats*. Parigi: Éditions de l'Harmattan.

Simone, Raffaele. 1996. Esistono verbi sintagmatici in italiano?. *Cuadernos de Filología Italiana*. 3. 47–61.

Simone, Raffaele. 2006a. Classi di costruzioni. In Grandi, Nicola & Iannàccaro, Gabriele (a cura di), *Zhi. Scritti in onore di Emanuele Banfi in occasione del suo 60° compleanno*, 383–409. Cesena/Roma: Caissa Italia.

Simone, Raffaele. 2006b. Nominales sintagmáticos y no-sintagmáticos. In De Miguel Aparicio, Elena & Palacios Alcaine, Azucena, & Serradilla Castaño, Ana Maria (Eds.), *Estructuras léxicas y estructuras del léxico*, 221–241. Berlin: Peter Lang.

Simone, Raffaele. 2007. Constructions and categories in verbal and signed languages. In Pizzuto, Elena & Pietrandrea, Paola & Simone, Raffaele (Eds.), *Verbal and Signed Languages*. *Comparing Structures, Constructs, and methodologies*, 198–252. Berlino-New York: Mouton De Gruyter.

Voghera, Miriam 1994. Lessemi complessi: percorsi di lessicalizzazione a confronto. *Lingua e Stile*, 29(2), 185–214.

Voghera, Miriam 2004. Polirematiche. In Grossmann, Maria & Rainer, Federica (a cura di), *La formazione delle parole in italiano*, 56–69. Tübingen: Max Niemeyer Verlag.

DOI: 10.17469/O2106SLI000026

LUCILLA PIZZOLI, MATTHIAS HEINZ

I vantaggi della ricerca su corpora per l'ampliamento e la verifica dei dati dell'OIM[1]

Nel contributo si intendono sottolineare le enormi potenzialità rappresentate dalla ricerca linguistica su corpora bilanciati rappresentativi dell'uso delle varietà linguistiche nelle quali sono stati censiti gli italianismi. Il controllo su corpora consente di rintracciare l'effettiva circolazione dei prestiti nell'uso e di misurarne il peso in modo più preciso rispetto a quanto documentato dalle fonti lessicografiche, nelle quali – soprattutto per il caso dei prestiti – possono intervenire fattori ideologici che sovrastimano o sottostimano l'impatto di una lingua straniera sulla lingua ricevente. Se il contributo di corpora sincronici risulta particolarmente prezioso nel caso dei neologismi, non sempre censiti in repertori lessicografici, anche l'indagine in prospettiva diacronica restituisce dati di grande interesse nell'indagine sugli scambi lessicali. La ricerca su corpora potrà dunque allargare la prospettiva di ricerca dei prestiti nelle tre direzioni di passato, presente e futuro.

Parole chiave: italianismi, corpora, lessicografia digitale, neologismi, arcaismi, archeologismi.

1. *Introduzione*

Non serve richiamare in questa sede come il fenomeno dei prestiti originati da parole italiane sia un tema di grande interesse, e che sia sempre più studiato da varie prospettive (anche in conseguenza della riscoperta del peso dell'italiano nel mondo, grazie alla constatazione dell'attrattività della lingua italiana come lingua di cultura, di studio, di lavoro)[2]. In quest'ambito si colloca il progetto OIM, che, come si

[1] Il presente testo è un prodotto collaborativo, la responsabilità redazionale per 1., 1.1, 1.3 e 2.1, 2.3 è di Lucilla Pizzoli e per 1.2 e 2.2 di Matthias Heinz.

[2] Si rimanda, per analoghe considerazioni, ai dati già richiamati in Pizzoli 2019; cfr. anche Mattarucco 2012 e Bombi & Orioles 2015, Bombi 2019; l'interesse per questo argomento è testimoniato dalla crescente mole di studi sulla diffusione dell'italiano nel mondo o in singole realtà geografiche e culturali. Basta vedere per esempio

potrà vedere, si può avvalere in modo particolarmente fruttuoso della ricerca su corpora.

1.1 Raccolte di prestiti nei repertori lessicografici

Partiamo da una considerazione generale: per lo più la ricerca di italianismi (e in generale di prestiti) si basa su estrazioni tratte dal lemmario di dizionari della lingua ospite, dal quale normalmente vengono selezionate le voci che hanno una marcatura specifica dal punto di vista etimologico, e quindi sono legate all'italiano come prestiti diretti, indiretti o calchi e quindi vengono isolati tramite questa trafila[3]. Tuttavia va precisato che la trafila etimologica può non essere ricostruita in modo omogeneo in tutte le lingue: l'accuratezza dei dati dipende dalla tradizione lessicografica di ogni lingua e soprattutto in molti dizionari potrebbe non essere aggiornato il comparto dei neologismi (perché prudentemente tenuti fuori dal lemmario in quanto frutto di contatto recente, legato a mode anche estemporanee); ancora, potrebbero esserci altri condizionamenti che determinano una

quanto pubblicato ogni anno nell'apposita sezione "L'italiano nel mondo" dalla rivista «Italiano LinguaDue» (https://riviste.unimi.it/index.php/promoitals/index/). Per i dati aggiornati sulla diffusione dell'italiano nel mondo si vedano i libri bianchi pubblicati a partire dal 2014 dal MAECI, a corredo delle edizioni degli *Stati generali della lingua italiana nel mondo*; 2014: *L'italiano nel mondo che cambia. Stati generali della lingua italiana nel mondo* (Firenze, Palazzo Vecchio, 21-22 ottobre 2014); 2016: *Italiano lingua viva. Stati generali della lingua italiana nel mondo* (Firenze, Palazzo Vecchio, 17-18 ottobre 2016); 2017: *L'italiano nel mondo che cambia. Stati generali della lingua italiana nel mondo* (Roma, Società Dante Alighieri, Palazzo Firenze, 18 ottobre 2017); 2018: *L'italiano nel mondo che cambia. Libro bianco degli Stati generali della lingua italiana nel mondo* (Roma, Villa Madama, 22 ottobre 2018); 2021: *L'italiano di domani*. *Stati Generali della Lingua e Creatività italiane nel Mondo* (Roma, Ministero degli Affari Esteri e della Cooperazione Internazionale – Sala Conferenze Internazionali, 29 novembre 2021).

[3] Tra le più recenti raccolte organiche, per esempio, si veda il ricco repertorio di italianismi in russo raccolto da Gherbezza 2019; si veda inoltre il DIFIT per inglese, tedesco e francese, e ancora Pinnavaia 2001 per l'inglese, Fábián-Szabó 2010 per l'ungherese, Gomez Gane 2012 per il catalano, Dashi 2013 per l'albanese, oltre a una cospicua mole di studi che indagano specifici ambiti semantici o epoche storiche, di cui è impossibile dare conto in questa sede e per la quale si rimanda alla raccolta bibliografica allestita proprio per il progetto OIM (consultabile online nel sito del progetto: https://www.italianismi.org/fonti).

attenzione ridotta – se non proprio l'esclusione – di alcune sezioni del lessico.

Nell'ottica di un osservatorio come l'OIM, concepito per misurare la profondità dell'impronta lessicale di una lingua sulle altre, conterà invece avere un quadro quanto più possibile esteso delle forme di contatto con la lingua ospite. Per questo motivo sarà utile avvalersi per la raccolta dei dati, oltre che delle banche dati lessicografiche, anche del tramite di altre fonti, non sempre prese in considerazione nella compilazione dei dizionari.

Per di più, all'interno del complesso paradigma della comunicazione nel mondo contemporaneo, hanno assunto un peso crescente le pratiche comunicative visibili nella sfera pubblica, in quella che viene classificata come "comunicazione sociale" e che include la fitta interazione tra gli individui per il tramite delle reti sociali e dell'ambiente esterno (Vespaziani 2018). In particolare, va notato che si riserva un'attenzione crescente, anche in linguistica, ai panorami linguistici urbani, proprio per monitorare la presenza di più lingue nel contesto multiculturale delle città[4]: i dati raccolti consentono dunque di misurare l'impronta semiotica di una o più lingue in un contesto specifico e dare conto del ruolo che rivestono le comunità linguistiche di antico o recente insediamento e degli equilibri che si vengono a creare tra diversi gruppi[5]. Altrettanto rilevante risulta, in questo quadro, il ruolo degli pseudoitalianismi, capaci di restituire l'immagine – sia pure semplificata e stereotipata – della lingua e della cultura italiana così come viene percepita nel mondo.[6]

[4] Gli studi sui *linguistic landscapes*, ben consolidati in area anglosassone (cfr. per es. Landry & Bourhis 1997 e Gorter 2006), sono stati applicati all'Italia e all'italiano grazie alle ricerche condotte nell'ambito dell'"Osservatorio linguistico nazionale dell'italiano diffuso fra stranieri e delle lingue immigrate in Italia" istituito presso l'Università per Stranieri di Siena nel 2000: cfr. Bagna & Barni 2006 e, da ultimo, per il panorama italiano negli ambienti urbani e scolastici, Bellinzona 2021.

[5] Indagini di questo tipo, avviate nell'Osservatorio senese, sono state documentate in studi concentrati su singole aree geografiche: cfr. per es. per il Giappone Vedovelli & Machetti 2006, per il Camerun Siebetcheu 2015.

[6] Per l'analisi di pseudoitalianismi in lingue e ambiti diversi si rimanda a Brincat 1991, Vedovelli 2005, Rieger 2008, Ille 2009, Vedovelli-Casini 2013, Casini 2015, 2017, 2018, Cassani 2015, Ferrini 2018 e ancora Siebetcheu 2015.

1.2 Un osservatorio sugli italianismi

Qui di seguito si darà qualche cenno sulla storia e lo sviluppo attuale del progetto di un osservatorio degli italianismi nelle lingue del mondo. L'idea nasce dalla pubblicazione nel 2008 in versione cartacea del *Dizionario degli italianismi in francese, inglese, tedesco* (DIFIT), concepito e redatto da H. Stammerjohann insieme a un gruppo internazionale di studiosi che ne curano le varie sezioni. Quasi nel contempo era in via di elaborazione un progetto editoriale di *Censimento* degli italianismi, curato da L. Serianni per l'editore UTET, che mirava alla documentazione di italianismi repertoriati in varie decine (oltre 70) di lingue del mondo intero. Gli autori del presente contributo erano già collaboratori dei rispettivi progetti. Del DIFIT si trae poi, per un complesso processo di retrodigitalizzazione, una prima versione in forma di banca dati interrogabile, messa a disposizione del pubblico dapprima all'interno del portale in rete VIVIT poi sul sito dell'*Osservatorio degli italianismi nel mondo* (OIM).[7]

Per ragioni di cambiamento di politica editoriale il progetto del *Censimento* non vede mai la luce in forma di opera a stampa (ne conseguono però alcuni trattamenti monografici basate sulle raccolte di italianismi come Gomez Gane 2012). Eppure, insieme con il nucleo dei dati del DIFIT una parte cospicua di tali materiali viene a costituire, sin dal 2014, la base della nuova piattaforma OIM, volta a fornire alla comunità scientifica uno strumento di lessicografia digitale atto a facilitare la ricerca sui prestiti passati dall'italiano in altre lingue. Definito come uno dei progetti strategici dell'Accademia della Crusca esso viene diretto da L. Serianni e M. Heinz e coordinato da L. Pizzoli, mentre la gestione della piattaforma informatica è affidata a M. Biffi e G. Salucci e l'elaborazione delle raccolte all'interno del gruppo di ricerca spetta a un numero crescente di unità operanti in varie sedi in Italia e all'estero.

Dalle tre lingue censite dal DIFIT si passa alle raccolte di italianismi in spagnolo, portoghese, catalano, polacco e ungherese mentre si stanno avviando i lavori di inserimento di ulteriori lingue di contatto quali il maltese, il neogreco, il macedone, il cinese mandarino e altre

[7] Il portale online VIVIT è stato progettato da F. Sabatini, N. Maraschio, D. De Martino, M. Biffi (https://www.viv-it.org/schede/crediti) mentre la banca dati DIFIT, opera di M. Biffi, G. Salucci, G. Seymer, M. Rago con la consulenza di H. Stammerjohann e M. Heinz (https://difit.italianismi.org/crediti), fa ormai parte dell'attuale sito dell'OIM (www.italianismi.org).

in parallelo alla revisione e estensione dei dati sulle lingue francese, tedesco, inglese, prendendone in considerazione varietà anche al di là dei confini nazionali.[8] Allo stato attuale sono presenti nella piattaforma di inserimento dati più di 12.500 voci in lavorazione di cui circa 1.500 con dati completi. Complessivamente sono più di 8.900 le voci provenienti dalle raccolte di italianismi del DIFIT ma va crescendo il numero di quelle provenienti da altre lingue accanto al numero delle voci di partenza ("etimi"). Le schede relative a quelle voci sono affidate alla redazione italiana che ne verifica la semantica e le trafile e le inserisce nella piattaforma.

1.3 La classificazione dei prestiti nell'OIM

Si è ritenuto utile registrare nell'OIM tutte le forme di contatto, anche rilevate attraverso fonti estemporanee (scritture esposte, etichette di prodotti commerciali, ecc.) o segnalazioni di singoli rilevatori, purché se ne possa garantire una minima diffusione nell'uso e non si tratti di meri occasionalismi.

Per assegnare il giusto valore al canale di raccolta utilizzato, nella piattaforma dell'OIM vengono classificate con una diversa marcatura le fonti usate per la raccolta, a seconda del loro livello di affidabilità: ovviamente il livello di maggiore autorevolezza viene assegnato alle banche dati lessicografiche (dizionari dell'uso, storici, specialistici e dei termini stranieri), considerate fonti "autorevoli" in quanto fondate su criteri scientifici e studi che documentino i dati raccolti; vengono poi censiti gli italianismi raccolti tramite fonti "divulgative" (opere tradotte dall'italiano, guide turistiche, libri di viaggio, trattati o studi di taglio divulgativo che descrivono aspetti specifici della cultura italiana, articoli giornalistici, ecc.), cioè fonti documentate che non hanno le caratteristiche di scientificità dei lavori di ricerca ma presentano comunque un margine di affidabilità. L'etichetta di fonti "estemporanee" si applica invece a tutte le voci provenienti da segnalazioni occasionali raccolte per mezzo di motori di ricerca o tramite l'osservazione diretta sul campo operata dai ricercatori e tutte le forme di rilevazione non tradizionale (come per l'appunto scritture esposte, insegne, menu

[8] Per il francese e l'inglese, come anche per lo spagnolo, hanno un interesse particolare le varietà extraeuropee, nelle quali molti italianismi si sono prodotti come frutto del contatto dell'italiano degli emigrati con le lingue locali.

di ristoranti, pubblicità di prodotti, ecc.), utili per monitorare fenomeni recenti e a volte recentissimi[9].

Tutte e tre le categorie individuate presentano dei limiti e dunque le raccolte che se ne ottengono non sono esenti da rischi, che varrà la pena di richiamare rapidamente.

I repertori lessicografici, ritenuti i più affidabili perché costruiti su ricerche condotte sulla base di criteri omogenei riscontrati in modo rigoroso, potrebbero essere influenzati dalla natura del corpus di partenza, non sempre rappresentativo della lingua nel suo complesso. Come sottolineava con qualche esempio Serianni (2008: 21-22), il dizionario potrebbe essere condizionato dalle caratteristiche specifiche delle singole raccolte lessicografiche, più sbilanciate sulla lingua scritta (e dunque poco rappresentative della diffusione di parole dell'uso vivo, anche recente): Serianni cita il caso del francese di Svizzera (ivi, p. 21, n.3), per il quale ci si basa sul *Dictionnaire suisse romand* (Thibault), fortemente incentrato sulla lingua scritta, specie amministrativa e politica. In altri casi il corpus di riferimento dei dizionari potrebbe essere falsato dall'esistenza precoce di repertori settoriali (è il caso di raccolte di termini musicali, come quelli pubblicati già nel Settecento per il francese; ancora Serianni 2008: 22); Ventura 2019 riporta l'esempio dell'inglese, arricchito da termini registrati in modo indebito da Florio e che hanno fatto entrare nei repertori inglesi voci tradotte da una presunta forma italiana mai usata[10]; *e converso*, molti repertori lessicografici difettano proprio di termini specialistici

[9] Sul tema della lingua italiana nel mercato globale si rimanda ai lavori avviati dal PRIN 2017 coordinato da Massimo Vedovelli su *Lingua italiana, mercato globale delle lingue, impresa italiana nel mondo: nuove dinamiche linguistiche, socioculturali, istituzionali, economico-produttive* (Università di Siena Stranieri; Università degli Studi di Firenze; Università degli Studi Internazionali di Roma – UNINT; Università Telematica IUL). Per l'italianità linguistico-culturale nella comunicazione commerciale cfr. ora Mori 2021.

[10] È il caso delle forme inglesi *countergabbion*, *to countergabbion*, traduzioni delle corrispondenti voci italiane *contra-gabbione*, *contragabbionare* (non registrate in italiano dal GDLI) e introdotte nell'OED con un'unica attestazione nella seconda edizione del *World of words* di John Florio (1611). Tali parole sono «da ascrivere piuttosto al fascino dell'italiano da cui il Florio si lascia talvolta trascinare. Il suo dizionario, prezioso perché notoriamente molto attento al lessico tecnico (e che nasceva proprio dall'intento di rendere leggibili per il pubblico inglese anche le opere italiane contemporanee non letterarie), va perciò attentamente vagliato sulla base di altre fonti, contemporanee e posteriori» (Ventura 2019: 192).

perché non hanno introdotto, nei corpora di riferimento, testi di carattere tecnico: è il caso per esempio delle voci di ambito militare (cfr. ancora Ventura 2019: 170).

Inoltre, la presenza o l'assenza di termini stranieri potrebbe essere influenzata da fattori esterni, in particolare di tipo ideologico: il repertorio potrebbe essere più o meno prescrittivo (più o meno orientato verso una qualche forma di difesa della purezza della lingua) e condizionato dalla considerazione dell'immagine che storicamente la lingua da cui provengono i termini ha guadagnato in un certo ambito (in altri termini, il valore simbolico di certe parole potrebbe portare a sovrastimarne l'impatto sulla lingua ospite, determinandone l'accoglimento nei relativi dizionari). Oppure, potrebbero essere sottostimati dati riguardanti termini di recente ingresso che pure potrebbero descrivere le tendenze in atto nelle relazioni tra l'Italia e un altro paese. Un interessante settore di indagine potrebbe essere per esempio quello della manifattura, nella quale come è noto l'Italia detiene un primato a livello europeo: le proposte di singole aziende potrebbero aver favorito l'esportazione di termini legati a prodotti di successo. Tali termini potrebbero non essere stati ancora registrati nei dizionari ed essere recuperabili solo tramite la consultazione di repertori terminologici di ambito settoriale o attraverso i cataloghi che descrivono prodotti brevettati in Italia e poi esportati all'estero.[11]

Infine, va considerato il fenomeno che Della Valle & Patota 2013 definiscono degli archeologismi, cioè parole che, a differenza degli arcaismi veri e propri, rappresentano dei "residui passivi", cioè termini mai circolati in una lingua (o circolati solo marginalmente) e che tuttavia vengono trasferiti in modo inerziale da un'edizione all'altra dei dizionari (anche di quelli sincronici, che dovrebbero invece fotografare la lingua realmente "in uso") senza che se ne controlli attentamente la reale consistenza. Questo fenomeno è probabilmente più accentuato nella lessicografia italiana, più restia, per il timore reverenziale nei confronti della tradizione letteraria, ad abbandonare le forme del passato. Tuttavia potrebbe verificarsi un fenomeno analogo anche nei dizionari delle lingue straniere, nei quali potrebbero continuare ad essere registrati come vitali prestiti ormai appartenenti a un momento della storia di quella lingua in cui il contatto con l'italiano era più stretto.

[11] Si rimanda per queste considerazioni a quanto già presentato in Pizzoli 2019, dove si propone, come esempio, il caso di *pergotenda*.

Naturalmente anche le fonti cosiddette "divulgative" potrebbero essere condizionate dal testo fonte, specie nel caso di traduzioni, o di guide turistiche o testi riguardanti l'Italia, e proporre parole italiane o calchi di espressioni non realmente diffuse nella lingua ospite, ma rese necessarie dall'esigenza di descrivere elementi culturospecifici. Analogamente, anche le fonti "occasionali" potrebbero raccogliere neologismi transeunti o veri e propri *hapax legomena* e devono dunque essere gestite con molta cautela.

Per ovviare a queste possibili deviazioni, potrà essere molto opportuno il riscontro su corpora bilanciati, che potrà restituire la reale consistenza della diffusione delle singole parole, anche monitorandone, dove possibile, il percorso nel breve e lungo periodo, un'esigenza sottolineata in più occasioni da tutti coloro che si sono misurati con la ricerca di prestiti italiani: oltre al già citato caso del lessico militare, si può ricordare quanto si è osservato per la presenza di italianismi nell'arte (Motolese 2012), nel commercio (Wilhelm 2013), nella musica (Bonomi & Coletti 2015).

I vantaggi del riscontro su corpora possono essere evidenziati soffermandosi in particolare su due – se non tre – possibili direzioni: nel passato (eliminare gli arcaismi) e nel presente (individuare i neologismi); ma forse anche verso il futuro (prevedere la stabilizzazione dei neologismi).

Per quanto riguarda il passato, controllare le parole su corpora che presentano testi distribuiti in diacronia ci consente di collocare al giusto posto ogni parola, fissando meglio le datazioni della prima attestazione. Individuare una retrodatazione è un lavoro che viene fatto di continuo e che ovviamente il ricorso a corpora di grande entità che comprendono testi antichi consente sempre di perfezionare. Ormai quasi un paio di decenni fa Luigi Matt proponeva, in modo piuttosto innovativo per l'epoca, la consultazione di corpora (in particolare del catalogo di SBN) per la retrodatazione di termini specialistici che avevano attestazioni tardive solo perché i testi che li registravano non erano stati considerati come campioni rappresentativi della lingua e dunque non comparivano nei repertori tradizionali ma si trovavano nei titoli di opere dei rispettivi settori disciplinari[12].

[12] Cfr. Matt 2004, che è riuscito a retrodatare oltre 1000 termini tecnici: «La relativa facilità con cui tale risultato è stato raggiunto permette di prevedere che un'applicazione a tappeto di questo metodo potrà fruttare un miglioramento sostanziale dei

Questo stesso controllo permette di monitorare l'arco di vita dei termini classificando eventualmente come arcaici o desueti termini non più attestati o scarsamente circolanti (e magari rimasti nei dizionari, anche sincronici, in modo inerziale). Documentare, anche per epoche passate, il rango di frequenza delle parole può inoltre servire a monitorare la vitalità dei termini e ricostruire la loro area di diffusione.

Per quanto riguarda il presente, il controllo su corpora sincronici consente di verificare l'esistenza e la consistenza della distribuzione nell'uso di parole che non sono state ancora censite nei dizionari di una lingua di interesse, magari lavorando in un'ottica comparata. Per esempio, partendo dal presupposto che alcuni termini potrebbero aver avuto una stessa trafila di diffusione in lingue geograficamente e culturalmente vicine, si può ipotizzare che la loro assenza in un repertorio dipenda dal fatto che la registrazione non è ancora stata completata ma che la parola sia comunque circolante. La conferma di questa ipotesi può essere garantita solo tramite il controllo su un corpus sincronico ampio e affidabile.

Dunque, a partire dalla lista di italianismi registrati in una lingua che può contare su una descrizione lessicografica particolarmente accurata, si potrebbe estendere il controllo su corpora rappresentativi di lingue che hanno avuto un analogo contatto con l'italiano per verificarne la presenza in queste lingue, specie ora che la globalizzazione rende così facili fenomeni di diffusione rapida di parole di moda.

Per quanto riguarda invece l'ultima possibile direzione, quella futura, il controllo su corpora di testi contemporanei consentirebbe di fare delle stime sulla stabilizzazione dei neologismi: l'analisi dell'andamento dei prestiti e del picco di frequenza, infatti, consentirebbe di ottenere elementi utili per valutarne la prospettiva di durata e l'opportunità di accoglierli nel lemmario di un dizionario.

Ancora una considerazione: il riscontro sistematico su corpora rappresentativi di testi autentici è in grado di restituire il dato del peso effettivo dell'italiano nella lingua ospite in modo molto più affidabile rispetto alla semplice registrazione lessicografica. Va ricordato infatti che i numeri di per sé non sono indicativi del contatto, né tantomeno della sua durata: molti italianismi potrebbero essere entrati in tem-

dati disponibili sui lessici settoriali» (p. 185). L'applicazione agli archivi elettronici è stata sperimentata anche da Telve 2002 per retrodatare voci onomatopeiche e interiettive in testi letterari. Cfr. ora anche Cortelazzo 2022.

pi remoti ed essere ora desueti (se non del tutto arcaici) nelle lingue ospiti, oppure risultare vitali solo in aree marginali della lingua[13]. Il controllo su corpora bilanciati e affidabili permette di assegnare all'italianismo un valore relativo rispetto alle diverse aree della lingua e misurare in modo preciso la diffusione del termine nella lingua ospite, eventualmente facendo un confronto più preciso anche della differenza tra una lingua all'altra (se i corpora sono confrontabili quantitativamente e qualitativamente).

Per l'OIM stiamo studiando un sistema di controllo che non si limiti solo al numero di prestiti registrati, ma consideri anche altre variabili: assegnando valori numerici fissati in base alla presenza della parola in un certo corpus e differenziando in base a parametri stabili si potrà misurare l'impatto dell'italiano sulle altre lingue[14]. In particolare, partendo dall'assunto che non tutti i prestiti valgono allo stesso modo, si potrebbe misurare il peso di variabili sia di tipo culturale (distinguendo tra prestiti legati alla vita privata e prestiti tecnici e culturali), sia di tipo più strettamente linguistico (analizzando cioè la presenza di termini concreti vs astratti, connotativi vs denotativi, permanenti vs effimeri).

Al numero di attestazioni (con la relativa percentuale nel campione di testi considerati) si potranno aggiungere dunque parametri di tipo culturale (rango di frequenza, ambito semantico, persistenza in diacronia, ecc.) e parametri di tipo linguistico (categorie grammaticali interessate dal contatto, livello di adattamento, produttività, ecc.). Il livello di penetrazione sarà massimo se la parola è entrata nel lessico di base, in categorie grammaticali meno effimere del nome (verbi, avverbi, ma anche affissi), se si è adattata e ha prodotto derivati (i cosiddetti "prestiti di secondo grado"); viceversa, l'impatto sarà meno rilevato se il prestito si limita ai nomi, si ferma alla periferia del repertorio di una

[13] Come è noto, Vidos 1965, ad esempio, elenca una serie di parole italiane passate ad altre lingue europee, soprattutto di àmbito marinaresco, commerciale, musicale basandosi su studi di inizio Novecento. Prima di registrare le voci come prestiti sincronici, andrà controllato se le voci sono tuttora citate da dizionari moderni; solo il controllo su corpora, però, confermerà se quelle voci sono oggi vitali, anche soltanto in àmbito tecnico; o se invece sono desuete (in tal caso ne dedurremo o che già le fonti di Vidos citavano le due forme per via libresca e inerziale o che queste erano vitali ancora ai primi del Novecento, ma sono oggi uscite d'uso).

[14] Rimandiamo alla proposta già abbozzata in Pizzoli 2017.

comunità (in un linguaggio specialistico), non presenta fenomeni di adattamento morfosintattico, ecc.[15]

2. *Il controllo sui corpora*

2.1 Esempi per lo spagnolo

Facendo un esperimento per lo spagnolo abbiamo provato a verificare la presenza di italianismi su due diversi corpora: quello offerto da Sketch engine (Spanish web 2018) e da Hemero[16]. Si può rilevare che la percentuale di occorrenze rispetto all'insieme delle parole raccolte nel corpus varia quando i numeri sono bassi. Per esempio, l'italianismo *arquitrabe*, un termine specialistico dell'arte attestato in spagnolo dal 1600 (cfr. DECH), risulta poco diffuso nella lingua comune ed è più soggetto ad oscillazioni nei valori percentuali nel confronto tra i due corpora (Sketch engine: 1853 (solo sing.) con una percentuale di 0.09 per million tokens; Hemero: 3 (2 sing; 1 pl. *arquitrabes*), con una percentuale di 0.00045 per million tokens). Uno scarto simile si ritrova anche in una parola come *cháchara* (dall'it. *chiacchiera*, attestato in spagnolo con i significati di "abbondanza di parole inutili", 1729; "conversazione frivola", 1551 e "oggetti di scarso valore", 1950: cfr. DECH; DLE 2014; NTLLE 2001): le occorrenze in Sketch engine sono 721 (sing.); + 277 (plur.), per una percentuale di 0.04 per million tokens + 0.01 per million tokens; in Hemero sono registrati 78 casi (73 sing; 5 pl. *chácharas*) per una percentuale di 0.12 per million tokens. Essendo un termine poco circolante, ha percentuali evidentemente casuali nei due corpora di controllo.

[15] Su problemi di questo tipo cfr. ancora quanto precisato da Serianni, che ha dimostrato come la qualità dell'apporto di una lingua vada valutata tenendo conto del comparto del lessico che viene coinvolto: limitando l'indagine al francese, una lingua pur profondamente legata all'italiano, si può misurare in poco più del 5% l'apporto totale degli italianismi al lessico francese fondamentale (Serianni 2008: 33).

[16] Sketch engine (https://auth.sketchengine.eu) è uno strumento che consente ricerche di tipo linguistico su corpora molto ampi di lingue diverse: uno dei corpus disponibile per lo spagnolo (Spanish web 2018) contiene testi provenienti da *Spanish Wikipedia, European Spanish web, American Spanish* raccolti tra febbraio e aprile 2018, per un totale di 17,553,075,259 parole; Hemero, purtroppo attualmente non più disponibile online, è un corpus di testi tratti da giornali spagnoli e ispanoamericani pubblicati tra il 1997 e il 2009, contenente oltre 660 milioni di parole.

Al contrario la parola *fiasco*, un termine di larghissima diffusione in molte lingue europee (sia con il significato di "bottiglia", sia con quello di "fallimento, insuccesso in uno spettacolo, specie teatrale", in spagnolo dal 1884, cfr. NTLLE), avendo numerose attestazioni mostra una percentuale più stabile: in Sketch engine: 28.694 attestazioni, per una percentuale di 1.41 per million tokens, in Hemero 900 attestazioni, per una percentuale di 1.36 per million tokens.

È inoltre particolarmente utile il riscontro su corpora per avere la conferma della scarsa o nulla circolazione di termini che già i dizionari definiscono desueti: nel DECH l'italianismo *fazoleto* viene definito «antic. y raro», e documentato dal 1611 in Covarrubias, con la precisazione che «Sólo este autor documenta el vocablo en castellano»[17].

Questo termine, prevedibilmente assente in entrambi i corpora di controllo, si può definire un vero e proprio archeologismo (già nel 1726 il *Diccionario de autoridades* della Real Academia Española precisa che «Tiene poco o nigun uso»): si tratta con ogni evidenza di una parola entrata per via lessicografica e poi rimasta senza una reale ragione nel repertorio.

2.2 Il caso di *quarantena*

Una parola il cui uso è aumentato sensibilmente dal 2020 in poi, in ovvio rapporto con la pandemia globale Covid19, è *quarantena*, il che vale per forme simili in varie lingue. Una vera e propria impennata delle occorrenze di *quarantena* in italiano si conferma ricorrendo ad una risorsa per la ricerche sul lessico giornalistico come *l'Archivio la Repubblica*. Così alle 1,967 occorrenze nel periodo compreso tra 1984 e inizio 2020 ne corrispondono ben 14,628 tra inizio 2020 e fine 2021.[18] Secondo studi recenti, una breve nota commento di Tomasin del 2020 cui fa seguito l'articolo dettgliato di Parenti/Tomasin 2021,

[17] L'informazione, per la quale ringraziamo Gloria Clavería Nadal, si ricava dal DECH, che ipotizza qualche spiegazione ulteriore: «de él pasó a Minsheu y a Aut., que lo declara sin uso. Covarr. explica que lo habían traído de Italia los cortesanos o los soldados. Para el origen de la voz italiana, controvertido, vid. Skok, ARom. V, 252-8; Gamillscheg, R.G. II, s. v.; Rohlfs, ASNSL CLXVIII, 257. Forma abreviada y usada realmente en España, pero sólo en germanía, es *fazo*, registrado por Juan Hidalgo (1609)».

[18] Simile l'aumento per forme analoghe come ted. *Quarantäne*, facilmente verificabile nei dizionari basati su corpora che offrono la funzionalità di visualizzare statistiche di frequenza (come DWDS, risorsa lessicografica per il vocabolario del tedesco).

sarebbe da escludere la pretesa origine veneziana del significato relativo a regolamenti sanitari[19], pur essendo stato coniato in ambito settentrionale nel sec. XVI e diffuso nel secolo successivo:

(1) "Nell'accezione sanitaria che ci interessa, il termine quarantena si è diffuso – verosimilmente a partire dall'italiano, nel corso del secolo XVII – in tutte le principali lingue europee, comprese quelle (come l'inglese, *quarantine*, o il tedesco, *Quarantäne*) nelle quali il legame con la parola che significa quaranta è ovviamente venuto meno." (Tomasin 2020)

La prima attestazione di questa accezione, 1630 secondo DELI, andrebbe retrodatata di vari decenni come dimostrano Parenti e Tomasin riassumendo l'intricata trafila etimologica del termine sanitario. Essa parte da forme del numerale collettivo *quarantena* (basate su forme latine contemporanee) già in uso per designare un intervallo di quaranta giorni (con evidenti rapporti al periodo canonico di purificazione e digiuno nella cultura giudaico-cristiana, cfr. 2021: 27, 32) per acquisire solo più tardi, con la pestilenza degli anni '70 del Cinquecento, il significato di '(periodo di) isolamento sanitario', nato con ogni probabilità in ambito settentrionale:

(2) "possiamo affermare che la voce può ben essersi formata nell'Italia settentrionale, perché anche in quest'area, come in Francia, erano in uso voci affini che possono essere chiamate in causa come premessa: anzitutto, come base di partenza, il numerale collettivo quarantena 'quarantina'" (2021: 33)

L'etimologia di *quarantena* nell'accezione sanitaria è alquanto discussa anche nella lessicografia delle singole lingue in cui esiste il termine. Per l'inglese *quarantine* OED propone una trafila etimologica che riferisce la forma ad una variante it. *quarantina* (accanto alla forma desueta *quarentina*), con rimando esplicito ad un *quarentena* di origine regionale con la precisazione "Venice" (Venezia):

(3) "In sense 4 probably < Italian *quarantina*, †*quarentina* (1630 in this sense, originally in the regional (Venice) form *quarentena*; a1311 denoting a set of forty (with reference to units of

[19] Il termine sinonimo veneziano è invece, come dimostrano i due studiosi con fior di esempi, *contumacia*, in uso corrente nei lazzaretti verosimilmente dalla seconda metà del sec. XVI (cfr. Parenti/Tomasin 2021: 24s.).

> time), 14th cent. denoting a period of forty days, originally specifically one set aside for penance; see below)." (OED, s.v.)

Il dizionario etimologico di Pfeifer (online, s.v.) menziona l'origine italiana delle varianti tedesche *Quarantena*, *Quarantia* (al plurale *Quarentennas*, *Quaranten*) che risalirebbero al primo Seicento mentre si sarebbe poi affermato, nel corso del Settecento, la forma francesizzante *Quarantaine* (oggi *Quarantäne*), soppiantando le forme di origine italiana. Ovviamente una tale ricostruzione etimologica si avvale di testimonianze scritte che la fanno sembrare l'unica plausibile. In questa prospettiva ted. *Quarantäne* sarebbe da considerare un italianismo indiretto, inoltrato, come avviene spesso nella storia del lessico tedesco, dal francese come lingua di mediazione, e così sarebbe da trattare anche dai dizionari dei prestiti.

In una risorsa lessicografica come l'OIM, che mira ad arricchire la documentazione degli italianismi lessicali, subentrano però funzionalità supplementari per dare conto del fatto che i fenomeni di contatto linguistico si manifestano, oltre le dimensioni del lessico e della semantica che interessano i prodotti stessi del contatto (prestiti), a più livelli sistematici della lingua (fonetica/morfosintassi/pragmatica). Infatti, al fine di offrire un quadro il più possibile completo del suo percorso diacronico, di ciascuna voce viene indicato, come parte integrante del corredo di informazioni microstrutturali, la trascrizione fonetica (con eventuali varianti foniche) e un file audio della pronuncia nella lingua di arrivo. In tedesco si notano due varianti fonetiche di *Quarantäne*, circoscritte nella loro diffusione diatopica. Da una parte esiste la variante principale [kaʀanˈtɛːnə] basata su una pronuncia francesizzante della sillaba iniziale di parola che si rifa a fr. [kaʀɑ̃ˈtɛn], diffusa in gran parte della Germania e percepita da molti parlanti come variante standard, mentre nell'area tedescofona meridionale, soprattutto nel dominio delle varietà bavaresi (sia della Germania che dell'Austria), è molto presente anche l'altra variante che invece dell'occlusiva velare sorda [k] inizia per il nesso della velare con la fricativa labiodentale sonora, [kvaʀanˈtɛːnə]. Nel maggiore dizionario di riferimento del tedesco moderno, DUDEN (s.v.), le varianti di pronuncia figurano di fianco, la seconda etichettata come "seltener" ('più rara') senza men-

zione di una distribuzione diatopica.[20] Una differenziazione areale che vale anche per le pronunce colte, con frequenza decisamente alta anche della seconda variante, specie nei parlanti oltre i 30 anni di età, è invece suggerita dalle inchieste recenti di Meier-Vieracker (2020) per le varietà della Germania e di Soukup (2021) per l'area austriaca. Essendo quest'ultima area linguistica anche in altri casi zona privilegiata del contatto con l'Italoromania, a differenza del resto della Germania[21], è altamente verosimile che qui la differenza di pronuncia non sia un effetto secondario (dovuto per es. alla grafia) bensì conseguenza delle diverse forme di partenza. Alla luce delle considerazioni sopraccitate di Parenti/Tomasin (2021) sono quindi da aggiornare le trafile etimologiche di ted. *Quarantäne*, ricostruite finora solo parzialmente dalla lessicografia tedesca:

(4) ted. standard [kaʀanˈtɛːnə] < fr. *quarantaine* (forse con precedenza dell'it. *quarantena*)

(5) ed. austriaco: [**kv**aʀanˈtɛːnə] < it. *quarantena* (forse con influsso a posteriori del fr. *quarantaine*)

I singoli fattori diatopici e sociolinguistici di una tale distinzione abbisognano senz'altro di futuri accertamenti attraverso i (rari) corpora di parlato con dettagliata annotazione sociofonetica. Prendendo in considerazione anche la fonetica degli adattamenti formali dei lemmi di partenza la banca dati OIM, che attinge a fonti lessicografiche, divulgative e corpora, si prefigge di documentare nelle sue schede i tratti sistematici degli italianismi nelle lingue di arrivo così da gettare le basi per ulteriori ricerche.

2.3 Il caso di *gabbione*

Per mostrare invece come può esser sfruttato un corpus di testi recenti per introdurre termini non ancora registrati dai dizionari, possiamo portare il caso di *gabbione*. In italiano la parola ha assunto il significa-

[20] DUDEN ne indica una terza, ancor più vicina a quella francese e ugualmente marcata 'più rara', con la vocale nasale [kaʀɑ̃ˈtɛnə], della quale però non si ha traccia in nessuno degli studi attuali. Sono inoltre pensabili realizzazioni variabili della vocale tonica (semiaperta o semichiusa).

[21] Cfr. ad es. la varietà di verdura chiamata *Melanzani* (< it. *melanzane*) nelle varietà dialettali di stampo bavarese-austriaco e *Aubergine* (< fr. *aubergine*) nelle altre varietà tedesche (inclusa quella standard; cfr. DWDS, s.vv.).

to specialistico nella terminologia militare di "elemento di fortificazione, costituito da un grosso cesto di vimini riempito di terra e sassi, usato per la costruzione di trincee e parapetti" e in quella idraulica di "elemento di difesa idraulica collocato a riparo di argini, ponti, scarpate fluviali e sim., costituito da un contenitore di rete metallica riempito di ciottoli e pietre" (GRADIT). Nel linguaggio recente il termine viene usato anche per designare la "gabbia in cui prendono posto gli imputati durante alcuni processi in tribunale" (ancora GRADIT; quest'ultima accezione non è registrata in Zingarelli 2022, nonostante sia l'uso più frequente nel linguaggio giornalistico: consultando l'archivio di Repubblica dal 1984 al 2021 si ottengono 154 occorrenze, prevalentemente concentrate in questo uso). Come prestito, però, la parola si è specializzata nella direzione della struttura portante riempita di sassi e ghiaia, usata con una certa frequenza specie per l'arredo dei giardini. In tedesco l'italianismo *gabbione* non è registrato fino a tempi recenti: non se ne trova traccia nel DWDS (che raccoglie testi fino al 1999); nel DIFIT era censito nella forma *Gabion* come prestito indiretto dal francese, dove aveva diversi significati (1. Fortificazione; 2. (region), Grande cesto con manici per il trasporto del letame, della terra. 3. sport / equit. Casotto per i cacciatori di selvaggina acquatica). In tempi recenti compare, come italianismo, in DUDEN, nella forma *Gabione*. La conferma che la parola ha acquisito una nuova vitalità, oltre che una nuova forma grafica, arriva proprio dal controllo su Sketch engine, che per il corpus tedesco (German web 2018, costituito da 5,346,041,196 parole) raccoglie oltre duemila occorrenze del prestito (343 sing. *Gabione* e 1758 pl. *Gabionen*), in gran parte concentrate in testi legati al giardinaggio.

Il successo di questa struttura e la sua graduale espansione anche al di fuori della Germania potrebbero far supporre un analogo radicamento anche in lingue di paesi vicini: *Gabion* ha oltre 7000 occorrenze nel corpus inglese (English Web 2020: 5004 sing., 2013 pl.); troviamo inoltre 1262 *gabion* in francese (French Web 2017), 3077 *gavión* in spagnolo (Spanish Web 2018), 677 *gabião* in portoghese (Portuguese Web 2011); 190 *gabion* in ungherese (Hungarian Web 2012). In queste ultime due lingue il numero più basso potrebbe essere collegato al fatto che i corpora raccolgono testi dei primi anni Dieci del XXI secolo, quando forse questo elemento di arredo non era ancora così diffuso.

Nell'OIM, per l'ungherese, la parola viene segnalata come prestito indiretto, mediato dal francese e dal tedesco, e se ne registra, accanto al significato abituale di fortificazione militare (indicato come neologismo nel 2010), anche con il significato più recente di elemento costruttivo usato specie nei recinti di giardini: la segnalazione arriva proprio per il tramite di fonti "estemporanee" (descrizioni storiche e – ultimamente – numerose occorrenze su Internet (p.es. https://www.pannongabion.hu/gyik) e di siti di venditori di materiali da costruzione)[22]: si potrebbe pensare a una nuova parola internazionale, utile a descrivere un'evoluzione moderna, forse un po' meno artistica, del tipico giardino all'italiana che decorava le ville rinascimentali.

Riferimenti bibliografici

Bagna, Carla & Barni, Monica. 2006. Per una mappatura dei repertori linguistici urbani: nuovi strumenti e metodologie. In De Blasi, Nicola & Marcato, Carla (a cura di), *La città e le sue lingue. Repertori linguistici urbani*. Napoli: Liguori. 1–43.

Bellinzona, Martina. 2021. *Linguistic landscape. Panorami urbani e scolastici nel XXI secolo*. Milano: FrancoAngeli.

Bombi, Raffaella & Orioles Vincenzo (a cura di). 2015. *Italiani nel mondo. Una Expo permanente della lingua e della cucina italiana*. Udine: Forum.

Bonomi, Ilaria & Coletti, Vittorio. 2015. *L'italiano della musica nel mondo*. Firenze: Accademia della Crusca.

Brincat, Giuseppe. 1991. Spigolando nel "Sunday Times": parole italiane e pseudoitalianismi nell'inglese di oggi. In Coveri, Lorenzo (a cura di), *L'italiano allo specchio. Aspetti dell'italianismo recente. Saggi di Linguistica Italiana. Atti del Primo Convegno della Società Internazionale di Linguistica e Filologia Italiana (Siena 28-31 marzo 1989)*. Torino: Rosenberg & Sellier. 7–14.

Casini, Simone. 2015. Italianismi e pseudoitalianismi nel mondo globale: il ruolo dell'enogastronomia. In Bombi, Raffaella & Orioles, Vincenzo (a cura di). *Italiani nel mondo. Una Expo permanente della lingua e della cucina italiana*. Udine: Forum. 89–102.

[22] L'informazione, raccolta da Zsuzsanna Fábián, è registrata nell'OIM (https://www.italianismi.org/scheda-italiano/gabbione/1888).

Casini, Simone. 2017. Italianismi e pseudoitalianismi a Toronto: una ricerca tra gli studenti di italiano del St. George Campus della University of Toronto. *Italica*. 94. 153–176.

Casini, Simone. 2018. Italianismi e pseudoitalianismi a Toronto: tra valori simbolici e prospettive di apprendimento. In Turchetta, Barbara & Vedovelli, Massimo (a cura di), *Lo spazio linguistico italiano globale: il caso dell'Ontario*. Pisa: Pacini. 225–254.

Cassani, Vanina. 2015. Italianismi e pseudoitalianismi a Londra: l'italiano tra vie tradizionali e innovative di diffusione. *Rivista Studi Italiani di Linguistica Teorica e Applicata* XLIV (1). 361–385.

Cortelazzo, Michele 2022. Retrodatazioni al "DELI" da traduzioni letterarie ottocentesche. *Studi di lessicografia italiana*. XXXIX, 247–312.

Covarrubias Orozco, Sebastián de. 1987. *Tesoro de la lengua castellana o española*, ed. de Martín de Riquer. Barcelona: Alta Fulla (rist. anast. dell'edizione del 1611).

Dashi, Brunilda. 2013. *Gli italianismi nella lingua albanese*. Roma: Nuova cultura.

DECH. 1980-1991. Coromines, Joan & Pascual, José Antonio. *Diccionario crítico etimológico castellano e hispánico*. Madrid: Gredos (edición en CD-ROM).

Della Valle, Valeria & Patota, Giuseppe. 2013. Residui passivi. Storie di archeologismi. *Studi di lessicografia italiana*. XXX. 133–164.

DIFIT. Harro Stammerjohann et al. (a cura di). 2008. *Dizionario di italianismi in francese, inglese, tedesco*. Firenze: Accademia della Crusca.

DLE. 2014[23] (y actualización). Real Academia Española. *Diccionario de la lengua española*. Madrid: Espasa-Calpe (https://dle.rae.es).

DRAE. 2003. *Diccionario de la lengua española*. 22ª ed. Madrid: Espasa-Calpe.

DUDEN. 1999. *Das große Wörterbuch der deutschen Sprache*, herausgegeben von Scholze-Stubenrecht, Werner unter Mitarbeit von Alsleben, Brigitte. Mannheim: Dudenverlag (https://www.duden.de/).

DWDS. *Digitales Wörterbuch der deutschen Sprache, Das Wortauskunftssystem zur deutschen Sprache in Geschichte und Gegenwart*, hrsg. v. d. Berlin-Brandenburgischen Akademie der Wissenschaften (https://www.dwds.de).

Fábián, Zsuzsanna & Szabó, Győző. 2010. *Dall'Italia all'Ungheria: parole di origine italiana nella lingua ungherese*. Udine: Forum Editrice.

Ferrini, Caterina. 2018. Italianismi e pseudoitalianismi nelle Little Italy di Toronto: il linguistic landscape come termometro per misurare la "febbre

da italiano". In Turchetta, Barbara & Vedovelli, Massimo (a cura di), *Lo spazio linguistico italiano globale: il caso dell'Ontario*. Pisa: Pacini. 255–292.

GDLI. 1961-2002. *Grande dizionario della lingua italiana*, fondato da Salvatore Battaglia. Torino: Utet.

Gomez Gane, Yorick. 2012. *Gli italianismi nel catalano. Dizionario storico-etimologico*. Roma: Aracne.

Gorter, Durk (ed). 2006. *Linguistic Landscape: A New Approach to Multilingualism*. Clevedon-Buffalo-Toronto: Multilingual Matters.

GRADIT. 1999-2003. *Grande dizionario italiano dell'uso*, diretto da Tullio De Mauro. Torino: Utet.

Heinz, Matthias (a cura di). 2017. *Osservatorio degli italianismi nel mondo: punti di partenza e nuovi orizzonti. Atti dell'incontro (Firenze, 20 giugno 2014)*. Firenze: Accademia della Crusca.

Ille, Karl. 2009. Italianismen und Pseudoitalianismen in der gastronomischen und kommerziellen Öffentlichkeit Wiens. In Ehmer, Josef & Ille, Karl (Hg.). *Italienische Anteile am multikulturellen Wien*. Innsbruck-Wien-Bozen: Studienverlag (Querschnitte, 27). 111–125.

Landry, Rodrigue & Bourhis, Richard Y. 1997. Linguistic landscape and ethnolinguistic vitality: an empirical study. *Journal of Language and Social Psychology* 16(1). 23–49.

Matt, Luigi. 2004. Retrodatazioni di tecnicismi da titoli di pubblicazioni. *Studi di lessicografia italiana* XXI. 183–246.

Mattarucco, Giada (a cura di). 2012. *Italiano per il mondo. Banca, commerci, cultura, arti, tradizioni*. Firenze: Accademia della Crusca.

Meier-Vieracker, Simon. 2020. Wie spricht man "Quarantäne" aus? Ergebnisse einer Umfrage. *Linguistische Werkstattberichte*. https://lingdrafts.hypotheses.org/1539

Mori, Laura, 2021. Italian Heritage come categoria interpretativa dell'italianità linguistico-culturale per l'analisi della comunicazione commerciale nel mercato economico internazionale. *Cultura&Comunicazione* XI (19). 36–48.

NTLLE. 2001. *Nuevo Tesoro Lexicográfico de la Lengua Española*, publicado por la Real Academia Española, 2 DVD. Madrid: Espasa Calpe.

OED. 1989[2]. *The Oxford English Dictionary*. by John A. Simpson & Edmund Weiner, 20 voll. Oxford: Clarendon press.

OIM e DIFIT: www.italianismi.org

Parenti, Alessandro & Tomasin, Lorenzo. 2021. Su Quarantena, preteso venezianismo, e su Contumacia. *Lingua nostra.* LXXXII/1-2. 23–33.

Pfeifer, Wolfgang (dir.). 1993². *Etymologisches Wörterbuch des Deutschen*. Berlin: Akademie-Verlag (online: https://www.dwds.de/d/wb-etymwb).

Pinnavaia, Laura. 2001. *The Italian Borrowings in the Oxford English Dictionary: A Lexicographical, Linguistic and Cultural Analysis*. Roma: Bulzoni.

Pizzoli, Lucilla. 2017. Per un dizionario degli italianismi nel mondo: rilancio di un progetto. *Testi e linguaggi* 11. 171–182.

Pizzoli, Lucilla. 2019. Italiano e italianismi nel mondo: osservazioni sulla ricerca di neologismi. In Bombi, Raffaella (a cura di), *Italiano nel mondo. Per una nuova visione*. Udine: Forum editrice. 151–158.

Repubblica Archivio (1/09/2021). https://ricerca.repubblica.it/repubblica/archivio/repubblica

Rieger, Marie A. 2008. 'Alles picco belli oder was?' Form und Funktion pseudo-italienischer Produktnamen im deutschen Lebensmittelmarkt. *Onoma*. 43. 149–175.

Serianni, Luca. 2008. Gli italianismi nelle altre lingue romanze: prime riflessioni, in *Italianismi e percorsi dell'italiano nelle lingue latine. Atti del Convegno (Treviso, 28 settembre 2007)*. Treviso-Paris: Fondazione Cassamarca-Unione Latina. 19–41.

Siebetcheu, Raymond. 2015. La lingua italiana nei panorami linguistici urbani delle città camerunensi. In Pallante, Gianna & Kuitche Talé, Gilles (a cura di), *20 anni d'insegnamento dell'italiano L2 in Camerun: bilancio e prospettive*. Italiano LinguaDue. 2. 59–70.

Soukup, Barbara. 2021. Geht Österreich in 'Karantäne' oder 'Kwarantäne'? Ergebnisse einer Umfrage. *Wiener Linguistische Gazette* 90. 265–307.

Telve, Stefano. 2002. Retrodatazioni di voci onomatopeiche e interiettive. Un esempio di applicazione lessicografica degli archivi elettronici. *Studi di lessicografia italiana* XIX. 229–277.

Tomasin, Lorenzo. 2020. Una *quarantena* può durare anche "solo" quattordici giorni. *Italiano digitale* XII/1. (online: https://id.accademiadellacrusca.org)

Vedovelli, Massimo & Casini, Simone. 2013. Italianismi e pseudoitalianismi in Giappone: le radici profonde di una consonanza culturale nel mondo globale. In Gesuato, Maria Katia & Peruzzi, Paola (a cura di), *La lingua italiana in Giappone 2*. Tokyo: Istituto italiano di cultura. 34–106.

Vedovelli, Massimo & Machetti, Sabrina. 2006. Italiano e lingue esotiche in contatto nella comunicazione sociale: il caso degli italianismi a Tokyo. In Banfi, Emanuele & Iannàccaro, Gabriele (a cura di), *Spazio linguistico italiano e le lingue esotiche: rapporti e reciproci influssi (Atti del XXXIX*

Congresso internazionale di studi della Società di linguistica italiana – SLI: Milano, 22-24 settembre 2005). Roma: Bulzoni. 181–206.

Vedovelli, Massimo. 2005. L'italiano nel mondo da lingua straniera a lingua identitaria: il caso 'freddoccino'. *Studi Italiani di Linguistica Teorica e Applicata* (XXXIV/3). 585–609.

Vedovelli, Massimo. 2014. L'italiano nel mercato globale delle lingue: prospettive, potenzialità, criticità. *Rapporto Italiani nel Mondo 2014*. 289–297.

Vespaziani, Alberto. 2018. Comunicazione Sociale. *Il Libro dell'Anno del Diritto Treccani*. 282–284.

Vidos, Benedek E. 1965. *Prestito, espansione e migrazione dei termini tecnici nelle lingue romanze e non romanze. Problemi, metodo e risultati*. Firenze: Olschki.

Wilhelm, Eva-Maria. 2013. *Italianismen des Handels im Deutschen und Französischen*. Berlin/Boston: De Gruyter.

Zingarelli 2022. *Vocabolario della lingua italiana*. Bologna: Zanichelli.

DOI: 10.17469/O2106SLI000027

ANDREA LISTANTI, LIANA TRONCI

Ordini di apprendimento di strutture VS in Italiano L2: Uno studio sul corpus LIPS

In italiano le costruzioni con soggetto post-verbale sono caratterizzate da diversi gradi di canonicità a seconda del tipo di verbo. La Teoria della Processabilità predice una sequenza di apprendimento dell'ordine VS in L2 basata sulla marcatezza sintattica e pragmatica. La sequenza prevede l'emergere del VS con verbi che lo selezionano come ordine pragmaticamente non marcato (Tipo 1), e la sua graduale estensione ad altri tre tipi di verbo (Tipi 2, 3, 4). Studi precedenti hanno testato l'ipotesi su un campione ristretto di apprendenti. Nel presente studio abbiamo analizzato produzioni di apprendenti di diversi livelli di competenza tratte dal corpus LIPS. I risultati rispecchiano le predizioni della Teoria della Processabilità nel caso dei verbi di tipo 1, 2 e 3, ma smentiscono le predizioni relative ai verbi di tipo 4. L'analisi qualitativa di queste strutture rivela un pattern di apprendimento che procede da un uso formulaico ad uno produttivo.

Parole chiave: italiano L2, soggetto post-verbale, teoria della processabilità, interfaccia sintassi-discorso, corpus LIPS.

1. *Introduzione*

Diversi studi sull'acquisizione dell'italiano come lingua seconda (L2) indagano lo sviluppo della sensibilità dell'apprendente nei confronti dei fattori lessicali e pragmatico-discorsivi che regolano la produzione di frasi con ordine dei costituenti non-canonico, cioè diverso da SV(O) (Salvi & Vanelli 2004: 297). In particolare, la produzione di frasi con soggetto posposto al verbo (VS) da parte di parlanti non nativi rappresenta un terreno di ricerca interessante, poiché le strutture VS in italiano non costituiscono un insieme omogeneo, ma implicano differenti gradi di marcatezza a seconda del tipo di verbo utilizzato.

In italiano la posizione pre-verbale è associata al *topic* e quella post-verbale al *focus* (Van Valin 2005; Belletti 2001, 2004). Di conseguenza, l'ordine VS è nella maggior parte dei casi pragmaticamente

marcato, in quanto il soggetto non ricopre la prototipica funzione topicale, bensì focale (Benincà *et al.* 1988). La relazione tra sintassi e struttura informazionale risulta evidente nell'uso del VS in costruzioni biargomentali in cui l'argomento diverso dal soggetto costituisce l'informazione già nota all'interlocutore, e il soggetto l'elemento nuovo del discorso:

(1) [chi ha comprato il libro?]
(il libro,) l'ha comprato Giacomo
#Giacomo ha comprato il libro

L'ordine VS è pragmaticamente marcato anche in diverse costruzioni monoargomentali, soprattutto con verbi inergativi. Come nel caso di (1), anche nell'esempio seguente la frase risulta appropriata soltanto all'interno di una situazione comunicativa che presuppone una domanda specifica:

(2) [chi ha parlato?]
ha parlato Giacomo
#Giacomo ha parlato

Tuttavia, in alcune costruzioni monoargomentali – soprattutto con verbi inaccusativi – l'ordine VS è appropriato anche in assenza di focus sul soggetto, cioè come risposta a domande generiche:

(3) [che cosa è successo?]
è arrivato Giacomo

Sebbene si discosti dall'ordine canonico SV(O), la frase (3) non risulta marcata dal punto di vista pragmatico, in quanto la posizione post-verbale del soggetto si deve alle sole proprietà lessicali del verbo (cf. Burzio 1986; Belletti 1988).

Facendo appello alla Teoria della Processabilità (*Processability Theory*, d'ora in avanti PT, cf. Pienemann 1998; Pienemann *et al.* 2005), possiamo individuare la sequenza di apprendimento del soggetto post-verbale in italiano L2 secondo l'ordine (inverso) delle costruzioni appena discusse. In altre parole, ogni apprendente (indipendentemente dalla propria L1) inizierà a produrre soggetti post-verbali come opzione pragmaticamente non marcata – come in (3) – e in seguito ne estenderà l'uso in modo progressivo a strutture sempre più complesse e marcate (cf. § 1.1). L'ipotesi è stata testata nello studio longitudinale di Bettoni *et al.* (2009) su due apprendenti adulti, e nel-

lo studio trasversale di Nuzzo (2015) su quindici bambini con diverse L1. I risultati di questi studi paiono confermare – seppur parzialmente – le predizioni della PT. Tuttavia, a detta stessa degli autori (cf. Bettoni *et al.* 2009: 169-170; Nuzzo 2015: 175), il campione di informanti e il numero di occorrenze analizzate risultano ancora troppo limitati per poter ricavare solide generalizzazioni.

Il presente studio intende testare le stesse ipotesi su un campione di apprendenti e occorrenze molto più ampio. I dati sono tratti dal corpus LIPS di italiano L2. Come vedremo in § 2.1, si tratta di uno strumento particolarmente adatto allo studio di sequenze di apprendimento, poichè i testi in esso contenuti sono divisi per livello di competenza.

1.1 Predizioni della PT

La PT afferma che la sequenza di apprendimento della sintassi in L2 dipende dall'allineamento tra ruoli tematici (agente/esperiente, paziente), funzioni grammaticali (soggetto, oggetto) e posizione dei costituenti all'interno della frase (Pienemann *et al.* 2005). I possibili allineamenti sono organizzati secondo una gerarchia di canonicità. In italiano l'allineamento canonico è SVO, con il soggetto/agente in posizione iniziale e l'oggetto/paziente in posizione finale. Di conseguenza, la PT predice che SVO emergerà per primo lungo la traiettoria acquisizionale degli apprendenti.

Successivamente, la capacità di produrre frasi con soggetto post-verbale emergerà in tempi diversi in base al grado di marcatezza sintattica o pragmatica. In base a quanto osservato in precedenza, essa dipende dal tipo di verbo utilizzato. La sequenza proposta da Bettoni *et al.* (2009) e Nuzzo (2015) è la seguente:

- *Verbi di tipo 1*: monoargomentali che selezionano VS come ordine pragmaticamente non marcato (i.e. inaccusativi; cf. (3)). L'apprendente deve aver maturato la consapevolezza che il soggetto può collocarsi in una posizione diversa da quella pre-verbale, e che le proprietà lessicali di questi verbi favoriscono la posizione post-verbale;
- *Verbi di tipo 2*: monoargomentali con SV come ordine canonico e con VS dettato da esigenze pragmatico-discorsive (i.e. inergativi; cf. (2)). L'apprendente deve essere consapevole che il soggetto può corrispondere al focus anziché al topic, e in questo caso occupa preferibilmente la posizione post-verbale;

– *Verbi di tipo 3*: biargomentali, con i quali due argomenti (anziché uno) occupano una posizione diversa rispetto all'allineamento canonico SVO per ragioni pragmatiche (i.e. transitivi; cf. (1)). L'apprendente deve avere acquisito l'uso del pronome clitico – come in (1) – o dei mezzi prosodici necessari a veicolare l'informazione non-nuova espressa dall'oggetto diretto (o da un altro argomento diverso dal soggetto).

Inoltre, come ultimo stadio di acquisizione del soggetto post-verbale in italiano L2, Bettoni *et al.* (2009) prevedono l'uso del VS con verbi "eccezionali" (*Tipo 4*). Si tratta di verbi come *piacere*, *divertire*, *interessare*, che presentano un allineamento doppiamente marcato, perché mappano il ruolo di esperiente sull'oggetto indiretto (spesso clitico; cf. (4)) in posizione pre-verbale, e il tema sul soggetto in posizione post-verbale (Pinker 1984):

(4) Gli interessa molto la storia d'Italia

Nuzzo (2015) non indaga la produzione del VS con questo tipo di verbi. I risultati del suo studio confermano una progressione nell'uso del soggetto post-verbale da verbi di tipo 1 a verbi di tipo 3, ma – data la scarsità di occorrenze – appaiono incerti rispetto al tipo 2. Per quanto riguarda lo studio longitudinale di Bettoni *et al.* (2009), la sequenza ipotizzata si riscontra in uno dei due apprendenti testati, mentre l'altro apprendente produce VS con tutti i tipi di verbo – compresi verbi di tipo 4 – già in corrispondenza del primo dei tre rilevamenti effettuati in ordine cronologico. Gli autori giustificano il risultato citando la relativa scarsità di dati e la possibilità che il livello di competenza degli apprendenti fosse già particolarmente avanzato all'inizio della rilevazione (Bettoni *et al.* 2009: 169).

Per ricavare un quadro maggiormente chiaro rispetto all'acquisizione del VS in italiano L2, nel presente studio indagheremo la produzione di soggetti post-verbali in un campione di 194 parlanti non-nativi adulti. Come nel caso di Nuzzo (2015), il nostro studio sarà trasversale e non longitudinale: per testare la sequenza di apprendimento proposta dalla PT analizzeremo l'uso dei quattro tipi di strutture VS da parte di apprendenti diversi, a differenti stadi di avanzamento nella conoscenza della lingua italiana.

2. *Lo studio*

2.1 Il corpus

I dati sono tratti dal corpus LIPS (Lessico Italiano Parlato da Stranieri; Vedovelli 2006), realizzato dall'Università per Stranieri di Siena. Il corpus contiene trascrizioni ortografiche di prove orali dell'esame CILS (Certificazione di Italiano come Lingua Straniera) svolte tra il 1993 e il 2006. Le trascrizioni sono divise per livello di competenza in base al Quadro Comune Europeo di Riferimento per le Lingue (da A1 a C2). In questo studio abbiamo analizzato i testi del 2002, cioè il primo anno contenente almeno una prova per ogni livello. Data l'esiguità del numero di prove relative ai livelli A1 e A2, abbiamo aggiunto al nostro dataset anche tutte le altre prove A1 e A2 presenti nel corpus. Nonostante questo, il numero di testi relativi a questi livelli rimane inferiore rispetto ai livelli successivi. Pertanto, le prove A1 e A2 sono state accorpate ed analizzate in un unico blocco (livello A).

2.2 Analisi e risultati attesi

Per la nostra analisi abbiamo considerato 289 testi orali prodotti da 194 apprendenti diversi. I testi selezionati in ciascun livello sono stati divisi in unità sintattiche, basate sull'occorrenza di un verbo finito (cf. Torregrossa *et al.* 2021). Nella nostra analisi, abbiamo considerato soltanto le unità con soggetto espresso in posizione post-verbale (VS). Il dataset è riassunto nella tabella seguente.

Tabella 1 - *Numero di informanti, trascrizioni, unità e occorrenze VS per ciascun livello di competenza*

	A	*B1*	*B2*	*C1*	*C2*
apprendenti	32	40	42	40	40
prove	45	60	61	60	63
unità	821	1759	2444	2255	2784
VS	66 (8%)	134 (7.6%)	155 (6.3%)	150 (6.6%)	148 (5.3%)

Le occorrenze di VS sono state annotate in base al tipo di verbo tra quelli menzionati in § 1.1:

– Tipo 1 (inaccusativi)
– Tipo 2 (inergativi)
– Tipo 3 (transitivi)
– Tipo 4 (tipo piacere)

In § 3 illustreremo la distribuzione dei quattro tipi di VS attraverso i cinque livelli di competenza considerati (A, B1, B2, C1, C2). La nostra analisi si basa sul rapporto percentuale tra ciascun tipo di struttura VS e il totale delle strutture VS prodotte nel livello corrispondente. Se le predizioni della PT sono esatte, ci aspettiamo che le strutture VS prodotte da apprendenti principianti (livello A) siano in larga parte associate a verbi di tipo 1, e che gli altri tipi di VS emergano in modo graduale nei livelli successivi. Inoltre, ci aspettiamo che la percentuale di VS di tipo 4 sia minima (o inesistente) nei livelli di competenza più bassi (A-B1), e massima nei livelli più alti (livelli C1-C2).

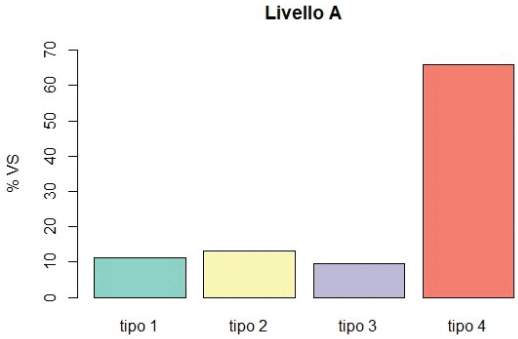
3. *Risultati*

La sezione dei risultati è divisa in sei sottosezioni. Le prime cinque descrivono la distribuzione dei VS in ciascun livello di competenza, dal più basso al più alto. La sesta è dedicata ad un'analisi qualitativa delle strutture VS con verbi di tipo 4.

3.1 Livello A

La Figura 1 mostra la distribuzione delle strutture VS prodotte da apprendenti di livello A in relazione al tipo di verbo. Contrariamente alle aspettative, gli apprendenti principianti producono in larga maggioranza VS con verbi di tipo 4 (66%). Il restante 34% si distribuisce uniformemente tra verbi di tipo 1 (11.3%), tipo 2 (13.2%) e tipo 3 (9.4%).

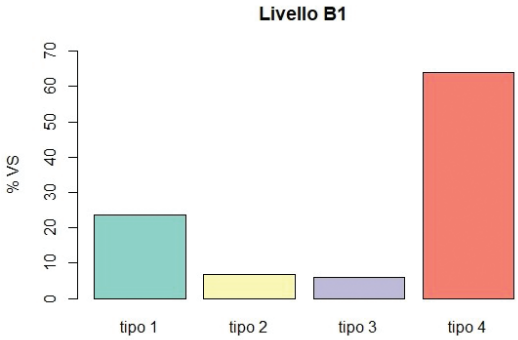
Figura 1 - *Percentuale di strutture VS associate a verbi di tipo 1, 2, 3 e 4 sul totale dei VS prodotti nel livello A*



3.2 Livello B1

La distribuzione delle strutture VS prodotte nel livello B1 non differisce molto rispetto al livello A. Come mostra la Figura 2, la percentuale di VS con verbi di tipo 1 aumenta (da 11.3% a 23.5%), ma la gran parte dei soggetti post-verbali rimangono associati ad un verbo di tipo 4 (63.9%).

Figura 2 - *Percentuale di strutture VS associate a verbi di tipo 1, 2, 3 e 4 sul totale dei VS prodotti nel livello B1*

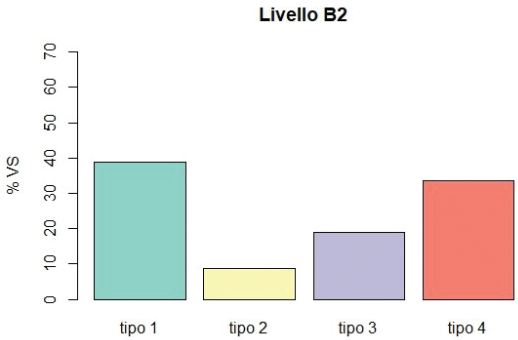


3.3 Livello B2

Al livello B2 aumenta ancora la percentuale dei VS con verbi di tipo 1 (da 23.5% a 38.8%), e si nota un aumento consistente di VS con verbi di tipo 3 (da 5.9% a 19%). Conseguentemente, si riduce l'altis-

sima percentuale di soggetti post-verbali associati a verbi di tipo 4 (da 63.9% a 33.6%), perché una loro quota "passa" ad altri tipi di verbo (cf. Figura 3).

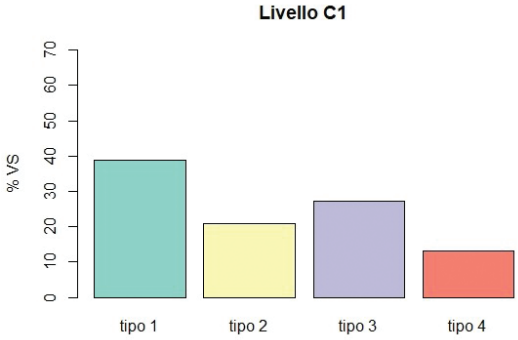
Figura 3 - *Percentuale di strutture VS associate a verbi di tipo 1, 2, 3 e 4 sul totale dei VS prodotti nel livello B2*



3.4 Livello C1

Al livello C1 si registra un aumento della percentuale dei VS con verbi di tipo 2 (da 8.6% a 20.9%) e di tipo 3 (da 19% a 27.1%), mentre la percentuale di VS con verbi di tipo 1 rimane stabile rispetto al livello precedente (38.8 %; cf. Figura 4). L'aumento dell'uso del soggetto post-verbale con verbi di tipo 2 e 3 a questo livello di competenza appare in linea rispetto alle ipotesi iniziali.

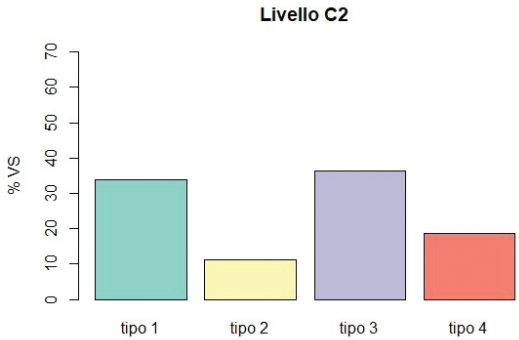
Figura 4 - *Percentuale di strutture VS associate a verbi di tipo 1, 2, 3 e 4 sul totale dei VS prodotti nel livello C1*



3.5 Livello C2

Il dato principale relativo al livello di competenza C2 è l'aumento della percentuale delle strutture VS con verbi di tipo 3 (da 27.1% a 36.3%). Si registra anche un leggero aumento della percentuale di VS con verbi di tipo 4 (da 13.2% a 18.5%). Conseguentemente, si riduce la percentuale dei VS con verbi di tipo 1 (33.8%) e con verbi di tipo 2 (11.3%).

Figura 5 - *Percentuale di strutture VS associate a verbi di tipo 1, 2, 3 e 4 sul totale dei VS prodotti nel livello C2*



3.6 Analisi qualitativa delle strutture VS con verbi di tipo 4

Il dato relativo alla distribuzione delle strutture VS con verbi di tipo 4 appare in netta contraddizione rispetto alle ipotesi iniziali. La PT predice che l'uso del soggetto post-verbale con questi verbi "eccezionali" emerga successivamente all'uso del VS con gli altri tipi di verbo. Pertanto, le predizioni della PT collocano l'emergere di questo tipo di struttura ad uno stadio particolarmente avanzato dell'apprendimento. Tuttavia, la nostra analisi ha rivelato che il VS con verbi di tipo 4 rappresenta circa due terzi dell'intera produzione di soggetti post-verbali da parte degli apprendenti ai livelli di competenza più bassi tra quelli considerati in questo studio (A e B1).

Per ricavare un quadro maggiormente chiaro rispetto a questo dato, le strutture VS con verbi di tipo 4 sono state distinte in base alle caratteristiche del costituente pre-verbale (esperiente) e del verbo

utilizzato. La Tabella 2 illustra la distribuzione di queste strutture attraverso i diversi livelli di competenza.

Al livello A, due terzi delle strutture VS con verbi di tipo 4 sono costituiti dall'espressione *mi piace X*, comprese due occorrenze in cui il pronome soggetto di prima persona singolare è comunque espresso (*io mi piace questo supermercato*; *io mi piace i film di avventura*, con mancanza di accordo di numero). In sette casi invece gli apprendenti di livello A producono VS con lo stesso pronome clitico *mi* seguito da un verbo diverso da *piacere* (per es. *mi basta*). Negli unici due casi in cui i principianti si allontanano dalla prima persona singolare o dal presente indicativo, commettono errori di accordo di numero (per es. *a lei piace i mosaici italiano*).

L'uso del VS nella forma *mi piace X* rappresenta più della metà delle occorrenze anche nei livelli B1 e B2. Tuttavia, in entrambi i casi si registra un aumento in termini percentuali nell'uso dell'espressione con una forma del verbo *piacere* diversa dalla prima persona singolare del presente indicativo (per es. *mi piaceva/piacevano*). Al livello B2 aumenta sensibilmente anche la percentuale di VS con un pronome clitico diverso da *mi* (per es. *gli piaceva*).

Infine, nei livelli di competenza più alti (C1-C2), l'uso del VS con verbi di tipo 4 è distribuito uniformemente tra *mi piace* e strutture sintatticamente analoghe con variazione nell'uso del clitico, del lemma e della flessione verbale (per es. *gli serviva*).

Tabella 2 - *Percentuale di strutture VS con verbi di tipo 4 divise per tipologia, attraverso i livelli di competenza*

	A (35)	B1 (76)	B2 (39)	C1 (17)	C2 (25)
Mi piace	23 (65.7%)	41 (53.9%)	21 (53.8%)	5 (29.4%)	8 (32%)
con flessione del verbo	3 (8.6%)	28 (36.8%)	9 (23.1%)	4 (23.5%)	10 (40%)
con altro verbo	7 (20%)	3 (3.9%)	4 (10.3%)	4 (23.5%)	3 (12%)
con altro clitico e verbo flesso	1 (2.9%)	4 (5.3%)	4 (10.3%)	1 (5.9%)	3 (12%)
con altro clitico e altro verbo flesso	1 (2.9%)	0	1 (2.6%)	3 (17.6%)	1 (4%)

4. *Discussione*

In questo studio abbiamo analizzato il corpus LIPS per indagare la produzione di frasi con soggetto post-verbale da parte di apprendenti di italiano a vari livelli di competenza. Lo scopo dell'indagine era verificare le predizioni della PT rispetto alla sequenza di apprendimento dell'ordine VS (verbo-soggetto) in italiano L2. La sequenza prevede che il VS emerga con verbi monoargomentali che selezionano questo ordine come pragmaticamente non marcato (verbi di tipo 1). Gli stadi di apprendimento successivi prevedono l'uso del VS in strutture monoargomentali in cui il soggetto segue il verbo quando corrisponde al *focus* del discorso (verbi di tipo 2), e in strutture biargomentali in cui entrambi gli argomenti appaiono dislocati rispetto alla loro posizione canonica (oggetto pre-verbale, soggetto post-verbale) per esigenze pragmatico-discorsive (verbi di tipo 3). Infine, l'ultimo stadio di apprendimento è costituito dall'uso del VS con verbi "speciali" che mappano l'esperiente sull'oggetto indiretto in posizione pre-verbale, e il tema sul soggetto in posizione post-verbale (verbi di tipo 4).

La nostra analisi si è basata sulla percentuale di strutture VS associate a ciascun tipo di verbo rispetto al totale delle strutture VS prodotte dagli apprendenti in ognuno dei cinque livelli considerati (A, B1, B2, C1, C2). Il primo dato da registrare è che tutti i livelli – compresi i più bassi – contengono almeno un'occorrenza per ogni tipo di VS. Non si tratta tuttavia di un dato particolarmente rilevante, data la natura trasversale del nostro studio. In assenza di dati longitudinali relativi ai singoli apprendenti, non possiamo escludere la possibilità che gli apprendenti che hanno prodotto strutture VS di tipo "avanzato" nei livelli più bassi di competenza non avessero comunque attraversato tutti gli stadi di sviluppo previsti dalla PT. Tuttavia, l'analisi della distribuzione delle diverse strutture VS attraverso i cinque livelli rivela alcune tendenze generali interessanti.

Il dato apparentemente più sorprendente riguarda le strutture VS con verbi di tipo 4. La PT predice che queste strutture emergano soltanto negli stati più avanzati dell'apprendimento, per via dell'allineamento non-canonico tra ruoli tematici, funzioni grammaticali e ordine dei costituenti all'interno della frase. Tuttavia, queste strutture rappresentano il tipo dominante nei livelli A e B1, con percentuali che superano la somma degli altri tre tipi. L'analisi qualitativa di queste occorrenze ha rivelato che tra gli apprendenti di livello A,

nella quasi totalità dei casi l'impiego di questa struttura corrisponde all'uso dell'espressione *mi piace X*, o espressioni analoghe con lo stesso pronome clitico dativo di prima persona singolare (per es. *mi serve X*, *mi basta X*). Sebbene in assenza di studi specifici su corpora di L1, possiamo ipotizzare che la struttura *mi piace X* abbia natura formulaica e alta frequenza nell'*input*. La PT individua nell'uso di formule fisse (e singole parole) una caratteristica universale della prima fase dell'apprendimento di una lingua seconda, definita pertanto "fase lessicale" (cf. Bettoni & Di Biase 2011). In questa fase iniziale le strutture utilizzate dall'apprendente non sono ancora grammaticalizzate, e vengono organizzate nel discorso secondo principi esclusivamente pragmatici. Nel nostro caso, una spia del fatto che le strutture VS con verbi di tipo 4 prodotte da apprendenti principianti corrispondono in realtà a formule non analizzate dal punto di vista sintattico è rappresentata dal tentativo di inserire la formula stessa all'interno di frasi che riproducono l'ordine basico SVO (per es. *io mi piace questo supermercato*), e dall'assenza di accordo di numero tra verbo e soggetto/tema post-verbale (per es. *mi piace i film d'avventura*; cf. § 3.6). Come mostrato nella Tabella 2 (§ 3.6), al crescere del livello di competenza degli apprendenti, aumenta l'uso produttivo di queste strutture. Il processo si nota già nel passaggio dal livello A al livello B1. Sebbene la quasi totalità di queste costruzioni rimanga rappresentata dalla formula *mi piace*, in circa un terzo dei casi gli apprendenti di livello B1 utilizzano una flessione diversa del verbo *piacere* (per es. *mi piacciono le chiese*; *mi piacerebbe qualcosa con le lingue*). In otto casi questo tentativo è associato ad errori di accordo di numero (per es. *mi piacevano sempre una compagnia piccola*). Il dato suggerisce che a questo stadio di apprendimento la struttura non è ancora completamente analizzata. Ancora al livello B2, *mi piace* rimane la struttura maggiormente usata, ma gli errori di accordo si riducono a tre occorrenze, e gli apprendenti fanno un uso più consistente di altri clitici e di altri verbi (per es. *gli interessano le stesse cose*). Questa tendenza viene consolidata ai livelli C1 e C2. In questi livelli, il processo di grammaticalizzazione di queste strutture appare completo. È pertanto ipotizzabile che a questi stadi di acquisizione affiori negli apprendenti la consapevolezza della natura marcata di queste costruzioni.

Per quanto riguarda l'ordine VS con gli altri tre tipi di verbo, i nostri risultati appaiono in linea rispetto alle ipotesi iniziali. In parti-

colare, il consolidamento dell'uso del VS con verbi di tipo 1 precede il consolidamento dell'uso del VS con verbi di tipo 2 e tipo 3. La tendenza è testimoniata dal fatto che la percentuale di queste occorrenze rispetto al totale dei VS prodotti aumenta tra A e B2, per poi stabilizzarsi nei livelli successivi. Al contrario, l'aumento significativo nell'uso del VS con verbi di tipo 2 avviene al livello C1, nel quale la percentuale di queste strutture rispetto al totale supera il 20%. Infine, l'uso del VS con verbi di tipo 3 aumenta in maniera costante attraverso tutti i livelli. Al massimo livello di competenza (C2), queste costruzioni rappresentano la maggioranza relativa di tutte le strutture VS prodotte (36.3%). In generale, la progressione nell'uso del VS in costruzioni inaccusative, inergative e transitive nel passaggio tra i livelli B, C1 e C2 conferma le predizioni della PT e offre maggiore supporto empirico ai risultati già individuati da Bettoni *et al.* (2009) e Nuzzo (2015).

I nostri dati suggeriscono che l'emergere del VS nel suo uso formulaico *mi piace X* nei livelli più bassi di competenza svolga il ruolo di *pivot* per uno schema sintattico che nei livelli successivi viene sviluppato in due direzioni. Da un lato, esso conduce al riconoscimento progressivo del costituente post-verbale come soggetto, e di conseguenza all'uso produttivo del VS nelle costruzioni inerentemente più complesse (con verbi di tipo 4). Dall'altro lato, lo schema viene progressivamente esteso ad altri tipi di verbo (cf. Ellis 2012).

Riferimenti bibliografici

Belletti, Adriana. 1988. The case of unaccusatives. *Linguistic inquiry* 19(1). 1-34.

Belletti, Adriana. 2001. "Inversion" as focalization. In Hulk, Aafke & Pollock, Jean-Yves (a cura di), *Subject inversion in Romance and the theory of Universal Grammar*, 60-90. Oxford: Oxford University Press.

Belletti, Adriana. 2004. Aspects of the low IP area. In Rizzi, Luigi (a cura di), *The structure of CP and IP: The cartography of syntactic structures, Vol. 2*, 16-51. Oxford: Oxford University Press.

Benincà, Paola & Salvi, Sergio & Frison, Lorenza. 1988. L'ordine degli elementi della frase e le costruzioni marcate. In Renzi, Lorenzo & Salvi, Sergio & Cardinaletti, Anna (a cura di), *Grande grammatica italiana di consultazione, Vol. 1,* 129-239. Bologna: il Mulino.

Bettoni, Camilla & Di Biase, Bruno. 2011. Beyond canonical order: The acquisition of marked word orders in Italian as a second language. *EUROSLA Yearbook* 11. 244-272.

Bettoni, Camilla & Di Biase, Bruno & Nuzzo, Elena. 2009. Postverbal subject in Italian L2: a processability theory approach. In Keßler, Jörg.U. & Keatinge, Dagmar (a cura di), *Research in Second Language Acquisition: Empirical Evidence across Languages*, 153-174. Cambridge: Cambridge Scholars.

Burzio, Luigi. 1986. *Italian syntax: A government-binding approach*. Dordrecht: Reidel.

Ellis, Nick C. 2012. Formulaic language and second language acquisition: Zipf and the phrasal teddy bear. *Annual review of applied linguistics* 32. 17-44.

Nuzzo, Elena. 2015. Ipotesi di sviluppo di ordini sintattici marcati in giovanissimi apprendenti di italiano L2. In Chini, Marina (a cura di), *Il parlato in (italiano) L2: Aspetti pragmatici e prosodici*, 166-176. Milano: FrancoAngeli.

Pienemann, Manfred. 1998. *Language Processing and Second Language Development: Processability Theory*. Amsterdam: John Benjamins.

Pienemann, Manfred & Di Biase, Bruno & Kawaguchi, Satomi. 2005. Extending Processability Theory. In Pienemann, Manfred (a cura di), *Cross-linguistic Aspects of Processability Theory*, 199-251. Amsterdam: John Benjamins.

Pinker, Steven. 1984. *Language Learnability and Language Development*. Cambridge, MA: Harward University Press.

Salvi, Sergio & Vanelli, Laura. 2004. *Nuova grammatica italiana*. Bologna: il Mulino.

Torregrossa, Jacopo & Andreou, Maria & Bongartz, Chris & Tsimpli, Ianthi Maria. 2021. Bilingual Acquisition of Reference. The role of language experience, executive functions and cross-linguistic effects. *Bilingualism: Language and Cognition* 24(4). 694-706.

Van Valin, Robert D. 2005. *Exploring the Syntax-Semantics Interface*. Cambridge: Cambridge University Press.

Vedovelli, Massimo. 2006. Il LIPS – Lessico di frequenza dell'italiano parlato dagli stranieri. In Bardel, Camilla & Nystedt, Jane (a cura di). *Progetto dizionario italiano-svedese. Atti del primo colloquio, Stoccolma, 10-12 febbraio 2005*, 55-78. Stoccolma: Intellecta Docusys.

Autrici e autori

Sandra Maria Aluísio – NILC, USP, São Carlos, sandra@icmc.usp.br

Silvia Ballaré – Dipartimento di Filologia classica e italianistica, Università di Bologna, Bologna, silvia.ballare@unibo.it

Manuel Barbera – Dipartimento di Lingue e letterature straniere e Culture moderne, Università di Torino, Torino, manuel.barbera@unito.it

Tony Berber Sardinha – LAEL, PUC-SP, São Paulo, tonycorpuslg@gmail.com

Marco Biffi – Dipartimento di Lettere e Filosofia, Università di Firenze, marco.biffi@unifi.it

Giorgina Cantalini – Civica Scuola di Teatro Paolo Grassi, Milano, giorgina@giorginacantalini.com

Flavio M. Cecchini – CIRCSE, Università Cattolica del Sacro Cuore, Milano, flavio.cecchini@unicatt.it

Chiara Celata– Dipartimento di Studi Umanistici, Università di Urbino 'Carlo Bo', Urbino, chiara.celata@uniurb.it

Massimo Cerruti – Dipartimento di Studi umanistici, Università di Torino, Torino, massimosimone.cerruti@unito.it

Francesca Cialdini – Dipartimento di Studi Linguistici e Culturali, Università di Modena e Reggio Emilia, francesca.cialdini@unimore.it

Federica Cominetti – Dipartimento di Lingue, Letterature e Culture Straniere, Università Roma Tre, Roma, federica.cominetti@uniroma3.it

Elisa Corino – Dipartimento di Lingue e letterature straniere e Culture moderne, Università di Torino, Torino, elisa.corino@unito.it

Emanuela Cresti – Dipartimento di Lettere e Filosofia, Università di Firenze, emanuela.cresti@unifi.it

Paolo D'Achille – Dipartimento di Studi Umanistici, Università Roma Tre, Roma, paolo.dachille@uniroma3.it

Mark Davies – Linguistics Department, Brigham Young University, Provo, mark@mark-davies.org

Anna-Maria De Cesare – Institut für Romanistik, Technische Universität Dresden, Dresden, anna-maria.decesare@tu-dresden.de

Francesca M. Dovetto – Dipartimento di Studi Umanistici, Università di Napoli Federico II, Napoli, dovetto@unina.it

Magali Duran – NILC, USP, São Carlos, magali.duran@gmail.com

Angela Ferrari – Istituto di italianistica, Università di Basilea, angela.ferrari@unibas.ch

Cláudia Freitas – Departamento de Letras, PUC-RJ, Rio de Janeiro, claudiafreitas@puc-rio.br

Charlotte Galves-Chambelland – Departamento de Linguística, UNICAMP, Campinas, charlotte.mgc@gmail.com

Vittorio Ganfi – Dipartimento di Scienze Mediche e Chirurgiche Materno-Infantili e dell'Adulto, Università di Modena e Reggio Emilia, Modena e Reggio Emilia, vittorio.ganfi@uniroma3.it

Mariafrancesca Giuliani – CNR, Opera del Vocabolario Italiano, Firenze, giuliani@ovi.cnr.it

Eugenio Goria – Dipartimento di Studi umanistici, Università di Torino, Torino, eugenio.goria@unito.it

Lorenzo Gregori – Dipartimento di Lettere e Filosofia, Università di Firenze, lorenzo.gregori@unifi.it

Raffaele Guarasci – ICAR-CNR, guarasci@gmail.com

Alessia Guida – Dipartimento di Studi Umanistici, Università di Napoli Federico II, Napoli, alessia.guida1@gmail.com

Matthias Heinz – Fachbereich Romanistik, Universität Salzburg, matthias.heinz@plus.ac.at

Claudio Iacobini – Dipartimento di Studi Umanistici, Università di Salerno, Fisciano (SA), ciacobini@unisa.it

Iørn Korzen – Department of Management, Society and Communication, Copenhagen Business School, iorn.bjarne@gmail.com

Anne Lacheret-Dujour – Université de Nanterre, MODYCO, UMR 7114, Nanterre, France, anne.lacheret@parisnanterre.fr

Letizia Lala – Sezione di italiano, Università di Losanna, letizia.lala@unil.ch

Sidney Leal – NILC, USP, São Carlos, sidleal@gmail.com

Andrea Listanti – Università per Stranieri di Siena, andrea.listanti@gmail.com

Edoardo Lombardi Vallauri – Dipartimento di Lingue, Letterature e Culture Straniere, Università Roma Tre, Roma, edoardo.lombardivallauri@uniroma3.it

Paola Manni – Dipartimento di Lettere e Filosofia, Università degli Studi di Firenze, Firenze, paola.manni@unifi.it

Carla Marello – Dipartimento di Lingue e letterature straniere e Culture moderne, Università di Torino, Torino, carla.marello@unito.it

Philippe Martin – LLF, UFRL, Université Paris Cité, Paris, philippe.martin@linguist.univ-paris-diderot.fr

Takehiko Maruyama – Department of International Communication, Senshu University, Tokyo, maruyama@isc.senshu-u.ac.jp

Caterina Mauri – Dipartimento di Lingue, letterature e culture moderne, Università di Bologna, Bologna, caterina.mauri@unibo.it

Heliana Mello – Faculdade de Letras, UFMG, Belo Horizonte, hmello@ufmg.br

Massimo Moneglia – Dipartimento di Lettere e Filosofia, Università di Firenze, massimo.moneglia@unifi.it

Rossella Mosti – CNR, Istituto "Opera del Vocabolario Italiano", Firenze, mosti@ovi.cnr.it

Naomi Nagy– Department of Linguistics, University of Toronto, Toronto, naomi.nagy@utoronto.ca

Carlota Nicolás – Dipartimento di Lettere e Filosofia, Università di Firenze, carlota.nicolas@unifi.it

Miguel Oliveira – Faculdade de Letras, UFAL, Maceió, miguel@fale.ufal.br

Cristina Onesti – Dipartimento di Lingue e letterature straniere e Culture moderne, Università di Torino, Torino, cristina.onesti@unito.it

Anna Chiara Pagliaro – Dipartimento di Studi Umanistici, Università di Napoli Federico II, Napoli, pagliaroannachiara@gmail.com

Alessandro Panunzi – Dipartimento di Lettere e Filosofia, Università di Firenze, alessandro.panunzi@unifi.it

Thiago Pardo – NILC, USP, São Carlos, taspardo@icmc.usp.br

Marco Passarotti – CIRCSE, Università Cattolica del Sacro Cuore, Milano, marco.passarotti@unicatt.it

Filippo Pecorari – Istituto di italianistica, Università di Basilea, filippo.pecorari@unibas.ch

Matteo Pellegrini – CIRCSE, Università Cattolica del Sacro Cuore, Milano, matteo.pellegrini@unicatt.it

Paola Pietrandrea – Université de Lille & CNRS UMR8163 STL, Lille, France, paola.pietrandrea@univ-lille.fr

Valentina Piunno – Dipartimento di Filosofia Comunicazione e Spettacolo, Università degli Studi Roma Tre, Roma, valentina.piunno@uniroma3.it

Lucilla Pizzoli – Facoltà di Interpretariato e Traduzione, Università degli Studi Internazionali di Roma - UNINT, Roma, lucilla.pizzoli@unint.eu

Lucia Raggio – Dipartimento di Studi Umanistici, Università di Napoli Federico II, Napoli, lucia.raggio@libero.it

Tommaso Raso – Faculdade de Letras, UFMG, Belo Horizonte, tommaso.raso@gmail.com

Assunta Sorrentino – Dipartimento di Studi Umanistici, Università di Napoli Federico II, Napoli, sundrasorrentino@gmail.com

Rachele Sprugnoli – DUSIC, Università degli Studi di Parma, rachele.sprugnoli@unipr.it

Fabio Tamburini – Dipartimento di Filologia Classica e Italianistica, Università di Bologna, fabio.tamburini@unibo.it

Mirko Tavoni – Accademia della Crusca, mirko.tavoni@gmail.com

Simona Trillocco – Dipartimento di Lettere e Filosofia, Università di Firenze, simona.trillocco@unifi.it

Liana Tronci – Università per Stranieri di Siena, tronci@unistrasi.it

Giulio Vaccaro – Istituto di Storia dell'Europa Mediterranea-CNR, Roma, giulio.vaccaro@isem.cnr.it

Gu Yueguo – The Chinese Academy of Social Sciences, Artificial Intelligence and Human Language Research Lab, BFSU, gyg@beiwaionline.com

finito di stampare
nel mese di novembre 2022
presso la Litografia Solari
Peschiera Borromeo (MI)

Questo volume raccoglie una selezione delle relazioni presentate al LIV Congresso Internazionale di Studi della Società di Linguistica Italiana "Corpora e Studi Linguistici", tenutosi ***online*** dall'8 al 10 settembre 2021 attraverso la piattaforma ***Underline***. Il Congresso, organizzato dall'Università di Firenze e dall'Accademia della Crusca, ha fornito dimostrazioni delle esperienze italiane e internazionali nel campo della costituzione e annotazione dei corpora e ha proposto studi linguistici sincronici e diacronici ***corpus-based***. I contributi sono stati sottoposti a doppia revisione anonima.

Emanuela Cresti, già professore ordinario di Grammatica italiana all'Università di Firenze, è Accademico corrispondente dell'Accademia della Crusca e collabora con il Laboratorio LIU dell'Università di Firenze. Si occupa di raccolta e analisi di corpora e ha sviluppato il quadro teorico per lo studio dell'orale noto come ***Teoria della Lingua in Atto***.

Massimo Moneglia è professore associato di Linguistica e di Semantica e Lessicologia all'Università di Firenze, si occupa di semantica e linguistica dei corpora e si è dedicato alla formazione di infrastrutture per la ricerca linguistica presso il Laboratorio LIU dell'Università di Firenze.

9 788897 657552