Language corpora of spontaneous speech in healthy and pathological ageing

Francesca M. Dovetto
(dovetto@unina.it)

Sheet for corpus DEMO

Keywords *pathological language - spontaneous spoken Italian - language disintegration*

Nome del corpus o dei corpora / Acronimo ***Corpus or corpora name(s) / Acronym(s)***	*Corpora di parlato spontaneo, normofasico e patologico, dell'età adulta e senile.* CIPPS, CIPP-ma, CIPP-mci *Language corpora of spontaneous speech, in healthy and pathological ageing.* CIPPS. CIPP-ma, CIPP-mci
Tipologia del corpus / corpora ***Corpus or Corpora typology***	corpora orali di parlato italiano, normofasico e patologico
Periodo storico a cui si riferiscono i dati ***Historical period of collected data***	CIPPS 2005-2007 (registrazioni); 2007-2012 (trascrizioni) CIPP-ma 2016-2017; 2019-... CIPP-mci 2018-...
Modalità di accesso e distribuzione alle risorse ***Access to the resources***	CIPPS DVD in Dovetto-Gemelli, *Il parlar matto. Schizofrenia tra fenomenologia e linguistica. Il Corpus CIPPS*, Roma, Aracne 2013 (2.nda ediz) CIPP-ma *in corso di acquisizione* CIPP-mci *in corso di acquisizione* I corpora trascritti ortograficamente e annotati saranno archiviati presso il Laboratorio LiSa, *Lingua e salute*, nell'Area di Ricerca *Processi e pratiche linguistiche*, Centro di Ricerca LUPT, Università di Napoli Federico II. Saranno liberamente accessibili entro la fine del 2021. <https://www.lupt.it/attivita/lisa.html>
Dimensione delle risorse ***Size of resources***	**CIPPS**: **17 ore di sonoro** per 4 soggetti [ogni registrazione corrisponde a una intera seduta di psicoterapia di ca 60 minuti: paziente A: 3 sedute per un totale di 2 ore e 30 minuti paziente; paziente B 4 sedute per un totale di 3 ore e 58 minuti; paziente C: 2 sedute per un totale di 2 ore e 8 minuti; paziente D: 1 seduta per un totale di 28 minuti]. Ca **10 ore** sono state **trascritte ortograficamente con il sistema CLIPS** (opportunamente adattato in particolare per l'annotazione dei fenomeni di disfluenza) per un totale di ca **59.000 tokens** Presso il Laboratorio LABLITA è attualmente in corso la segmentazione e annotazione prosodico-informativa del corpus. **CIPP-ma**: ca **10 ore di sonoro** - **40 pazienti + 40 controlli** [ogni registrazione corrisponde all'esecuzione di più compiti: descrizione orale di una figura complessa e produzione di parlato (semi)spontaneo stimolato con domande

	su diversi argomenti (1. elementi presenti nella vignetta; 2. attività giornaliera; 3. famiglia)]. Attualmente (agosto 2021) sono state acquisite le registrazioni di 20 pazienti; di 2 pazienti sono state acquisite 2 registrazioni in tempi successivi. Le registrazioni sono di durata variabile, anche in relazione alla gravità della patologia, e vanno da un minimo di ca 3 minuti a un massimo di ca 16 minuti; per lo stesso motivo vi è una notevole variabilità individuale anche in termini di tokens prodotti (*si prevedono nuove acquisizioni nel periodo 2020-2021, in particolare si acquisiranno nuove registrazioni degli stessi pazienti a stadi ulteriori della malattia)*. È stata completata la **trascrizione ortografica con il sistema CLIPS** (opportunamente adattato in particolare per l'annotazione dei fenomeni di disfluenza) per ca **17.000 tokens** e **ca 2.400 types**. Le trascrizioni ortografiche sono state sottoposte a **codifica XML** ed è in corso la segmentazione e annotazione prosodico-informativa. In via sperimentale, è stata inoltre sviluppata un'**analisi** pilota **delle emozioni** relativa alle trascrizioni della paziente C. Per l'analisi qualitativa è stato utilizzato il modello teorico della tematizzazione delle emozioni; per l'analisi qualitativa è stato utilizzato il lessico emotivo ITEM oltre a strumenti computazionali per l'analisi automatica dei testi. In entrambi i casi i risultati (provvisori) delle analisi attestano un incremento del lessico emotivo al progredire della malattia, con particolare evidenza per le emozioni positive (gioia), in controtendenza rispetto a quanto atteso e descritto in letteratura relativamente alla incapacità dei soggetti affetti da Alzheimer di riconoscere ed esprimere emozioni. **Gruppo di controllo per CIPP-ma**: attualmente (agosto 2021) sono state acquisite le interviste di **16 soggetti** (+ 2 acquisizioni longitudinali), corrispondenti a un sonoro di ca **2 ore e 17 minuti**, di cui è stata ultimata la trascrizione ortografica con il sistema CLIPS (adattato) con un totale di ca **20.100 tokens** e ca **2.900 types** **CIPP-mci**: ca **16 ore di sonoro** per 2 gruppi di registrazioni [ogni registrazione corrisponde all'esecuzione di più compiti: produzione di parlato (semi)spontaneo stimolato con domande su diversi argomenti (1. elementi presenti nella vignetta; 2. attività giornaliera; 3. famiglia) e somministrazione della batteria SAND (*Screening for Aphasia in NeuroDegeneration*), che si compone di 9 *task* di valutazione delle abilità linguistiche (1. Denominazione di immagini; 2. Comprensione di frasi; 3. Comprensione di singole parole; 4. Ripetizione di parole; 5. Ripetizione di frasi; 6. Lettura; 7. Descrizione scritta; 8. Associazione semantica; 9. Descrizione orale di una vignetta]: gruppo a/pazienti mci (20 soggetti); gruppo b/controlli (20 soggetti). Si stima un totale finale di ca **105.000 tokens**. Attualmente (agosto 2021) sono state acquisite le registrazioni di **6 coppie** (pazienti/controlli) per un totale **di 4 ore e 45 minuti**, di cui è stata ultimata la trascrizione ortografica con il sistema CLIPS (opportunamente adattato in particolare per l'annotazione dei fenomeni di disfluenza). Il parlato trascritto corrisponde a ca **31.600 tokens** e **ca 4.400 types**. Le trascrizioni ortografiche sono state sottoposte a **codifica XML** ed è in corso la segmentazione e annotazione prosodico-informativa.

	Il sistema di trascrizione dei tre corpora è manuale, basata sul confronto e accordo tra più trascrittori/annotatori. **L'Unità** che lavora al progetto è composta da: - Francesca M. Dovetto (responsabile, Univ. di Napoli Federico II) - Alessia Guida (dottoranda, Univ. di Napoli Federico II) - Raffaele Guarasci (RTD-CNR, Napoli) - Anna Chiara Pagliaro (assegnista Postdoc, Univ. di Napoli Federico II) - Lucia Raggio (laureanda magistrale, Univ. di Napoli Federico II) - Simona Schiattarella (dottoranda, Univ. di Napoli Federico II) - Sundra Sorrentino (borsista, Univ. del Molise) - Simona Trillocco (borsista, Univ. di Firenze)
Descrizione degli obiettivi e del design del corpus / corpora ***Description of the objectives and the design of the corpus / corpora***	Obiettivo prioritario è la costruzione di corpora di italiano parlato relativi a diverse patologie del linguaggio dell'età adulta e senile, trascritti e annotati in analogia con altri corpora di parlato italiano normofasico. La comparabilità dei dati tende a risolvere una lacuna negli studi sulle concrete manifestazioni linguistiche patologiche nonché nella classificazione stessa delle alterazioni linguistiche fondate su corpora e dati tratti quasi esclusivamente da lingue diverse dall'italiano (prevalentemente l'inglese). Particolare salienza è assegnata ai fenomeni di disfluenza verbali e vocali non verbali. Obiettivo immediatamente subordinato è l'individuazione di predittori linguistici diagnostici (ad es. correlati linguistici in grado di discriminare tra invecchiamento fisiologico e MCI) e prognostici (ad es. correlati linguistici che permettano di distinguere soggetti MCI che svilupperanno l'Alzheimer e soggetti che non svilupperanno la malattia) a partire dall'eloquio spontaneo dei pazienti. L'individuazione di predittori linguistici della malattia nel parlato (spontaneo) dei pazienti può costituire una risorsa utile ed economica in alternativa all'individuazione dei biomarcatori verso cui si è rivolta la ricerca più recente, che richiede esami costosi e invasivi. Ulteriore obiettivo è l'elaborazione di proiezioni relative alle manifestazioni linguistiche conseguenti all'avanzamento della malattia. La raccolta di subcorpora longitudinali consente di ottenere dati importanti relativi alle diverse tappe del deterioramento progressivo dell'eloquio. Obiettivo più generale è la promozione della interdisciplinarità, con il coinvolgimento di competenze tecniche multidisciplinari: mediche, linguistiche e, più in generale, delle scienze dei comportamenti psicopatologici. Il **design** del corpus (raccolta di corpora) comprende un numero di ore di registrazione (almeno 10) per diverse **patologie dell'età adulta e senile**: dai disturbi dello spettro della schizofrenia e dei disordini psicotici alle patologie neurodegenerative a carattere progressivo (tra cui MCI, Demenza di Alzheimer). La struttura del corpus è aperta: i corpora saranno implementati con l'acquisizione di registrazioni relative a ulteriori disturbi del linguaggio

	associati al deterioramento cognitivo di natura dementigena.
Riferimenti bibliografici ***References***	**CIPPS** Francesca M. Dovetto & Monica Gemelli, *Marcatori discorsivi nel parlato schizofrenico*, in Barbara Gili Fivela & Carla Bazzanella (a cura di), *Fenomeni di intensità nell'italiano parlato*, Firenze, Franco Cesati Editore, 2009, pp. 181-193 - Francesca M. Dovetto & Monica Gemelli, *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, 2013 [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) - Francesca M. Dovetto, *Schizofrenia e deissi*, «Studi e Saggi Linguistici», LII (2014), pp. 101-132 - Francesca M. Dovetto, *Uso delle parole nella schizofrenia*, in Laura Mariottini (a cura di), *Identità e discorsi. Studi offerti a Franca Orletti*, Roma, Roma TrE-Press, 2015, pp. 161-174 - Francesca M. Dovetto, Emanuela Cresti & Bruno Rocha, *Schizofrenia tra prosodia e lessico. Prime analisi*, «Studi Italiani di Linguistica Teorica e Applicata (*SILTA*)», Numero tematico: *Tra linguistica medica e linguistica clinica. Il ruolo del linguista*, a cura di Franca Orletti, Anna Cardinaletti & Francesca M. Dovetto, XLIV/3, Nuova Serie (2015), pp. 486-507 - Emanuela Cresti, Francesca M. Dovetto & Bruno Rocha, *Schizophrenia and Prosody. First Investigations*, in Claudia Manfredi (ed.), IX International Workshop *Model and Analysis of Vocal Emissions for Biomedical Applications* (MAVEBA), September 2-4 2015, *Proceedings*, Firenze, Firenze University Press, 2015, pp 139-142, ISBN 978-88-6655-792-0 - Francesca M. Dovetto, *Usi della prima persona plurale nel testo schizofrenico*, in Patrizia Sorianello (a cura di), *Il parlato disturbato. Modelli, strumenti e dati empirici*, Roma, Aracne, 2017, pp. 49-66 - Francesca M. Dovetto, Alessandro Panunzi & Lorenzo Gregori, *Sull'annotazione di un corpus orale mistilingue non standard (patologico schizofrenico)*, in Anna De Meo & Francesca M. Dovetto (a cura di), *La Comunicazione parlata / Spoken Communication. Napoli 2016*, Collana "La comunicazione parlata", Roma, Aracne, 2017, pp. 345-361 - Emanuela Cresti & Massimo Moneglia, *Prosodic Monotony and Schizophrenia*, in Francesca M. Dovetto, a cura di, *Lingua e patologia. Le frontiere interdisciplinari del linguaggio*, Collana "Linguistica delle differenze" n.2, Roma, Aracne, 2017, pp. 147-197 - Simona Trillocco, *Allineamento testo suono di Corpus di parlato patologico CIPPS*, borsa semestrale di ricerca (2020-2021), Dipartimento di Lettere e

	Filosofia, Università di Firenze (tutor Massimo Moneglia) **CIPP-ma** - Marina Melone, Francesca M. Dovetto, Simona Schiattarella, Alessia Guida & Cinzia Coppola), *Parola, linguaggio ed emozioni nelle malattie neurodegenerative. Dalla fisiopatologia agli studi clinici con uno studio pilota sulla tematizzazione delle emozioni*, in Francesca M. Dovetto, a cura di, *Lingua e patologia. I sistemi instabili*, Collana "Linguistica delle differenze" n.5, Roma, Aracne (ics) - Falco Mariacristina, Guarasci Raffaele, Maisto Alessandro, Pelosi Serena, *Marcovaldo ovvero Le Stagioni in città: un esperimento computazionale*, «Rassegna Italiana di Linguistica Applicata (RILA)» 2/3, 2017, pp. 47-64 - Trotta, Daniela, Albanese Teresa, Stingo Michele, Guarasci Raffaele, Elia Annibale, *Multi-Word Expressions in spoken language*, PoliSdict, Quarta Conferenza Italiana di Linguistica Computazionale (CLiC-it 2018) - Alessia Guida, *Costruzione e annotazione di un corpus dell'età senile in pazienti affetti da disturbi neurocognitivi*, tesi di dottorato 35° ciclo, Dottorato in Filologia moderna, Università Federico II di Napoli (tutor Francesca M. Dovetto) - Anna Chiara Pagliaro, *Linguistica dei corpora e corpora non standard*, assegno di ricerca post-doc 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) **CIPP-mci** - Francesca M. Dovetto, Simona Schiattarella, Alessia Guida, Cinzia Coppola, Marina Melone), *Relationships between cognition, emotion and language in Dementia Syndrome, from interdisciplinary to transdisciplinary research: A case study.* 34th International Conference of Alzheimer's Disease International (ADI) (Singapore, 10-12 December 2020) (poster). - Simona Schiattarella, *Competenze lessicali e sintattiche in età senile e in contesti normali e turbati*, tesi di dottorato 34° ciclo, Dottorato in Filologia moderna, Università Federico II di Napoli (tutor Francesca M. Dovetto)