Workshop 5: ***Corpora di parlato: verso l'individuazione di pratiche condivise***

Soci proponenti: Eugenio Goria, Università degli Studi di Torino; Simone Ciccolone, Libera Università di Bolzano

N.B.: La scadenza per l'invio di proposte è prorogata al 6 marzo 2019

Breve presentazione del contenuto atteso:

A più di dieci anni dalla pubblicazione del corpus CLIPS (cfr. Albano Leoni 2007a, 2007b), il panorama delle ricerche sul parlato in Italia risulta notevolmente cambiato (cfr. Crocco 2015 per una rassegna). Proprio progetti come CLIPS, C-ORAL-ROM, o più recentemente VoLIP (cfr. Cresti/Moneglia 2005, 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. Parallelamente, la disponibilità di nuove metodologie di lavoro e di nuovi strumenti di supporto nelle varie fasi dell'elaborazione del dato linguistico hanno facilitato notevolmente l'accesso a dati reali. Questo ha determinato a sua volta una crescita notevole dell'interesse verso il parlato, anche in prospettive teoriche tradizionalmente meno legate all'uso di dati empirici, come tipologia linguistica e grammatica generativa (cfr. Kortmann 2008, Mauri/Sansò 2018, D'Alessandro 2018).

Ciononostante, 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. Fino ad ora, infatti, gran parte della riflessione metodologica sui corpora si è concentrata su varietà scritte, per le quali sono inoltre disponibili 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 etc.). 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.

È perciò lecito chiedersi se proprio 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), che si dovrebbe estendere anche a quell'insieme di risorse "sommerse" e archivi sonori non (ancora) accessibili o disponibili alla comunità scientifica (cfr. Calamai/Ginouvés/Bertinetto 2016).

Date queste premesse, il workshop si propone di contribuire al dibattito relativo alle attuali risorse per lo studio di varietà orali, con particolare attenzione agli elementi di complessità che esse comportano. Si intende in particolare riflettere su una serie di questioni metodologiche fondamentali: l'obiettivo principale è di aprire un confronto critico tra prospettive teoriche diverse, che permetta di individuare soluzioni comuni nel trattamento dei dati di parlato, anche alla luce delle convenzioni adottate dalla comunità scientifica a livello internazionale (cfr. ad esempio i sistemi CHAT, Dubois, Jefferson, GAT, ecc.). Scopo ultimo della discussione è dunque quello di individuare un possibile standard qualitativo che faciliti la condivisione di risorse, strumenti e prassi di ricerca che tengano conto della natura "elastica" della modalità parlata (cfr. Voghera 2017: 189-198).

Pertanto, verranno privilegiati contributi che supportino una prospettiva ecologica alla ricerca e al dato linguistico, riflettendo sull'individuazione di buone pratiche capaci di rendere le risorse a disposizione del linguista *aperte* (ovvero condivisibili secondo una filosofia *open access*), *modulari* (organizzate in componenti dedicati a problemi specifici), *flessibili* (trasversali rispetto a singoli interessi di ricerca e adattabili ad approcci teorici diversi) ed *efficienti* (che prevedano cioè uno sforzo via via minore da parte della comunità scientifica per la loro implementazione e il loro mantenimento).

Le relazioni dei partecipanti dovranno dedicarsi non tanto alla presentazione di risultati di un'analisi specifica, quanto al proporre riflessioni o confrontare soluzioni su questi temi, cercando di porsi il problema di come i propri dati, il proprio corpus, le proprie trascrizioni e annotazioni possano essere utilizzati per ricerche con prospettive e obiettivi specifici diversi, incentivando l'analisi su più livelli dello stesso insieme di dati (si vedano l'esemplare esperimento pubblicato in Albano Leoni/Giordano 2005 e, più recentemente, Guerini 2016).
Si incoraggia quindi l'invio di abstract che propongano riflessioni intorno ai seguenti temi:

1. Raccolta dati in prospettiva ecologica:
1.1. qualità della registrazione;
1.2. equilibrio tra accuratezza e sostenibilità della raccolta dati;
1.3. struttura dei metadati su parlanti e situazione comunicativa;
1.4. pro e contro dell'utilizzo di risorse ottenute in *crowdsourcing*, tramite web e/o realizzate da non linguisti;

2. Trascrizione del parlato:
2.1. costi e benefici delle convenzioni diffuse a livello internazionale (GAT, Jefferson, Dubois, CLAN) e loro applicabilità alla situazione italiana;
2.2. affidabilità dei sistemi di trascrizione automatica di libero accesso;
2.3. equilibrio tra granularità e computabilità della trascrizione;
2.4. rappresentazione di varietà substandard o *dachlos*, di varietà di apprendimento, del parlato bi/plurilingue e della variazione interna all'evento comunicativo;

3. Annotazione dei corpora di parlato:
3.1. strumenti a disposizione per il trattamento automatico o semi-automatico di dati di parlato (POS tagging, lemmatizzazione, analisi acustica etc.) e trasponibilità degli strumenti disponibili per i corpora scritti;
3.2. riutilizzabilità e modularità dell'annotazione;

4. Trasversalità e condivisione delle risorse:
4.1. adattabilità di risorse costruite per scopi specifici a un modello condiviso;
4.2. possibile adattamento dei corpora di parlato ad ambiti tradizionalmente basati su materiali scritti o varietà standard (ad es. glottodidattica);
4.3. formato e infrastrutture per la condivisione di corpora di parlato (software, strumenti di interrogazione, allineamento con audio/video, accesso ai metadati);

5. Aspetti etici, deontologici e giuridici: problemi e proposte di soluzione riguardo a
5.1. liberatorie;
5.2. anonimizzazione dei dati;
5.3. pubblicazione e proprietà dei dati.

Relatore invitato: Lorenzo Spreafico (Libera Università di Bolzano)

Comitato Scientifico per la selezione delle proposte di intervento:
Cecilia Andorno, Gaetano Berruto, Silvia Calamai, Massimo Cerruti, Francesco Cutugno, Silvia Dal Negro, Gabriele Iannàccaro, Caterina Mauri, Lorenzo Spreafico, Alessandro Vietti, Miriam Voghera, Eugenio Goria, Simone Ciccolone

Bibliografia di riferimento
Albano Leoni, Federico (2007a), Presentazione. In: *Corpus CLIPS*. (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 / Giordano, Rosa (a c. di) (2005), *Italiano parlato: analisi di un dialogo (con un cdrom contenente il materiale audio variamente elaborato e altri materiali)*. Liguori, Napoli.

Andorno, Cecilia (2008), Segnali discorsivi: tra struttura dell'enunciato e struttura dell'interazione. In: Bernini, Giuliano / Spreafico, Lorenzo / Valentini, Ada (a c. di), *Competenze lessicali e discorsive nell'acquisizione di lingue seconde*. Guerra, Perugia: 481-510.
Andorno, Cecilia / Rastelli, Stefano (a c. di) (2009), *Corpora di Italiano L2: tecnologie, metodi, spunti teorici*. Guerra, Perugia.
Backus, Ad (2008), Data Banks and Corpora. In: Wei, Li / Moyer, Melissa G. (eds.), *The Blackwell Guide to Research Methods in Bilingualism and Multilingualism*. Blackwell, Oxford: 232–248. (https://doi.org/10.1002/9781444301120.ch13)
Baker, Paul (2006), *Using Corpora in Discourse Analysis*. Continuum, London/New York.
Baldry, Anthony (2005), *Multimodal transcription and text analysis: a multimedia toolkit and coursebook*. Equinox, London.
Basile, Pierpaolo / Cutugno, Francesco / Nissim, Malvina / Patti, Viviana / Sprugnoli, Rachele (eds.) (2016), *Proceedings of the 5th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian (EVALITA 2016)*. Accademia University Press, Torino.
Bazzanella, Carla (1994), *Le facce del parlare: un approccio pragmatico all'italiano parlato*. La Nuova Italia, Scandicci.
Berruto, Gaetano (2017), System-oriented and speaker-oriented approaches in Italian sociolinguistics. *Sociolinguistic Studies*, 11(2–4): 271–290. (https://doi.org/10.1558/sols.32864)
Bird, Steven / Klein, Ewan / Loper, Edward (2009), *Natural Language Processing with Python. Analyzing Text with the Natural Language Toolkit.* O'Reilly, Beijing/Cambridge.
Boersma, Paul / Weenink, David (2018), *Praat: doing phonetics by computer* [Computer program]. (http://www.praat.org).
Bowern, Claire (2008), *Linguistic fieldwork: a practical guide*. Palgrave Macmillan, Houndmills/ New York.
Calamai, Silvia / Ginouvès, Veronique / Bertinetto, Pier Marco (2016), Sound Archives Accessibility. In: Borowiecki, Karol Jan / Forbes, Neil / Fresa, Antonella (eds.), *Cultural Heritage in a Changing World*. Springer, Berlin/Heidelberg/New York: 37–54. (https://doi.org/ 10.1007/978-3-319-29544-2_3)
Chini, Marina (a c. di) (2004), *Plurilinguismo e immigrazione in Italia. Un'indagine sociolinguistica a Pavia e Torino*. Franco Angeli, Milano.
Cresti, Emanuela (2000), *Corpus di italiano parlato*. Accademia della Crusca, Firenze.
Cresti, Emanuela / Moneglia, Massimo (2005), *C-ORAL-ROM: integrated reference corpora for spoken Romance languages*. Benjamins, Amsterdam/Philadelphia.
Crocco, Claudia (2015), Corpora e testi di italiano contemporaneo. In: Iliescu, Maria / Roegiest, Eugeen (eds.), *Manuel des anthologies, corpus et textes romans / Manual of Romance Anthologies, Corpora, and Texts*, vol. 7. de Gruyter, Berlin: 509-534.
Cutugno, Francesco / Romano, Sara (2011), Time- and Text-Aligned Annotations: the SpLaSH Data Model. In: Yuan, J. (eds.), *Proceedings of New Tools and Methods for Very-Large-Scale Phonetics Research*. University of Pennsylvania, Philadelphia: 115–118.
Cutugno, Francesco / Voghera, M. (2004), Analisi sintattica e annotazione XML a contatto. In: Albano Leoni, Federico / Cutugno, Francesco / Pettorino, Massimo / Savy, Renata (a c. di), *Il parlato italiano. Atti del convegno nazionale - Napoli, 13-15 febbraio 2003*. D'Auria Editore, Napoli: 50–52.
Dal Negro, Silvia (2013), Dealing with bilingual corpora: Parts of speech distribution and bilingual patterns. *Revue Française de Linguistique Appliquée*, 18/2: 15-28.
D'Alessandro, Roberta (2018), Il progetto Microcontact: l'eredità linguistica dei dialetti italiani. (http://www.treccani.it/magazine/lingua_italiana/speciali/Microcontact/D_Alessandro.html)
D'Alessandro, Roberta / van Oostendorp, Marc (2017), On the Diversity of Linguistic Data and the Integration of the Language Sciences. *Frontiers in Psychology*, *8*. (https://doi.org/ 10.3389/fpsyg.2017.02002)
Du Bois, John W. (1991), Transcription design principles for spoken discourse research. *Pragmatics*, 1(1): 71–106. (https://doi.org/10.1075/prag.1.1.04boi)
Eskenazi, Maxine / Levow, Gina-Anne / Meng, Helen / Parent, Gabriel / Suendermann, David (a c. di) (2013), *Crowdsourcing for speech processing: applications to data collection, transcription, and assessment*. Wiley, Chichester.

Goria, Eugenio / Mauri, Caterina (in stampa), Il corpus KIParla: una nuova risorsa per lo studio dell'italiano parlato. In: Masini, Francesca / Tamburini, Fabio (a cura di), *CLUB Working Papers in Linguistics,* Vol. 2. Bologna: CLUB – Circolo Linguistico dell'Università di Bologna.
Grenoble, Lenore A. / Furbee-Losee, Louanna (eds.) (2010), *Language documentation: practice and values*. Benjamins, Amsterdam/Philadelphia.
Guerini, Federica (a cura di) (2016), *Italiano e dialetto bresciano in racconti di partigiani*. Aracne, Roma.
Hennecke, Inga (2018), Petits corpus oraux bilingues et plurilingues – enjeux théoriques et méthodologiques. *Corpus*, 18: 1-19. (http://journals.openedition.org/corpus/3451)
Iannàccaro, Gabriele (2000), Per una semantica più puntuale del concetto di dato linguistico: un tentativo di sistematizzazione epistemologica. *Quaderni di semantica*, 2000/1. (https://doi.org/ 10.1400/97181)
Jefferson, Gail (2004), Glossary of transcript symbols with an introduction. In: Lerner, Gene (a c. di), *Conversation Analysis: studies from the fist generation.* Pragmatics & Beyond new series. Amsterdam, Benjamins: 13-31.
Kisler, Thomas / Reichel, Uwe / Schiel, Florian (2017), Multilingual processing of speech via web services. *Computer Speech & Language*, *45*, 326–347. (https://doi.org/10.1016/j.csl.2017.01.005)
Kortmann, Bernd (ed.) (2008), *Dialectology meets Typology. Dialect Grammar from a Cross-Linguistic Perspective*. de Gruyter, Berlin. (https://www.degruyter.com/view/product/36468)
Léglise, Isabelle / Alby, Sophie (2016), Plurilingual corpora and polylanguaging, where corpus linguistics meets contact linguistics. *Sociolinguistic Studies*, 10(3): 357–381. (https://doi.org/ 10.1558/sols.v10i3.27918)
Linell, Per (2005), *The Written Language Bias in Linguistics: Its nature, origins and transformations*. Routledge, London.
LIPPS Group (2000), The LIDES Coding Manual: A document for preparing and analyzing language interaction data Version 1.1—July, 1999. *International Journal of Bilingualism*, *4*(2): 131–132. (https://doi.org/10.1177/13670069000040020101)
MacWhinney, Brian (2000), *The CHILDES Project: Tools for Analyzing Talk*. Lawrence Erlbaum, Mahwah.
Magnini, Bernardo / Cutugno, Francesco / Falcone, Mauro / Pianta, Emanuele (eds.) (2013), *Evaluation of Natural Language and Speech Tools for Italian. International Workshop, EVALITA 2011. Rome, January 24-25, 2012. Revised Selected Papers*. Springer, Berlin/Heidelberg.
Mauri, Caterina / Sansò, Andrea (2018), Strategie linguistiche per la costruzione on-line di categorie: un quadro tipologico. In: Brincat, Giuseppe / Caruana, Sandro (a c. di), *Tipologia e 'dintorni': il metodo tipologico alla intersezione di piani di analisi*. Bulzoni, Roma: 209-232.
Müller, Cornelia / Cienki, Alan J. / Fricke, Ellen / Ladewig, Silva H. / McNeill, David / Tessendorf, Sedinha (eds.) (2013), *Body - language - communication: an international handbook on multimodality in human interaction*. de Gruyter, Berlin/New York
Pons Bordería, Salvador (ed.) (2014), *Discourse segmentation in Romance languages*. Benjamins, Amsterdam/Philadelphia.
Reppen, Randi / Fitzmaurice, Susan M. / Biber, Douglas (eds.) (2002), *Using corpora to explore linguistic variation*. Benjamins, Amsterdam/Philadelphia.
Rossini Favretti, Rema (2001), La linguistica dei «corpora» in Europa: prospettive di analisi. *Lingua e Stile,* 2: 367-382. (https://doi.org/10.1417/11723)
Savy Renata (2010), Pr.A.T.I.D: a coding scheme for pragmatic annotation of dialogues. In: *Proceedings of LREC 2010*. Malta, 19-21 maggio 2010: 2141-2148.
Savy, Renata / Cutugno, Francesco (2009), CLIPS. Diatopic, diamesic and diaphasic variations in spoken Italian. In: *Proceedings of the Corpus Linguistics Conference 2009 (CL2009)*: 213/1-24.
Schiel, Florian (1999), Automatic phonetic transcription of non-prompted speech. In: *Proceedings of the ICPhS 1999*: 607–610.
Schiel, Florian / Stevens, Mary / Reichel, Uwe / Cutugno, Francesco (2013), Machine Learning of Probabilistic Phonological Pronunciation Rules from the Italian CLIPS Corpus. In: *Proc. Interspeech*: 1414–1418.
Schmidt, Thomas / Schütte, Wilfried / Winterscheid, Jenny (2015), *cGAT. Konventionen für das computergestützte Transkribieren in Anlehnung an das Gesprächsanalytische Transkriptionssystem 2 (GAT2)*. IDS, *Datenbank für Gesprochenes Deutsch (DGD), FOLK*(http://dgd.ids-mannheim.de).

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)*.
Sornicola, Rosanna / Calamai, Silvia (2014), Sound archives and linguistic variation. The case of Phlegraean diphtongs. In: Celata, Chiara / Calamai, Silvia (eds.), *Advances in Sociophonetics*. Benjamins, Amsterdam/Philadelphia: 169-185.
Spina, Stefania (2014), Il Perugia Corpus: una risorsa di riferimento per l'italiano. Composizione, annotazione e valutazione. *Politica*, *1*: 021.
Tagliamonte, Sali A. (2006), *Analysing Sociolinguistic Variation*. Cambridge University Press, Cambridge.
Vietti, Alessandro / Anselmi, Vittorio / Spreafico, Lorenzo (2015), Verso un sistema di riconoscimento automatico del parlato tramite immagini ultrasoniche. In: Vayra, Mario / Avesani, Cinzia / Tamburini, Stefano (a c. di), *Il farsi e il disfarsi del linguaggio. Acquisizione, mutamento e destrutturazione della struttura sonora del linguaggio*. AISV, Milano: 477–489.
Voghera, Miriam (2017). *Dal parlato alla grammatica. Costruzione e forma dei testi spontanei*. Carocci, Roma.
Voghera, Miriam / Basile, Grazia / Cutugno, Francesco / Fiorentino, Giuliana (2005), Sintassi in AN.ANA.S. In: Albano Leoni, Federico / Giordano, Rosa (a c. di), *Italiano parlato: analisi di un dialogo*. Liguori, Napoli: 189-211.
Voghera, Miriam / Iacobini, Claudio / Savy, Renata / Cutugno, Francesco / De Rosa, Aurelio / Alfano, Iolanda (2014). VoLIP: a searchable Italian spoken corpus. In: Veselovská, Ludmila / Janebová, Markéta (eds). *Complex Visibles Out There. Proceedings of the Olomouc Linguistics Colloquium: Language Use and Linguistic Structure*. Palacký University, Olomouc: 628-640.
Winkelmann, Raphael / Harrington, Jonathan / Jänsch, Klaus (2017), EMU-SDMS: Advanced speech database management and analysis in R. *Computer Speech & Language*, *45*: 392–410. (https://doi.org/10.1016/j.csl.2017.01.002)

Indicazioni per l'invio di proposte

Gli abstract dovranno avere una lunghezza massima di 500 parole circa, bibliografia esclusa, e dovranno essere inviati a Simone Ciccolone (simone.ciccolone@unibz.it) e a Eugenio Goria (egoria@unito.it) entro il **6 marzo 2019**.

Si ricorda che tutti i relatori al momento d'inizio del workshop dovranno essere soci regolari della SLI.