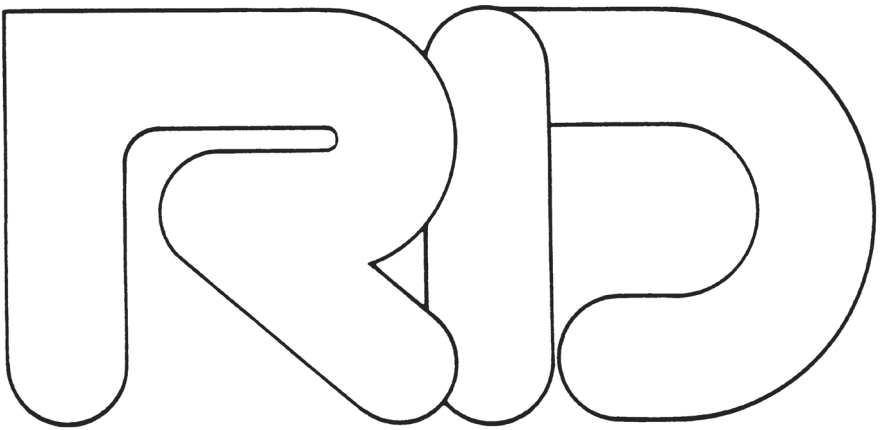

44

RIVISTA ITALIANA DI DIALETTOLOGIA

lingue dialetti società

RIVISTA ITALIANA DI DIALETTOLOGIA. Lingue dialetti società

«RID. Rivista Italiana di Dialettologia» è una rivista internazionale con referaggio anonimo (*blind peer review*), pubblicata annualmente.
«RID. Rivista Italiana di Dialettologia» is a blind peer-reviewed international journal published once a year.

Edizione e amministrazione
Edizioni Pendragon, via Borgonuovo 21/a, 40125 Bologna - tel. 0039 051 267869
www.pendragon.it – RID@pendragon.it
Periodico annuale. Abbonamento: € 39,00 (Italia); € 54,00 (Estero).

Modalità di pagamento / Terms of payment

Italia: versamento sul c.c.p. n. 25317405 intestato a Edizioni Pendragon srl, via Borgonuovo 21/a, 40125 Bologna, specificando la causale.
Bonifico bancario: Edizioni Pendragon srl, IBAN IT50C0558402402000000014154, specificando la causale.
Foreign countries: International cheque or postal money order, in euro, to Edizioni Pendragon srl, via Borgonuovo 21/a, 40125 Bologna
Bank transfers: IBAN IT50C0558402402000000014154
cod. SWIFT BPMIITMM754

Chi richiede fattura di abbonamento deve specificare nella causale o per lettera o all'email RID@pendragon.it l'Ente a cui intestare la fattura, con tutti i dati necessari all'emissione.
L'abbonamento si considera tacitamente rinnovato per l'anno successivo se non viene disdetto entro il mese di dicembre.
Tutta la corrispondenza, i periodici in cambio e i libri per recensione possono essere inviati al direttore editoriale presso Edizioni Pendragon srl, via Borgonuovo 21/a, 40125 Bologna. I libri, periodici, estratti ed ogni altro materiale riguardante le singole regioni ai rispettivi corrispondenti regionali (se ne veda l'indirizzario a fine del fascicolo).
I dattiloscritti pervenuti alla rivista, anche se non pubblicati, non vengono restituiti.

Registrazione presso il Tribunale di Bologna n. 4630 del 6.3.1978
Direttore responsabile: Lorenzo Coveri

Finito di stampare nel mese di XXXXX 2021 a cura di NW presso LegoDigit s.r.l. - Lavis (TN)

RIVISTA ITALIANA DI DIALETTOLOGIA
Lingue dialetti società

Anno XLIV (2020), numero unico [= RID 44]

INDICE

9 Ines Loi Corvetto (Università di Cagliari), *Ricordo di Antonietta Dettori*

RID/MONOGRAFICA
Corpora di italiano parlato nel panorama italiano: verso l'individuazione di pratiche condivise, a cura di Eugenio Goria e Simone Ciccolone

17 Presentazione

19 Eugenio Goria (Università di Torino), Simone Ciccolone (Università di Cagliari), *Individuare pratiche condivise per i corpora di parlato in Italia: prospettive di ricerca*
37 Lorenzo Spreafico (Università di Bergamo), *Corpora di parlato o corpora di ascoltato?*
53 Silvia Ballarè e Caterina Mauri (Università di Bologna), *La creazione del corpus* KIPARLA: *criteri metodologici e prospettive future*
71 Daniela Mereu, Alessandro Vietti (Libera Università di Bolzano), *Studiare la variazione fonetica nel parlato spontaneo dialogico: il corpus* DIA (DIALOGIC ITALIAN)
89 Stefania Spina, Luciana Forti, Fabio Zanda (Università per stranieri di Perugia), *Verso un corpus di riferimento dell'italiano parlato dialogico: il modello* BNC2014
107 Marco Angster (Università di Zara), Raffaele Cioffi, Marco Bellante, Livio Gaeta (Università di Torino), *Corpora e varietà minoritarie: le isole walser in Italia*
127 Lorenzo Ferrarotti, Aline Pons, Sara Racca (Università di Torino), *Prospettive di studio del parlato dialettale a partire dagli etnotesti dell'*ALEPO (Atlante Linguistico ed Etnografico del Piemonte Occidentale)

151 Pierangela Diadori, Elena Monami (Università per stranieri di Siena), CLODIS (CORPUS DI LINGUA ORALE DEI DOCENTI DI ITALIANO PER STRANIERI): *una banca dati multimodale per la formazione dei docenti di italiano L2*

RID/RICERCA

Saggi e studi

173 Marta Maddalon (Università della Calabria), *Lingua e dialetto nei monologhi di Andrea Pennacchi (El Poiana): analisi etno-sociolinguistica*
203 Daniela Mereu (Libera Università di Bolzano), Nicoletta Puddu (Università di Cagliari), *Il 'gergo' di Cagliari a novant'anni dalle prime inchieste*
233 Francesco Laterza (Università di Torino), *L'italiano della divulgazione scientifica. I casi di* Focus *e* Le Scienze
261 Silvia Natale (Università di Berna), *Note sui 'nuovi' repertori linguistici degli emigrati italiani nella Svizzera tedesca*
289 Margherita Di Salvo (Università 'Federico II' di Napoli), Sara Matrisciano (Vienna University of Economics and Business), *Uso e forme dell'inglese come marcatore identitario tra espatriati e migranti*

Testi e documenti

315 Paolo D'Achille, Claudio Giovanardi (Università di Roma 3), Michele Loporcaro (Università di Zurigo), Vincenzo Faraoni (Università La Sapienza di Roma), *La lettera* E *del* 'Vocabolario del romanesco contemporaneo'

Note rassegne discussioni

335 Salvatore Claudio Sgroi (Università di Catania), *L'errore, le rubriche linguistiche e gli oroscopi*

RID/SCHEDARIO

347 *Generalità,* a cura di Immacolata Tempesta (Università di Lecce)
359 *Friuli*, a cura di Federico Vicario (Università di Udine)
371 *Marche*, a cura di Fabio Aprea (Università La Sapienza di Roma-Università di Vienna)

395 *Lazio*, a cura di Paolo D'Achille (Università di Roma 3)
443 *Abruzzo. Molise,* a cura di Francesca Guazzelli (Università 'G.D'Annunzio' di Chieti-Pescara
457 *Sardegna*, a cura di Nicoletta Puddu (Università di Cagliari)

465 Notizie sui Collaboratori
471 Istruzioni per i Collaboratori
474 Elenco dei Corrispondenti di RID/Schedario

Rivista Italiana di Dialettologia. Lingue dialetti società a. XLIV (2020), Pendragon Bologna

Eugenio Goria (Università di Torino),
Simone Ciccolone (Università di Cagliari)

INDIVIDUARE PRATICHE CONDIVISE PER I CORPORA DI PARLATO IN ITALIA: PROSPETTIVE DI RICERCA*

***

1. I corpora di parlato in Italia oggi

A quasi quindici anni dalla pubblicazione del corpus CLIPS (cfr. Albano Leoni 2007a, 2007b), e a meno di trenta da quella del LIP (cfr. De Mauro *et al.* 1993), il panorama delle ricerche sul parlato in Italia risulta notevolmente cambiato (si veda Crocco 2015 per una rassegna).

I contributi qui raccolti, incentrati su *Corpora di parlato nel panorama italiano: verso l'individuazione di pratiche condivise*, si posizionano cronologicamente a una distanza relativamente breve da una fase di notevole fermento e attività nel campo della ricerca scientifica sul parlato in Italia, che ha prodotto progetti di grande portata come appunto i già citati CLIPS e LIP, e in un periodo in cui l'interesse verso il parlato appare ancora fortemente in crescita, benché rappresentato da iniziative più contenute e con obiettivi più specifici. Al contempo, e forse proprio in virtù di questa frammentazione degli sforzi della ricerca scientifica nazionale nel settore, non appare ancora pienamente raggiunto un paradigma condiviso, sia nei termini di un insieme comune e unitario di strumenti per la raccolta e il trattamento di dati linguistici di parlato, sia per quanto riguarda il raggiungimento di obiettivi condivisi, come la costruzione di un corpus di riferimento per il parlato in Italia.

Proprio iniziative come CLIPS, C-ORAL-ROM (cfr. Cresti, Moneglia 2005), o più recentemente VoLIP (cfr. Voghera *et al.* 2014) hanno permesso di raccogliere e mettere a disposizione della comunità scientifica un campione sempre più ricco e variegato di dati per la descrizione e l'analisi del parlato, sollecitando continuamente la discussione scientifica sul piano teorico e metodologico. Accanto a questi sono fioriti numerosi progetti con obiettivi più specifici, che hanno permesso uno studio più approfondito e dettagliato di particolari varietà del repertorio linguistico italo-romanzo, o specifiche situazioni sociolinguistiche nel territorio italiano. Si pensi ad es. alla *Banca dati di italiano L2* di Pavia (cfr. Andorno 2001), oppure al *Corpus di italiano Televisivo* (CiT, cfr. Spina 2005), o infine al corpus di parlato bilingue in Alto Adige *Kontatto* (cfr. Dal Negro, Ciccolone 2018).

Parallelamente, la disponibilità di nuovi strumenti in termini sia di metodi di annotazione e analisi (come AN.ANA.S o PraTiD) sia di supporti al lavoro del ricercatore nelle varie fasi di trattamento del dato linguistico (ad esempio Praat, ELAN, WebMaus, EMU), oltre a stimolare e facilitare l'analisi di fenomeni linguistici propri della modalità parlata (come segnali discorsivi, prosodia e comunicazione multimodale), ha anche aperto a un possibile cambio di prospettiva nello studio dei sistemi linguistici *tout court*. Infatti, uno degli effetti della maggiore diffusione di strumenti informatici dedicati alla gestione e al trattamento di dati linguistici orali è sicuramente rappresentato da una considerevole proliferazione di studi, anche fra le tesi di laurea e di dottorato, che poggiano sull'osservazione empirica di materiali orali raccolti *ad hoc* dal ricercatore.

Tutto ciò introduce una maggiore oggettività nella ricerca linguistica, in quanto fa sì che si ponga al centro della riflessione l'osservazione empirica del comportamento linguistico dei parlanti, superando metodologie basate esclusivamente sull'introspezione, e dunque sulla produzione di esempi fittizi, o su osservazioni estemporanee. Inoltre, la maggiore disponibilità di corpora ha reso più comune anche l'utilizzo di metodi quantitativi nella ricerca linguistica (cfr. *inter alia* Vietti 2003, 2005; Berruto, Cerruti 2015; nonché la rassegna in Iannàccaro, Ciccolone 2017), e, più in generale, ha contribuito a una maggiore consapevolezza delle differenze fra approcci qualitativi e quantitativi (cfr. Pallotti 2016). Bisogna infine osservare che, anche in virtù delle risorse di dati e dei nuovi strumenti a disposizione, l'interesse verso il parlato è cresciuto in misura considerevole non solo nei campi di ricerca programmaticamente legati all'oralità (come fonetica, dialettologia, sociolinguistica, linguistica acquisizionale, analisi conversazionale), ma anche in prospettive teoriche tradizionalmente meno legate all'uso di dati empirici, come la tipologia linguistica e la grammatica generativa (cfr. Kortmann 2008; Mauri, Sansò 2018; D'Alessandro 2018).

Tuttavia, come anticipato, a tale proliferare di studi su dati di parlato non sembra sempre accompagnarsi una riflessione organica e (soprattutto) condivisa sulle prassi da adottare nella raccolta e nel trattamento dei dati e, più in generale, su tutte quelle operazioni che preludono alla costruzione di un corpus e ne rendono possibile l'utilizzo. Fino ad ora, infatti, gran parte della riflessione metodologica sui corpora si è concentrata su varietà scritte, con problematiche chiaramente diverse da quelle che caratterizzano i corpora orali; sono inoltre disponibili per lo scritto risorse esponenzialmente più grandi rispetto al parlato, corredate tra l'altro di vari livelli di descrizione del dato linguistico nonché di strumenti automatici per l'annotazione (lemmatizzazione, *POS tagging*, *parsing* sintattico ecc.). Per il parlato questi strumenti sono assenti, non applicabili o ancora in elaborazione (cfr. Magnini *et al.* 2013, Basile *et al.* 2016), innanzitutto proprio in virtù della maggiore complessità del dato stesso: la compresenza di più partecipanti, ciascuno associato a metadati propri, l'alternarsi di più varietà e strutture concorrenti (nel parlato

bilingue, nell'interlingua, nella variazione diafasica in generale), nonché di codici semiotici diversi (intonazione, gestualità) rappresentano ancora oggi una sfida non risolta. In altre parole, sembra si possa osservare anche nelle risorse attualmente disponibili la presenza di un *written language bias* nei termini di Linell (2005).

Ed è proprio in virtù di questa evidente disparità fra le fonti scritte e quelle orali che pare lecito chiedersi se la maggiore complessità del dato non debba spingere con maggior vigore la riflessione scientifica verso una più sistematica condivisione sia di un patrimonio metodologico comune, sia di strumenti e fonti di dati riutilizzabili e ulteriormente implementabili, in una prospettiva ecologica già ventilata da Voghera *et al.* (2014). Allo stesso modo, un dibattito organico sulla condivisione di risorse non può ignorare la presenza di risorse "sommerse" e archivi sonori non (ancora) accessibili o disponibili alla comunità scientifica; il che pone ulteriori questioni metodologiche ed etiche legate al trattamento dei dati audio successivo alla loro raccolta (cfr. Calamai *et al.* 2016) che, ancora una volta, dovranno essere affrontate in maniera organica.

Intorno a questi temi, si è tenuto nel corso del LIII congresso della Società di Linguistica Italiana, presso l'Università dell'Insubria, un workshop tematico dal titolo *Corpora di parlato: verso l'individuazione di pratiche condivise*, i cui contributi sono raccolti nella presente sezione monografica. Rispetto al titolo originario, si è optato per una maggiore specificità, concentrando la riflessione sui corpora di parlato *nel panorama italiano*. Ciò si deve non tanto alla natura della riflessione, la cui portata è spesso più generale e applicabile in linea di principio anche a scenari diversi, quanto ai problemi connessi più specificamente con la natura delle lingue e varietà di lingua discusse dai vari autori: sono infatti discussi dati di varietà regionali di italiano, dialetti e lingue di minoranza appartenenti al continuum italoromanzo e, in un caso, una lingua alemannica di minoranza. Per questo motivo, è parso opportuno insistere con maggiore forza su uno dei denominatori comuni a tutti i lavori contenuti nel volume.

2. Prospettive a confronto

Nell'intenzione dei curatori, la presente sezione monografica intende mettere a sistema una serie di riflessioni teorico-metodologiche ricorrenti, che caratterizzano uno o più aspetti della costruzione di corpora orali. Tali riflessioni sono state adottate da linguisti provenienti da ambiti di ricerca anche molto diversi tra loro, come la fonetica, la sociolinguistica, la dialettologia, la documentazione linguistica e l'insegnamento di lingue seconde. Nonostante le evidenti differenze nelle esigenze e nelle finalità di utilizzo dei corpora di cui si discute, numerose problematiche legate alla natura "elastica" della modalità parlata (cfr. Voghera 2017: 189-198) ritornano nei lavori di tutti gli autori e tutte le autrici.

Inoltre, un importante punto di contatto fra i lavori e i corpora di cui si discute all'interno del volume è il loro carattere di assoluta novità: in alcuni casi le risorse presentate sono di recente realizzazione, e in questo senso la discussione delle metodologie adottate nella loro costruzione alimenta il confronto fra prospettive teoriche diverse e fornisce un contributo prezioso per la condivisione di strumenti e prassi di ricerca, anche in direzione di uno standard condiviso. Altri contributi, invece, rispecchiano esperienze di ricerca che non nascono con l'obiettivo primario di produrre un corpus delle varietà orali indagate, ma si trovano a dover risolvere questioni metodologiche analoghe. Parte della riflessione, in questi casi, si concentra inoltre sul tema a cui si è già accennato delle cosiddette "risorse sommerse", e di come, adottando una prospettiva ecologica, sia possibile rifunzionalizzare i dati raccolti durante esperienze precedenti a una fruizione sotto forma di corpus orale.

Il primo articolo di questa raccolta, ad opera di Lorenzo Spreafico, presenta una riflessione teorico-metodologica relativa alle prassi di trascrizione fonetica, che in molti casi rappresentano una delle prime operazioni che il ricercatore effettua durante la costruzione di un corpus. Spreafico sottolinea l'importanza del ruolo giocato dalla percezione del trascrittore durante la prassi trascrittoria, facendo riferimento alla natura "ascoltata" dei corpora di parlato, e discutendo delle implicazioni teoriche che derivano dalla scelta di una specifica convenzione di trascrizione.

La seconda parte è dedicata alla presentazione di tre corpora di italiano parlato, senza distinzione fra risorse che sono state ideate sin dal principio per una fruizione estesa da parte della comunità dei linguisti e risorse che invece rimangono, al momento, a disposizione esclusiva delle Università che hanno dato avvio alla raccolta. L'elemento comune che ricorre in tutti e tre i contributi che formano questa sezione è una riflessione sistematica sulle metodologie di raccolta e costruzione dei vari corpora, in relazione al tipo di dati e al tipo di parlato che essi contengono. Così, Mauri e Ballarè discutono le caratteristiche del corpus KIParla a partire dalla fase di *corpus design*, fino ad arrivare ai possibili ampliamenti della risorsa. Mereu e Vietti riflettono su come conciliare l'elevato grado di dettaglio dei dati richiesto nell'indagine sociofonetica con le vaste possibilità offerte dai corpora di parlato conversazionale: entrambi gli aspetti coesistono infatti nel corpus DIA (*DIalogic Italian*), di cui illustrano le caratteristiche. Infine, Spina e colleghi presentano la loro proposta di come applicare il modello dei *reference corpora* disponibili per altre lingue (l'esempio adottato in questo caso è il *British National Corpus*) all'italiano, presentando le caratteristiche del CIP-PG (*Corpus di Italiano Parlato raccolto a Perugia*).

Un ulteriore nucleo tematico riguarda le problematiche che emergono quando l'esigenza di documentare le varietà orali mira a includere, oltre all'italiano, anche dialetti e lingue di minoranza parlati sul territorio nazionale. Nonostante i

due ambiti siano tradizionalmente distinti all'interno delle scienze del linguaggio, una riconciliazione sarebbe auspicabile per vari motivi. In primo luogo, come è noto, italiano, dialetti, lingue di minoranza e lingue di recente immigrazione si trovano a coesistere nello scenario sociolinguistico italiano, rendendo, di fatto, molti corpora di parlato dei corpora plurilingui, le cui scelte in sede di trascrizione e annotazione dei dati dovranno necessariamente tenere conto della presenza di più codici (cfr. il caso del ParlaTo discusso in Cerruti, Ballarè in stampa). Inoltre, la riflessione sulle prassi di raccolta e di elicitazione dei dati linguistici, così come sulle modalità di trascrizione e archiviazione del parlato, possiedono da sempre un ruolo di primaria importanza proprio in ambiti di ricerca come la *language documentation* (cfr. il contributo di Mereu e Vietti in questo numero). Per contro, discipline come la dialettologia o la *language documentation*, pur non avendo lo scopo primario di produrre corpora annotati paragonabili a quelli in uso per lingue nazionali di maggiore diffusione, si confrontano con problematiche simili a quelle che caratterizzano le varietà orali di queste lingue, soprattutto per quanto riguarda l'esigenza di rappresentare varietà di lingua caratterizzate da un elevato grado di variabilità.

Ad alcune di queste problematiche forniscono una soluzione Angster e colleghi, che si concentrano sulle modalità di trattamento informatico dei dati, adottate nei progetti Archiwals e Diwals, dedicati alla documentazione delle varietà walser. Infatti, nonostante il progetto si sia occupato fino ad ora di testi scritti, la presenza di numerose varianti grafiche trova un interessante parallelismo nella variazione che caratterizza le varietà orali. Ferrarotti e colleghe riflettono invece sulla possibilità di utilizzare sotto forma di corpus gli etnotesti, ovvero i materiali orali di varia lunghezza raccolti durante le inchieste dialettologiche che hanno condotto alla redazione dell'ALEPO (*Atlante Linguistico ed Etnografico del Piemonte Occidentale*).

Infine, chiude questa raccolta di articoli il lavoro di Diadori e Monami, in cui si presenta il CLODIS (*Corpus di Lingua Orale dei Docenti di Italiano per Stranieri*). Il corpus si distingue dalle altre risorse presentate nel volume per un obiettivo più specifico rispetto a queste, che mirano infatti alla documentazione del parlato per scopi legati all'indagine scientifica. Al contrario, il CLODIS si propone come risorsa da impiegarsi nella formazione dei futuri insegnanti di lingua italiana a stranieri, come supporto alla didattica frontale.

3. Verso l'individuazione di pratiche condivise: un programma di ricerca "diffuso"

All'inizio della nostra riflessione su questo tema, in fase di preparazione della proposta di workshop per il congresso della Società di Linguistica Italiana,

ci eravamo posti una serie di obiettivi che speravamo di raggiungere e di temi, di cui speravamo di poter discutere con i relatori e gli altri partecipanti, non solo di carattere teorico e metodologico generale ma anche molto specifici e pratici. Il confronto che è scaturito in quest'occasione, a cui hanno contribuito linguisti e linguiste con *background* teorici anche molto diversi tra loro, ha permesso di affrontare alcuni di questi temi, evidenziando insospettati elementi di somiglianza e mettendo a fuoco una serie di problemi ricorrenti e questioni teorico-metodologiche parzialmente irrisolte.

In qualità di curatori di questa sezione monografica, il nostro auspicio è di aver contributo a fissare almeno parte di questi interrogativi rimasti aperti, individuando i punti chiave del dibattito attuale e contribuendo a individuare le nuove sfide su cui la ricerca sui corpora orali dovrà confrontarsi. Anche a chiusura di questo lavoro editoriale, non possiamo certo dire di averli raggiunti tutti, né di aver esaurito gli argomenti da discutere – tantomeno, e non solo da parte dei curatori, l'interesse scientifico e la voglia di discuterli collettivamente sono diminuiti. Anzi, sono stati alimentati dal confronto concreto con le iniziative di ricerca e i problemi ancora aperti e in attesa di soluzione. Ci pare dunque opportuno concludere il nostro saggio introduttivo con una rassegna delle prospettive di ricerca che si delineano a partire dalle riflessioni contenute negli articoli che seguono.

In primo luogo, come discute ampiamente già il primo contributo di Spreafico, lavorare su corpora di parlato induce a un'imprescindibile riflessione sulla natura del dato nella ricerca linguistica: da un lato è opportuna una delimitazione del campo di osservazione, in quanto con assoluta evidenza nel parlato subentrano altri codici semiotici a complessificare l'oggetto di studio; ma ancor più radicalmente bisogna riflettere sulla natura del processo percettivo e interpretativo che è alla base di qualsiasi tentativo di elaborazione della materia acustica in una qualche forma di codifica (per necessità discreta, sistematica e normalizzante), che è il primo passaggio per la costituzione di un qualsiasi corpus di parlato (o di "ascoltato", come in realtà suggerisce Spreafico).

La consapevolezza dei notevoli vincoli epistemologici che poniamo alla materia linguistica per sottoporla a una qualche forma di controllo, per renderla in altre parole un dato (e un dato, per di più, computabile, di cui si possa osservare la distribuzione), pone costantemente il riflettore sul rischio di ogni operazione di trasposizione e trattamento del dato linguistico che non sia ancorata all'evidenza originaria, al parlato registrato. La qualità acustica delle registrazioni rappresenta quindi un aspetto pratico non di second'ordine, cruciale per l'affidabilità stessa dei dati e indispensabile per permetterne l'uso, ad esempio anche in analisi fonetiche. Nel contempo, la registrazione di materiali audio in alta qualità ha implicazioni anche molto pesanti per il ricercatore, relative ai costi e alla trasportabilità dei registratori professionali, alla possibilità concreta di utilizzarli in ogni situazione e, infine, alla grande quantità di spazio di archiviazione richiesta da questo tipo

di dato digitale. Pertanto, molto spesso il lavoro del linguista "di campo" consiste nell'individuazione del miglior *trade-off* possibile tra la qualità del dato auspicata, la strumentazione disponibile e le limitazioni imposte dai singoli contesti di ricerca.

Inoltre, un aspetto pratico che ha risvolti determinanti sull'affidabilità del dato, e dev'essere quindi tenuto in considerazione nell'impianto metodologico di qualsiasi campagna di raccolta di dati di parlato, riguarda la selezione e conduzione dell'evento comunicativo registrato, che deve tendere a un punto di equilibrio tra la desiderata naturalezza del dato e la controllabilità del *setting*, ad esempio preferendo interazioni con un numero di partecipanti facilmente gestibile e con voci chiaramente riconoscibili, per garantire un corretto ancoraggio ai metadati sia relativi ai parlanti coinvolti che ad eventuali modificazioni del *setting* nel corso dell'interazione. Il protocollo adottato per il corpus DIA (presentato in Mereu, Vietti) rappresenta un ottimo esempio di come tali compromessi possano essere raggiunti: gli autori forniscono generosamente abbondanti dettagli riguardo alle procedure per la raccolta dati e il trattamento semi-automatico delle trascrizioni, realizzando così un obiettivo centrale della nostra iniziativa: la condivisione di buone pratiche e di protocolli di ricerca "aperti".

Riguardo ai limiti dell'osservazione linguistica, Spreafico problematizza la scelta se includere o meno nella raccolta dei dati il canale visivo: includere il video (come ad es. avviene per il CLODIS per permettere l'osservazione dettagliata dell'interazione, cfr. Diadori, Monami) comporta l'inclusione o quantomeno la considerazione, nella descrizione/trasposizione del dato originario, anche dei codici semiotici sul solo canale visivo, come cinesica e prossemica, e quindi una notevole complessificazione del dato. Viceversa, una raccolta di sole registrazioni audio potrebbe più facilmente rischiare di perdere segnali rilevanti, all'interno delle interazioni registrate, in merito a modificazioni del *setting* o a rimandi al campo indicale fisico dell'evento comunicativo.

La trascrizione del parlato deve inoltre considerare quantomeno un inventario minimo delle sue caratteristiche, come pause piene e vuote, cambi di progetto enunciativo e tratti intonativi pertinenti (o in ogni caso, deve permetterne l'implementazione anche in un secondo momento). Quest'osservazione sembra confermata anche dalla condivisione di alcune scelte metodologiche in molti dei contributi qui inclusi (cfr. Mauri, Ballarè; Mereu, Vietti; Spina *et al.*; Diadori, Monami), come quella dell'adozione di una trascrizione approssimativamente ortografica, arricchita di notazioni in linea, spesso ispirate ai principi in uso nella *Conversation Analysis* (anche senza ricorso a *markup* esplicito, come invece avviene con la codifica XML proposta in Spina *et al.*) per marcare pause, disfluenze, allungamenti o altri tratti del parlato. La scelta di una trascrizione ortografica, benché possa sembrare in certa misura "anacronistica", o "meno scientifica", rispetto alle potenzialità offerte ad esempio da una trascrizione strumentale automatica su base

acustica (ma su ciò si veda ancora Spreafico), rappresenta in realtà un ottimo compromesso in termini di efficienza e uniformità del dato, facilitandone l'utilizzo e la condivisione anche da parte di trascrittori meno esperti. Proprio in virtù della sua uniformità, la trascrizione ortografica apre spesso la strada alle successive fasi di elaborazione e analisi; tutto ciò, però, purché il trascritto rimanga costantemente ancorato e immediatamente riconducibile al dato originario, alla registrazione.

Un ulteriore problema legato alla trascrizione del parlato riguarda la sua granularità, in merito alla quale siamo chiamati a individuare un ulteriore punto di equilibrio, stavolta tra la precisione della rappresentazione grafica del parlato e il grado di astrazione proprio della trascrizione scientifica, attraverso il quale il ricercatore individua (e permette di individuare) una serie di unità, tipi (fonologici, morfologici, lessicali) prevedibili. Come osservano acutamente Angster *et al.*, una certa polinomia grafica può caratterizzare, in modi simili benché difformi, sia il testo scritto di lingue meno diffuse o "a bassa densità" (in relazione a grafie diverse o a scritture spontanee) sia la trascrizione del parlato (in virtù di variazioni del significante acustico o anche di usi diversi da parte di trascrittori diversi), rappresentandone quindi un tratto saliente, non alienabile per una riproduzione più fedele della variabilità osservabile. Una soluzione potrebbe prevedere una progressione più graduale verso una normalizzazione del dato originario, permettendo una maggiore variabilità nel primo livello di codifica e una maggiore uniformità a livelli di annotazione successivi, e in particolare con la lemmatizzazione.

Arriviamo così a un altro obiettivo centrale nella nostra iniziativa (purtroppo ancora da raggiungere), ovvero la condivisione di uno schema generale di annotazione dei dati di parlato. In un'ottica ecologica del dato, ovvero nella prospettiva di una condivisione dei dati di corpora e di un loro utilizzo per i più svariati scopi di ricerca, un tale schema di annotazione, per rappresentare un modello efficace e veramente condivisibile, deve necessariamente avere struttura modulare, espandibile e adattabile a diversi approcci teorici (quindi, in un certo senso, *theory-agnostic*) e a diversi programmi di ricerca, anche non previsti al momento della raccolta dati.

Uno schema di annotazione comune ad analisi linguistiche e interazionali deve permettere un certo grado di autonomia tra i diversi tipi di informazione, prevedendo un struttura modulare ed espandibile in base ai diversi canali e al diverso numero di partecipanti all'interazione; al contempo, deve vincolare queste informazioni in una struttura gerarchica (come avviene in ELAN, cfr. Mauri, Ballarè), o ancor meglio di tipo relazionale (come avviene in un database multistrato, cfr. Angster *et al*.). Sul dettaglio delle categorie da utilizzare, ad es. per il *POS tagging*, o sulla forma dell'annotazione sintattica, che può complicare notevolmente lo schema, l'accordo potrebbe essere più difficile e anche non necessariamente auspicabile; tuttavia, lo schema complessivo di annotazione deve permettere un riconoscimento e un accesso intuitivo alle informazioni linguistiche implementate

nel corpus, facilitando l'individuazione di flussi di lavoro nel processo di annotazione largamente condivisi e sostenibili.

Lo stesso dovrebbe avvenire per i metadati, che, anche alla luce delle recenti modifiche alla normativa europea in materia di protezione dei dati personali (GDPR, regolamento UE 2016/679), dovrebbero essere raccolti con prassi trasparenti e ampiamente documentate, e organizzati secondo una struttura condivisibile e, soprattutto, sostenibile in termini di costi e tempi. Su questo esistono già degli standard di riferimento (cfr. *Dublin Core*, citato anche in Angster *et al.*, di cui tuttavia sarebbe ancora da osservare e discutere la reale condivisione e diffusione). Occorre però riflettere sulla complessità e sulle difficoltà dell'adozione di tali strutture di metadati, proprio in relazione alla sostenibilità nel tempo, e in progetti "diffusi", di tale impianto in una raccolta dati di parlato.

Il tempo rappresenta un fattore determinante, che può compromettere o rendere inaccessibili anche progetti altrimenti impeccabili, e che quindi deve essere preso opportunamente in considerazione, in particolare in relazione alla longevità degli strumenti informatici e dei formati digitali da adottare per la raccolta, archiviazione ed elaborazione dei dati di parlato. Ad esempio, per quanto riguarda la cosiddetta "codifica zero", non si può prescindere dall'adozione di Unicode, che rappresenta senza dubbio uno standard affidabile, benché relativamente recente. Risorse digitali obsolescenti dovrebbero quindi essere ricodificate in Unicode, individuando però strategie che permettano di automatizzare tale processo, proprio per evitare una perdita di informazioni non più recuperabile una volta abbandonato il formato di partenza (cfr. la discussione in Ferrarotti *et al.*).

Sui metodi di codifica della struttura del "metatesto" della trascrizione, invece, i contributi raccolti qui mostrano scelte divergenti: CIP-PG adotta uno schema XML-TEI generico; corpora che utilizzano ELAN come strumento principale per l'annotazione (ad es. KIParla, cfr. Mauri, Ballarè) usano anch'essi uno schema XML-TEI, tuttavia molto meno trasparente e "leggibile" dall'uomo senza l'ausilio del software. Ciò nondimeno, il ricorso a XML permette una notevole interoperabilità del dato, anche col supporto di specifici convertitori automatici (cfr. Mauri, Ballarè; Mereu, Vietti). Il formato TextGrid usato da Praat può allo stesso modo essere reso interoperabile, come d'altronde avviene con le trascrizioni in Chat-CLAN. Più complessa invece può risultare la scelta sull'architettura di un database relazionale, che infatti molto spesso tende a rispecchiare più da vicino le specifiche esigenze di un dato progetto di ricerca. La proposta di Angster *et al.* risulta in tal senso innovativa, altro esempio di possibile condivisione di una *best practice* che infatti è già accolta da Ferrarotti *et al.* come modello per la loro iniziativa di estrazione di dati di parlato dagli etnotesti raccolti per l'atlante ALEPO.

I progetti raccolti qui segnalano un ulteriore aspetto in comune, di rilievo proprio in relazione agli strumenti digitali a cui ricorrono: la costituzione di un corpus di parlato comporta molto spesso, nei suoi vari passaggi dalla raccolta dati,

all'archiviazione, fino all'analisi, un flusso di lavoro che coinvolge più software (per la trascrizione e annotazione del file multimediale, la gestione del corpus e la sua distribuzione tramite interfaccia web, la gestione delle *queries*, o ancora per l'analisi statistica). Vari sono gli strumenti proposti, tra l'altro quasi tutti *open source* (ELAN, Praat, WebMAUS, R, EMU). Questo evidenzia ancor di più l'importanza cruciale della interoperabilità e trasponibilità dell'informazione digitale tra formati e sistemi di codifica diversi, nell'ottica del rispetto dei principi FAIR (*Findable, Accessible, Interoperable, Reusable*) evocati anche in Ferrarotti *et al.* La scelta per il CLODIS di usare Nvivo, al contrario, potrebbe sul lungo termine rappresentare un ostacolo, principalmente a causa del formato proprietario dei documenti, non accessibile al di fuori dello specifico software e quindi potenzialmente più prono all'obsolescenza digitale. È importante quindi sottolineare il valore aggiunto di formati digitali "aperti" per i dati, leggibili ed esplorabili dall'essere umano anche senza necessità di un software specifico.

Infine, non possiamo ignorare un'ulteriore difficoltà materiale già segnalata in occasione di precedenti progetti, come nel caso del CLIPS, e che è ancora più presente ora: la gestione di iniziative che richiedono grande partecipazione, e il mantenimento di risorse e attività sul medio-lungo termine, con poche o scarse ricadute nella fase iniziale in termini di "prodotti" della ricerca. Soprattutto in una fase storica nella ricerca universitaria italiana, in cui si tende a capitalizzare in tempi brevi progetti scientifici piuttosto specifici e mirati, l'idea di un programma di ricerca articolato e complesso come quello della costituzione di un corpus di parlato di dimensioni estese fa ancora più fatica a decollare.

Anche per questo motivo risulta di cruciale importanza stabilire un protocollo condiviso, trasmissibile, dettagliato e al contempo modulare, flessibile, adattabile al contesto e alle specifiche esigenze di ricerca. Struttura modulare può avere anche il programma di ricerca nella sua interezza, prendendo come modello operativo quello del corpus KIParla e dei suoi moduli, al momento due ma in attesa di essere integrati con altre aggiunte più recenti. Una implementazione efficace di questo modello richiede però che ogni modulo sia sufficientemente esteso da permettere un'analisi circoscritta al modulo stesso, permettere di testarne l'affidabilità e di validarne l'utilizzo per ricerche a più ampio raggio. La struttura modulare, in sostanza, non può prescindere dalle necessità di affidabilità del campionamento dei dati per i corpora linguistici, per quanto più lasche e permissive per il parlato spontaneo in particolare e per corpora orali in generale.

Una prospettiva ecologica del dato linguistico richiede da un lato lo sviluppo di iniziative sostenibili, in termini economici e di tempo, e dall'altro una efficace e diffusa riutilizzabilità di dati raccolti per scopi diversi. Il ricorso a dati altrui richiede necessariamente la condivisione di uno schema di annotazione comune, come segnalato sopra; comune, ma comunque modulare e aperto a successive integrazioni.

Se l'obiettivo che ci eravamo posti in origine, ovvero l'individuazione di pratiche condivise per i corpora di parlato, era di certo ambizioso, non di meno risulta necessario un ulteriore slancio di entusiasmo e di impegno condiviso per poter raggiungere un obiettivo ancor più arduo: la creazione di un corpus di riferimento del parlato per il panorama italiano, nato dall'unione e dalla messa in rete degli sforzi dei singoli gruppi di ricerca impegnati in questo campo, in una sorta di progetto "diffuso" che sappia integrare prospettive diverse e finalità specifiche in un nuovo paradigma di indagine del variegato diasistema italiano.

Appendice
Rassegna sintetica di strumenti e risorse per i corpora di parlato

A complemento di questa nostra breve riflessione introduttiva alla sezione monografica, inseriamo di seguito una rassegna, inevitabilmente parziale, di alcuni strumenti e risorse disponibili sul tema dei corpora di parlato in Italia, sia in termini di fonti bibliografiche fondamentali (a cui non solo noi, ma anche molti degli autori dei contributi di questa sezione hanno fatto riferimento, qui o durante i lavori del workshop di Como), sia in termini di banche dati e corpora liberamente accessibili, nonché software e risorse digitali relative al trattamento di dati di parlato e alla creazione di corpora di parlato.

Ci teniamo a segnalare che quest'appendice non è solo frutto dell'intenso scambio di idee e discussioni tra gli autori, ma anche, e in misura considerevole, del confronto scaturito con tutti gli studiosi che hanno partecipato a vario titolo ai lavori del workshop e di questa sezione monografica, compresi i componenti del Comitato Scientifico del workshop e i revisori anonimi dei contributi, a cui esprimiamo la nostra gratitudine.

In particolare, in quest'appendice sono state integrate alcune voci riprese dai capitoli successivi, segnalandole con l'indicazione degli autori tra parentesi quadre, in modo da offrire al lettore uno spunto per possibili approfondimenti.

1. Corpora di parlato in Italia: alcuni riferimenti essenziali

1.1. LIP: De Mauro *et al.* 1993; il testo è accompagnato da due *floppy disk* con il corpus. [2.2]

1.2. CLIPS: si veda Albano Leoni 2007a per una presentazione del progetto, *id.* 2007b per una breve storia del progetto. Si vedano anche Savy, Cutugno 2009; Altre fonti utili: http://www.clips.unina.it/it/ [1.3] [2.3]

1.3. Albano Leoni, Giordano 2005: esperimento esemplare di analisi di un dialogo da più punti di vista (fonetico, prosodico, morfologico, sintattico

e pragmatico). Il dialogo è estratto dal corpus CLIPS (v. sopra).

1.4. ParVa: corpus realizzato a partire da interviste non raccolte per scopi di ricerca linguistica. Cfr. Guerini 2016, che si ispira al modello di [1.3], raccogliendo analisi autonome sullo stesso set di dati di partenza.

1.5. LABLITA (Cresti 2000) e C-ORAL-ROM (Cresti, Moneglia 2005); cfr. anche Cresti *et al.* 2018.

1.6. Corpus di italiano televisivo (CiT): Spina 2005.

1.7. Per una riflessione più ampia su analisi del parlato e sua grammaticalità, cfr. Voghera 2017.

1.8. Su strumenti di analisi quantitativa e qualitativa, cfr. *inter alia* Pallotti 2016; Iannàccaro, Ciccolone 2017.

1.9. Altre riflessioni di notevole rilievo metodologico: Vietti 2003; Vietti 2005; Spreafico in stampa.

1.10. Su corpora di italiano L2: Andorno, Rastelli 2009. [2.6]

1.11. Rassegne sui corpora di italiano: Crocco 2015; Cresti, Panunzi 2013.

2. Corpora di parlato in Italia: risorse online

2.1. *Parlaritaliano*: il sito raccoglie risorse, informazioni e riferimenti bibliografici relativi a corpora di parlato e ricerche linguistiche sul parlato nel panorama italiano. http://www.parlaritaliano.it/

2.2. Corpus VoLIP: permette di esplorare i dati del LIP, associando gli audio originali alla trascrizione. Cfr. Voghera *et al.* 2014. http://www.parlaritaliano.it/index.php/it/volip [1.1]

2.3. Corpus CLIPS: l'intero corpus è disponibile per analisi linguistiche. http://www.clips.unina.it/it/ [1.2]

2.4. AN.ANA.S_MT: corpus annotato sintatticamente di *map task* in italiano, inglese e spagnolo. Cfr. Cutugno, Voghera 2004; Voghera *et al.* 2005. http://www.parlaritaliano.it/index.php/it/corpora-di-parlato/716-corpus-ananas-multilingue-ananasmt [4.2]

2.5. PraTiD: corpus con annotazione pragmatica di dialoghi *task oriented* estratti dal corpus CLIPS. Cfr. Savy 2010. http://www.parlaritaliano.it/index.php/it/corpora-di-parlato/645-corpus-pratid [4.3]

2.6. *Banca dati di italiano L2* (Pavia): Andorno 2001. [1.10]

2.7. *Perugia Corpus*: Spina 2014. https://www.unistrapg.it/cqpwebnew/ [Spina *et al.*]

2.8. Corpus KIParla: Goria, Mauri 2018; https://kiparla.it/ [Ballarè, Mauri]

2.9. *Kontatto*: corpus di parlato bilingue (Alto Adige); Dal Negro, Ciccolone 2018. Accessibile su richiesta insieme a *Kontatti*. http://kontatti.projects.unibz.it/

2.10. DiWaC e ArchiWals (cfr. Angster *et al.*, ivi): http://www.archiwals.org/ [Angster *et al.*]
2.11. *Alpine Laboratory of Phonetic Sciences* (ALPS): https://alps.projects.unibz.it/ [Mereu, Vietti]
2.12. *Progetto CLODIS*: https://sites.google.com/site/progettoclodis/ [Diadori, Monami]
2.13. Su corpora di scritto e WaC (*web-as-corpus*) per l'italiano, cfr. Baroni *et al.* 2009 e le risorse disponibili su https://corpora.dipintra.it/
2.14. EVALITA: iniziativa relativa alla valutazione di strumenti e risorse computazionali per l'italiano. Cfr. *inter alia* Magnini *et al.* 2013; Basile *et al.* 2016. http://www.evalita.it/
2.15. Vademecum per il trattamento delle fonti orali: in preparazione; pubblicato in versione provvisoria sul sito dell'AISV; versione definitiva prevista nel corso del 2021. https://www.aisv.it/

3. Linee guida generali per la raccolta di dati di parlato

3.1. *Text Encoding Initiative* (TEI): consorzio per lo sviluppo e il mantenimento di standard per la codifica digitale di testi; TEI Consortium 2020. https://tei-c.org/guidelines/p5/
3.2. *Dublin Core Metadata Initiative*: protocolli standard per i metadati. https://dublincore.org/
3.3. Su accessibilità degli archivi sonori, cfr. Calamai *et al.* 2016. [2.15]
3.4. Per un manuale metodologico sulla ricerca sociolinguistica, cfr. Tagliamonte 2006.
3.5. Gibbon, Dafydd, Roger Moore, Richard Winski 1997, *Handbook of standards and resources for spoken language systems*. Berlin, Mouton de Gruyter. [Spreafico]
3.6. Wilkinson, Mark D. *et al.* 2016, "The FAIR Guiding Principles for scientific data management and stewardship". *Scientific Data* 3:160018. https://www.nature.com/articles/sdata201618. [Ferrarotti *et al.*]

4. Protocolli e convenzioni per la trascrizione e l'annotazione di dati di parlato

4.1. CLIPS: Savy 2007a, *Specifiche per la trascrizione ortografica annotata dei testi raccolti* (http://www.clips.unina.it/it/documenti/11_specifiche_trascrizione_ortografica.pdf); Savy 2007b, *Specifiche per l'etichettatura dei livelli segmentali* (http://www.clips.unina.it/it/documenti/12_specifiche_di_etichettatura.pdf) [1.2] [2.3]

4.2. AN.ANA.S.: protocollo di annotazione e analisi sintattica; Cutugno, Voghera 2004. http://www.parlaritaliano.it/index.php/it/strumenti/717-ananas-4 [2.4]
4.3. PraTiD: protocollo di annotazione pragmatica; Savy 2010. http://www.parlaritaliano.it/index.php/it/strumenti/668-pratid [2.5]
4.4. Jefferson 2004: convenzioni di trascrizione per l'analisi conversazionale.
4.5. GAT: sistema di trascrizione per l'analisi conversazionale. Couper-Kuhlen, Barth-Weingarten 2011; Schmidt *et al.* 2015. http://agd.ids-mannheim.de/gat.shtml
4.6. Chat-CLAN: software di trascrizione e annotazione (CLAN) e sistema per la trascrizione e codifica delle interazioni orali (CHAT). Cfr. MacWhinney 2000 per una descrizione dettagliata; per il manuale del sistema di trascrizione CHAT: https://talkbank.org/manuals/CHAT.pdf; altre risorse disponibili su https://childes.talkbank.org/ [5.3]
4.7. Crowdy, Steven 1994, "Spoken corpus transcription". *Literary and Linguistic Computing* 9: 25-28. [Spreafico]
4.8. Du Bois, John W., Susanna Cumming, Stephan Schuetze-Coburn, Danae Paolino 1992, *Santa Barbara Papers in Linguistics. IV. Discourse transcription*. Santa Barbara, Department of Linguistics (UCSB). [Ballarè, Mauri]
4.9. Nagy, Naomi, Devyani Sharma 2013, "Transcription". In: Robert J. Podesva, Devyani Sharma (eds), *Research Methods in Linguistics*. Cambridge, Cambridge University Press: 235-256. [Angster *et al.*]
4.10. Paternostro, Giuseppe 2007, "La trascrizione conversazionale". In: Vito Matranga (a cura di), *Trascrivere. La rappresentazione del parlato nell'esperienza dell'Atlante Linguistico della Sicilia*. Palermo, Centro di studi filologici e linguistici siciliani: 103-136. [Ferrarotti *et al.*]
4.11. Jenks, Christopher, 2011, *Transcribing Talk and Interaction.* Amsterdam/Philadelphia, John Benjamins. [Diadori, Monami]

5. Software e risorse digitali per il trattamento di dati di parlato e la creazione di corpora

5.1. ELAN: software per la trascrizione e annotazione multimodale del parlato. Cfr. Sloetjes, Wittenburg 2008. https://archive.mpi.nl/tla/elan
5.2. Praat: software per la trascrizione, annotazione e analisi acustica del parlato. Cfr. Boersma 2001; Boersma, Weenink 2018. https://www.fon.hum.uva.nl/praat/
5.3. CLAN: software di trascrizione e annotazione del parlato, usato dal progetto TalkBanks; usa la codifica CHAT. https://dali.talkbank.org/clan/ [4.7]

5.4. WebMaus: software per la segmentazione fonetica di un file multimediale a partire da una trascrizione ortografica. Cfr. Schiel 1999. Accessibile direttamente via web insieme a una serie di altri strumenti software messi a disposizione dal *Bavarian Archive for Speech Signals* (BAS), su cui cfr. Kisler *et al.* 2015, Kisler *et al.* 2017. https://clarin.phonetik.uni-muenchen.de/BASWebServices/interface [Mereu, Vietti]

5.5. EMU: per la creazione di database relazionali che permette di associare dati di corpora linguistici con i metadati. Si integra con R e Praat. Winckelmann *et al.* 2017. [Mereu, Vietti]

5.6. R: software di analisi statistica, con numerosi "pacchetti" di funzioni dedicate al *natural language processing* e al trattamento di dati di corpora linguistici. Tra i manuali più interessanti: Baayen 2008; Levshina 2015. https://www.r-project.org/

5.7. Natural Language Toolkit (NLTK): pacchetto di funzioni in linguaggio Python per il *natural language processing*. Bird *et al.* 2009. https://www.nltk.org/

5.8. McCarty, Christopher 2011, *EgoNet*. Software per la raccolta dati sulle reti sociali. https://sourceforge.net/projects/egonet/ [Mereu, Vietti]

Note

* Il presente contributo e la riflessione scientifica da cui deriva sono il frutto di una stretta collaborazione tra i due autori avviata con l'organizzazione del workshop *Corpora di parlato: verso l'individuazione di pratiche condivise*, tenutosi a Como il 20 settembre 2019 nel corso del LIII congresso della Società di Linguistica Italiana (SLI). Tuttavia, per quanto riguarda la redazione del testo, il paragrafo 1 è da attribuire a Eugenio Goria, il paragrafo 3 è da attribuire a Simone Ciccolone, mentre il paragrafo 2 è stato redatto da entrambi.

Riferimenti bibliografici

Albano Leoni, Federico 2007a, "Presentazione". In: www.clips.unina.it.

Albano Leoni, Federico 2007b, "Un frammento di storia recente della ricerca (linguistica) italiana. Il corpus CLIPS". *Bollettino d'Italianistica* (n.s.) 4: 122-130.

Albano Leoni, Federico, Rosa Giordano (a cura di) 2005, *Italiano parlato: analisi di un dialogo (con un cd-rom contenente il materiale audio variamente elaborato e altri materiali)*. Napoli, Liguori.

Andorno, Cecilia (a cura di) 2001, *Banca dati di italiano L2. Progetto di Pavia*. CD-ROM, Dipartimento di Linguistica, Università di Pavia.

Andorno, Cecilia, Rastelli, Stefano (a cura di) 2009, *Corpora di italiano L2. Tecnologie, metodi, spunti teorici*. Perugia, Guerra.

Baayen, Harald 2008, *Analyzing Linguistic Data. A practical introduction to statistics*. Cambridge, Cambridge University Press.

Baroni, Marco, Silvia Bernardini, Adriano Ferraresi, Eros Zanchetta 2009, "The WaCky wide web: a collection of very large linguistically processed web-crawled corpora". *Language Resources and Evaluation* 43/3: 209-226.

Basile, Pierpaolo, Franco Cutugno, Malvina Nissim, Viviana Patti, Rachele Sprugnoli (eds) 2016, *Proceedings of the 5th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian (EVALITA 2016)*. aAcademia University Press.

Berruto, Gaetano, Massimo Cerruti 2015, "Un esercizio di analisi variazionista: l'accordo verbale nel costrutto locativo-esistenziale-presentativo". In: Maria Grazia Busà, Sara Gesuato (a cura di), "Lingue e contesti. Studi in onore di Alberto M. Mioni", CLEUP, Padova, 609-620.

Bird, Steven, Ewan Klein, Edward Loper 2009, *Natural Language Processing with Python. Analyzing Text with the Natural Language Toolkit*. Beijing-Cambridge, O'Reilly.

Boersma, Paul 2001, "Praat, a system for doing phonetics by computer". *Glot International* 5/9-10: 341-345.

Boersma, Paul, David Weenink 2018, *Praat: doing phonetics by computer* (software: http://www.praat.org).

Calamai, Silvia, Veronique Ginouvès, Pier Marco Bertinetto 2016, "Sound Archives Accessibility". In: Karol Jan Borowiecki, Neil Forbes, Antonella Fresa (eds), *Cultural Heritage in a Changing World*. Berlin, Springer: 37-54.

Cerruti, Massimo, Silvia Ballarè in stampa, "Il parlato di Torino: introduzione al corpus ParlaTO". In: *Bollettino dell'Atlante Linguistico Italiano (BALI)* 44.

Couper-Kuhlen, Elizabeth, Dagmar Barth-Weingarten 2011, "A system for transcribing talk-in-interaction: GAT 2". *Gesprächsforschung - Online-Zeitschrift zur verbalen Interaktion* 12: 1-51.

Cresti, Emanuela 2000, *Corpus di italiano parlato*. 2 voll. Firenze, Accademia della Crusca.

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

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

Cresti, Emanuela, Massimo Moneglia, Alessandro Panunzi 2018, "The LABLITA Corpus & the Language into Act Theory: Analysis of Viterbo Excerpts". In: Amedeo De Dominicis (ed.), *Speech Audio Archives: Preservation, Restoration, Annotation Aimed at Supporting the Linguistic Analysis*. Roma, Bardi Edizioni: 47-63.

Crocco, Claudia 2015, "Corpora e testi di italiano contemporaneo". In: Maria Iliescu, Eugeen Roegiest (eds), *Manuel des anthologies, corpus et textes romans / Manual of Romance Anthologies, Corpora, and Texts*, vol. 7. Berlin, de Gruyter: 509-534.

Cutugno, Francesco, Miriam Voghera 2004, "Analisi sintattica e annotazione XML a contatto". In: Federico Albano Leoni, Francesco Cutugno, Massimo Pettorino, Renata Savy (a cura di), *Il parlato italiano*. Atti del convegno nazionale (Napoli, 13-15 febbraio 2003). Napoli, D'Auria Editore: 50-52.

D'Alessandro, Roberta 2018, "Il progetto *Microcontact*: l'eredità linguistica dei dialetti italiani". URL: https://www.treccani.it/magazine/lingua_italiana/speciali/Microcontact/D_Alessandro.html.

Dal Negro, Silvia, Simone Ciccolone 2018, "Il parlato bilingue: italiano e tedesco a contatto in un corpus sudtirolese". In: Felisa Bermejo Calleja, Peggy Katelhön (a cura di), *Lingua parlata. Un confronto fra l'italiano e alcune lingue europee*. Bern, Lang: 385-407.

De Mauro, Tullio, Federico Mancini, Massimo Vedovelli, Miriam Voghera 1993, *Lessico di frequenza dell'italiano parlato*. Milano, ETAS LIbri.

Goria, Eugenio, Caterina Mauri 2018, "Il corpus KIParla: una nuova risorsa per lo studio dell'italiano parlato". In: Francesca Masini, Fabio Tamburini (a cura di), *CLUB Working Papers in Linguistics*. *II*. Bologna, CLUB – Circolo Linguistico dell'Università di Bologna: 76-95.

Guerini, Federica (a cura di) 2016, *Italiano e dialetto bresciano in racconti di partigiani*. Roma, Aracne.

Iannàccaro, Gabriele, Simone Ciccolone 2017, "Italian quantitative Sociolinguistics: research directions". *Sociolinguistic Studies* 11/2-4: 365-387.

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

Kisler, Thomas, Florian Schiel, Uwe Reichel, Christoph Draxler 2015, "Phonetic/linguistic Web Services at BAS". *Interspeech* 2015: 2609-2610.

Kisler, Thomas, Uwe Reichel, Florian Schiel 2017, "Multilingual processing of speech via web services". *Computer Speech & Language* 45: 326-347.

Kortmann, B. (ed.) 2008, *Dialectology meets Typology. Dialect Grammar from a Cross-Linguistic Perspective*. Berlin, de Gruyter.

Levshina, Natalia 2015, *How to do Linguistics with R*. Amsterdam-Philadelphia, Benjamins.

Linell, Per 2005, *The Written Language Bias in Linguistics: Its nature, origins and transformations*. London, Routledge.

MacWhinney, Brian 2000, *The CHILDES Project: Tools for Analyzing Talk*. Mahwah, Lawrence Erlbaum.

Magnini, Bernardo, Francesco Cutugno, Mauro Falcone, Emanuele Pianta (eds) 2013, *Evaluation of Natural Language and Speech Tools for Italian. International Workshop, EVALITA 2011. Rome, January 24-25, 2012. Revised Selected Papers*. Berlin-Heidelberg, Springer.

Mauri, Caterina, Andrea Sansò (eds) 2018, *Linguistic strategies for the construction of ad hoc categories: synchronic and diachronic perspectives* (= *Folia Linguistica* 39/1).

Pallotti, Gabriele 2016, "Qualitativo/quantitativo: ripensare la distinzione". In: Francesca Gatta (a cura di), *Parlare insieme. Studi per Daniela Zorzi*. Bologna, Bononia University Press: 105-117.

Savy, Renata 2010, "Pr.A.T.I.D: a coding scheme for pragmatic annotation of dialogues". In: *Proceedings of LREC 2010*: 2141-2148.

Schiel, Florian 1999, "Automatic phonetic transcription of non-prompted speech". In: *Proceedings of the ICPhS 1999*: 607–610.

Schmidt, Thomas, Wilfried Schütte, Jenny Winterscheid 2015,"cGAT. Konventionen für das computergestützte Transkribieren in Anlehnung an das Gesprächsanalytische Transkriptionssystem 2 (GAT2)". In: *Datenbank für Gesprochenes Deutsch (DGD), FOLK*. Mannheim, IDS (http://dgd.ids-mannheim.de).

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

Spina, Stefania 2005, "Il Corpus di Italiano Televisivo (CiT): struttura e annotazione". In: Elisabeth Burr (a cura di), *Tradizione e innovazione. Il parlato. Teoria, corpora, linguistica dei corpora*. Atti del VI convegno della Società Internazionale di Linguistica e Filologia Italiana (Gerhad-Mercator Universität, Duisburg, 28 giugno - 2 luglio 2000). Firenze, Cesati: 413-426.

Spina, Stefania 2014, "Il Perugia Corpus: una risorsa di riferimento per l'italiano. Composizione, annotazione e valutazione". In: Roberto Basili, Alessandro Lenci, Bernardo Magnini (a cura di), *Proceedings of the First Italian Conference on Computational Linguistics CLiC-it 2014* (Pisa 9-10 dicembre 2014). Pisa, Pisa University Press: 354-359.

Spreafico, Lorenzo in stampa, "La trascrizione strumentale del significante". In: Giuliano Bernini, Ada Valentini, Lorenzo Spreafico, Jacopo Saturno (a cura di) in stampa, *Superare l'evanescenza del parlato. Un vademecum per il trattamento digitale di dati di lingua parlata*. Bergamo, Sestante.

Tagliamonte, Sali A. 2006, *Analysing sociolinguistic variation*. Cambridge, Cambridge University Press.

TEI Consortium 2020, *TEI P5: Guidelines for Electronic Text Encoding and Interchange* [versione 4.0]. http://www.tei-c.org/Guidelines/P5/.

Vietti, Alessandro 2003, "Come costruire una intervista 'ecologica': Per una interpretazione contestualizzata dei dati". In: Ada Valentini, Piera Molinelli, Pierluigi Cuzzolin, Giuliano Bernini (a cura di), *Ecologia linguistica*. Atti del XXXVI Congresso della Società di Linguistica Italiana. Roma, Bulzoni: 161-184.

Vietti, Alessandro 2005, "Approcci quantitativi nell'analisi della variazione sociolinguistica: Il caso di GOLDVARB 2001". *Linguistica e Filologia* 20: 31-69.

Voghera, Miriam 2017, *Dal parlato alla grammatica. Costruzione e forma dei testi spontanei*. Roma, Carocci.

Voghera, Miriam, Grazia Basile, Francesco Cutugno, Giuliana Fiorentino 2005, “Sintassi in AN.A-NA.S.”. In: Federico Albano Leoni, Rosa Giordano (a cura di), *Italiano parlato: analisi di un dialogo*. Napoli, Liguori: 189-211.

Voghera, Miriam, Claudio Iacobini, Renata Savy, Francesco Cutugno, Aurelio De Rosa, Iolanda Alfano 2014, “VoLIP: a searchable Italian spoken corpus”. In: Ludmila Veselovská, Markéta Janebová (eds), *Complex Visibles Out There. Proceedings of the Olomouc Linguistics Colloquium: Language Use and Linguistic Structure*. Olomouc, Palacký University: 628-640.

Winkelmann, Raphael, Jonathan Harrington, Klaus Jänsch 2017, “EMU-SDMS: Advanced speech database management and analysis in R”. *Computer Speech & Language* 45: 392-410.