ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

10-12 settembre 2026

LIX Congresso internazionale di studi della Società di Linguistica Italiana

I DATI NELLE SCIENZE DEL LINGUAGGIO

Riassunti delle comunicazioni – Sessione generale

Sommario

ANTONIO BIANCO
Umorismo e comunicazione politica su X: un modello di annotazione per i tweet dei politici italiani 3

ARIANNA BIENATI, PAOLO BRASOLIN E FEDERICA COMINETTI
Dati, metodi e modelli per la verifica dell'ipotesi della semplificazione nell'italiano politico repubblicano 6

ELISABETTA BONVINO
Osservare l'apprendimento. I dati nella linguistica educativa 10

LORENZA BRASILE, FABIO ARDOLINO, SILVIA CALAMAI E ROSALBA NODARI
Archivi orali di ieri, dati linguistici di oggi: il caso della Sestino (AR) degli anni '80 12

BEATRICE COLCUC
Mappatura percettiva: terminologia, tipi di percezione, metodi di elicitazione e analisi dei dati 15

ALESSANDRA FAGGIOTTO
Il dato fonologico in prospettiva multimodale: il *cued speech* come rappresentazione visiva del segmento 18

TOMMASO LAMARRA
L'astrazione lessicale attraverso metodi sperimentali: dall'elaborazione psicolinguistica di parole in isolamento al loro uso in conversazioni naturalistiche 20

ALESSANDRO LENCI
Quali dati, quali teorie? Sfide e prospettive per la linguistica nell'era dei Large Language Model 23

LUCA LÒTANO
GoPro e creazione partecipata del dato: dal punto di vista incarnato alla video-analisi condivisa 24

KRISTEN LEONE
Elicitare il non standard: due studi di caso all'interfaccia tra semantica e sintassi 27

MARTA MAFFIA E MASSIMO PETTORINO

Quando i dati mancano: il parlato dal web nella ricerca sui *biomarker* ritmici del Parkinson 30

BORBALA SAMU ED ELISABETTA SANTORO
Il *Discourse Completion Task* come strumento di raccolta dati per la pragmatica (inter)linguistica e cross-culturale 34

SILVIA SILLERESI, ITAI BASSI, ABIGAIL ANNE BIMPEH, IMKE DRIEMEL, JOHNSON FỌLỌRUNṢỌ ILỌRI, ANASTASIA NUWORSU, GERALD OKEY NWEYA E MARIA TERESA GUASTI
Rethinking acceptability judgments in linguistic research: the case of logophoric pronouns in three West African languages 36

GIORGIA TROIANI, EUGENIO GORIA E BEATRICE BERNASCONI
Challenges and practices of corpus design in multilingual societies 39

LUISA TRONCONE, FLAVIO PISCIOTTA, IOLANDA ALFANO, ALFONSINA BUONICONTO, GIOVANNI DI PAOLA, GIUSEPPE PISEGNA, CARMELA SAMMARCO, ANDREA SANSÒ E MIRIAM VOGHERA
L'annotazione del dato come pratica interpretativa: il caso dei segnali discorsivi 42

PYTTI VALVERDE ROCHA DINIZ SILVA, LEONARDO WELPEL TEIXEIRA AND MARLON NUNES SILVA
School small data and synthetic data: retention, interoperability, and ethics for a situated linguistic data science 46

ROSSELLA VARVARA
Dati distribuzionali per la semantica derivazionale 49

ALESSANDRO VIETTI
L'incertezza strutturata: il ruolo della probabilità in linguistica 53

ANTONIO BIANCO

Umorismo e comunicazione politica su X: un modello di annotazione per i tweet dei politici italiani

Lo studio presenta un modello di annotazione multilivello di tweet umoristici pubblicati su X da politici italiani. Il modello è stato applicato a un corpus di 7.550 tweet (estratti automaticamente) pubblicati, durante la campagna elettorale del 2022, da 38 politici italiani (bilanciati per genere e appartenenza politica). Studi recenti hanno, difatti, esaminato l'uso dell'umorismo in altri scenari politici (es. Mendiburo-Seguel et al. 2022 per il contesto cileno). In ambito italiano, invece, gli studi si sono prevalentemente concentrati sul riconoscimento automatico di una specifica tipologia di umorismo: l'ironia (es. Cignarella et al. 2018).

Il modello di annotazione si articola in tre livelli di analisi, che forniscono una descrizione linguistica completa dei tweet umoristici e delle loro funzioni comunicative, consentendo analisi sia qualitative sia quantitative (Bianco 2026).

Il primo livello distingue i tweet umoristici da quelli non umoristici, adottando i criteri della *General Theory of Verbal Humor* (Attardo & Raskin 1991). In particolare, un testo è considerato umoristico quando presenta simultaneamente due o più *scripts* (idee, concetti) in opposizione tra loro (Attardo 2020). Il secondo livello riguarda l'annotazione delle cinque funzioni comunicative dell'umorismo (Martin et al. 2003): coesiva, aggressiva, difensiva, auto-denigratoria e (auto-) celebrativa. Nello specifico, un singolo tweet può contenere una sola istanza umoristica oppure più istanze con funzioni comunicative identiche o differenti. Questo livello consente, ad esempio, di osservare se i politici e i partiti utilizzino l'umorismo in modo differente nella comunicazione su X.

Inoltre, per i tweet umoristici è annotata anche la presenza o assenza di *humor markers* (Burgers & van Mulken 2017), classificati in: (a) markers tipografici (emoji, grassetto, etc.); (b) *markers* morfosintattici (segni di punteggiatura, morfologia valutativa, forme avverbiali); (c) casi di ripetizione e *code-switching*; (d) figure retoriche (iperbole, metafora, domanda retorica, litote). L'inclusione degli *humor markers*, non analizzati da Mendiburo-Seguel et al. (2022), ha permesso di individuare correlazioni statisticamente significative tra *markers* e specifiche funzioni comunicative (es. auto-deprecazione ed emoji).

Il modello è stato utilizzato per annotare i tweet sia in Excel sia nella piattaforma INCEpTION (Klie et al. 2018). La Figura 1 mostra l'annotazione di un tweet di D. Santanchè (FDI).

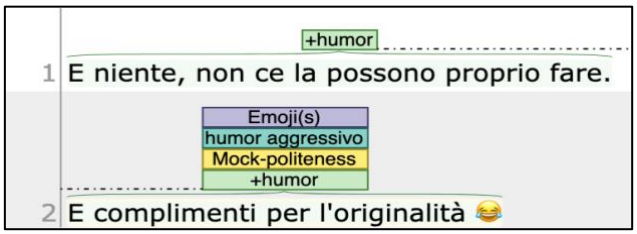


Figura 1: esempio di tweet annotato.

Il tweet umoristico (+humor) presenta un enunciato umoristico ("E complimenti per l'originalità 😂"), segnalato dall'emoji sorridente, interpretabile come una forma di *mock-politeness* (Taylor 2016) con funzione aggressiva. A ciascun enunciato possono naturalmente essere attribuiti più valori, ad esempio più di una funzione comunicativa.

Inoltre, saranno discussi i risultati del *task* di annotazione. Difatti, tre annotatori hanno applicato lo schema a un campione del 10% dei dati (Krippendorff 2019),[1] al fine di calcolare gli indici di accordo tra annotatori (Cohen's Kappa e Fleiss' Kappa) e individuare punti di forza e criticità del modello adottato. Lo schema è stato, inoltre, utilizzato per confrontare l'annotazione condotta da due annotatori umani e da due *Large Language Models* (GPT-4.0; Llama 3.3): le prestazioni di questi ultimi risultano soddisfacenti, in particolare nell'individuazione dei tweet umoristici con *humor markers*, mentre maggiori difficoltà emergono nel riconoscimento delle funzioni comunicative (Bianco, Brocca & Garassino 2026).

Il modello potrebbe essere applicato ad altri tipi di dati, inclusi testi con limiti di lunghezza maggiori rispetto a X (es. WhatsApp) o dati di parlato. Un'ulteriore prospettiva di ricerca riguarda l'estensione del modello di annotazione ad altre lingue. Infatti, l'applicazione dello schema ad altri generi testuali e ad altre lingue potrebbe rafforzarne replicabilità e la validità complessiva.

Riferimenti bibliografici

Attardo, Salvatore & Raskin, Victor. 1991. Script theory revis(it)ed: Joke similarity and joke representation model. *Humor: International Journal of Humor Research* 4(3–4). 293–347. https://doi.org/10.1515/humr.1991.4.3-4.293

Attardo, Salvatore. 2020. *The linguistics of humor: An introduction*. Oxford: Oxford University Press.

[1] Il primo annotatore ha annotato l'intero corpus (7.550 tweet).

Bianco, Antonio. 2026. *Humor and political discourse: A corpus-based study of Italian politicians on X*. Bergamo: Università di Bergamo. (Tesi di dottorato.)

Bianco, Antonio, Brocca, Nicola & Garassino, Davide. 2026. Assessing the Potential of LLM-assisted Annotation for Corpus Pragmatics: The Case of Humor. *Corpus Pragmatics* 10. 33. https://doi.org/10.1007/s41701-026-00235-7

Burgers, Christian & van Mulken, Margot. 2017. Humor Markers. In Attardo, Salvatore (a cura di), *The Routledge Handbook of Language and Humor*, 385–399. London: Routledge.

Cignarella, Alessandra Teresa, Bosco, Cristina, Patti, Viviana & Lai, Mirko. 2018. Application and Analysis of a Multi-layered Scheme for Irony on the Italian Twitter Corpus TWITTIRÒ. In Calzolari, Nicoletta, Choukri, Khalid, Cieri, Christopher, Declerck, Thierry, Goggi, Sara, Hasida, Koiti, Isahara, Hitoshi, Maegaard, Bente, Mariani, Joseph, Mazo, Hélène, Moreno, Asuncion, Odijk, Jan, Piperidis, Stelios & Tokunaga, Takenobu (a cura di), *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, *Miyazaki, 7-12 maggio 2018,* 4204–4211. Paris: European Language Resources Association (ELRA). https://aclanthology.org/L18-1664/

Klie, Jan-Christoph, Bugert, Michael, Boullosa, Beto, Eckart de Castilho, Richard & Gurevych, Iryna. 2018. The INCEpTION Platform: Machine-Assisted and Knowledge-Oriented Interactive Annotation. In Zhao, Dongyan (a cura), *Proceedings of System Demonstrations of the 27th International Conference on Computational Linguistics (COLING 2018)*, *Santa Fe (New Mexico) 20-26 agosto 2018,* 5–9. Stroudsburg: Association for Computational Linguistics. https://aclanthology.org/C18-2000/

Krippendorff, Klaus. 2019. *Content Analysis: An Introduction to Its Methodology*. New York: SAGE Publications.

Martin, Rod A., Puhlik-Doris, Patricia, Larsen, Gwen, Gray, Jeanette & Weir, Kelly. 2003. Individual differences in uses of humor and their relation to psychological well-being: Development of the Humor Styles Questionnaire. *Journal of Research in Personality* 37(1). 48–75. http://doi.org/10.1016/S0092-6566(02)00534-2

Mendiburo-Seguel, Andrés, Alenda, Stephanie, Ford, Thomas, Olah, Andrew, Navia, Patricio & Argüello-Gutiérrez, Catalina. 2022. *#funnypoliticians*: How Do Political Figures Use Humor on *Twitter*? *Frontiers in Sociology* 7. 788742. http://doi.org/10.3389/fsoc.2022.788742

Taylor, Charlotte. 2016. *Mock-Politeness in English and Italian*. Amsterdam: John Benjamins.

ARIANNA BIENATI, PAOLO BRASOLIN E FEDERICA COMINETTI

Dati, metodi e modelli per la verifica dell'ipotesi della semplificazione nell'italiano politico repubblicano

L'ipotesi che l'italiano politico repubblicano si sia semplificato nel corso del tempo costituisce un assunto molto diffuso nella letteratura scientifica sul tema (cfr. Antonelli 2007; Bolasco, Giuliano & Galli de' Paratesi 2006; Gualdo & Dell'Anna 2004) e nell'opinione pubblica. Solo recentemente tale ipotesi ha iniziato a ricevere conferme empiriche attraverso analisi *corpus-based* su larga scala (Bienati, Brasolin & Cominetti *in stampa*; Brasolin, Cominetti & Bienati *in revisione*; Cominetti, Bienati & Brasolin *in stampa*).

Il presente contributo estende questa linea di ricerca sfruttando la base di dati già analizzata per capire se e in che modo le stime probabilistiche dei *large language models* (LLM) possono essere impiegate per misurare la complessità linguistica in prospettiva diacronica. Questa domanda di ricerca si iscrive in una questione metodologica più ampia, e in questo momento fondamentale (Futrell & Mahowald 2025): quale ruolo hanno i *large language models* nello sviluppo e nella convalida empirica di teorie linguistiche?

Nello specifico, consideriamo un tipo particolare di dato estraibile dagli LLMs, il *surprisal*. Espresso matematicamente come il logaritmo dell'inverso della probabilità che il modello assegna all'occorrere di una parola dato il suo contesto precedente, esso si riferisce alla sua imprevedibilità o "sorpresa", la quale correla con la difficoltà di *processing* umana. Studi empirici hanno osservato infatti una correlazione lineare, significativa e stabile a livello cross-linguistico tra il *surprisal* estratto da modelli neurali e lo sforzo cognitivo (Smith & Levy 2013; Wilcox et al. 2024). Tuttavia, se per l'inglese sono stati esplorati in modo sistematico i parametri del modello (grandezza, regime di *pre-training*) che garantiscono il miglior allineamento con misure di difficoltà cognitiva (Oh & Schuler 2023), per l'italiano questo tipo di esplorazioni ancora manca.

Sulla base del *surprisal* si possono calcolare metriche di complessità basate sulla teoria dell'informazione come (i) la cross-entropia, intesa come misura dello scarto tra la distribuzione dell'informazione in un testo e quella appresa dal modello (Jurafsky & Martin 2026), e (ii) l'indice di *information fluctuation complexity* (Bates & Shepard 1993), che quantifica la variabilità locale del *surprisal*. Un testo può essere considerato tanto semplice quanto più si conforma alla distribuzione probabilistica appresa dal modello (cross-entropia bassa) e quanto più uniformemente è distribuita l'informazione (*fluctuation complexity* bassa). Per verificare l'ipotesi della semplificazione

nell'italiano politico repubblicano, queste metriche sono state calcolate sui testi del corpus IMPAQTS (Cominetti et al. 2024), composto da 1500 discorsi politici italiani dal 1946 a oggi.

I risultati preliminari ottenuti con il modello Minerva 350M (Orlando et al. 2024) mostrano (i) una riduzione della cross-entropia dalla Prima alla Terza Repubblica e (ii) una maggiore uniformità nella distribuzione dell'informazione, in accordo con l'ipotesi della *Uniform Information Density* (Jaeger 2010). Tali evidenze precisano e rafforzano i risultati precedenti (Bienati, Brasolin & Cominetti *in stampa*; Cominetti, Bienati & Brasolin *in stampa*), distinguendo con maggiore chiarezza l'andamento diacronico dell'entropia e le differenze tra diversi tipi di monologo politico. Di fatto, aggiungono un'ulteriore conferma dell'ipotesi della semplificazione in termini di riduzione della difficoltà cognitiva.

Il contributo, infine, discute criticamente le potenzialità e i limiti dell'approccio proposto: in particolare, il grado di dipendenza dei risultati dal modello scelto, la loro interpretabilità e la loro generalizzabilità alla varietà linguistica dell'italiano politico repubblicano. Più in generale, il lavoro si propone di offrire una serie di buone pratiche per l'uso di LLM nella ricerca linguistica, sottolineando l'importanza di trasparenza, replicabilità e consapevolezza epistemologica (Liesenfeld & Dingemanse 2024) nell'impiego di modelli neurali per la misurazione della complessità.

Riferimenti bibliografici

Antonelli, Giuseppe. 2007. *L'italiano nella società della comunicazione*. Bologna: Il Mulino.

Bates, John E. & Shepard, Harvey K. 1993. Measuring complexity using information fluctuation. *Physics Letters A* 172(6). 416–425. https://doi.org/10.1016/0375-9601(93)90232-O

Bienati, Arianna, Brasolin, Paolo & Cominetti, Federica. *In stampa*. Il potere in parole povere: un'analisi diacronica corpus-based sulla semplificazione dell'italiano politico repubblicano. In Valenti, Iride, Emmi, Tiziana, Fontana, Sabina & Menza, Salvatore (a cura di), *Varietà parlate, segnate, scritte: nuovi paradigmi interpretativi tra sincronia e diacronia*. *Atti del LVII Congresso della Società di Linguistica Italiana*, *Catania, 19-21 settembre 2024*. Milano: Officinaventuno.

Bolasco, Sergio, Giuliano, Luca & Galli de' Paratesi, Nora. 2006. *Parole in libertà. Un'analisi statistica e linguistica*. Roma: Manifesto Libri.

Brasolin, Paolo, Cominetti, Federica & Bienati, Arianna. *In revisione*. Legge di Heaps e ricchezza lessicale nella storia della Repubblica Italiana. In Leone, Paola & Chiusaroli, Francesca (a cura

di), La semplificazione linguistica per la comunicazione: tecnologie in contesto. *Atti del XXV Congresso dell'Associazione Italiana di Linguistica Applicata, Macerata, 19-21 febbraio 2025*. Milano: Officinaventuno.

Cominetti, Federica, Bienati, Arianna & Brasolin, Paolo. *In stampa*. La semplificazione lessicale e stilistica dei comizi politici nella storia repubblicana: un'analisi qualitativa e quantitativa. In Riccio, Anna (a cura di), *"La Comunicazione Parlata / Spoken Communication", Atti del congresso internazionale del Gruppo di Studio sulla Comunicazione Parlata (GSCP), Foggia, 19-21 giugno 2025.* Roma: Aracne.

Cominetti, Federica, Gregori, Lorenzo, Lombardi Vallauri, Edoardo & Panunzi, Alessandro. 2024. IMPAQTS: a multimodal corpus of Italian political discourse pragmatically annotated per implicitly conveyed questionable content. In Fišer, Darja, Eskevich, Maria & Bordon, David (a cura di), *Proceedings of the IV Workshop on Creating, Analysing, and Increasing Accessibility of Parliamentary Corpora (ParlaCLARIN) @ LREC-COLING 2024*, *Torino, 20 maggio 2024,* 101–109. Paris: European Language Resources Association (ELRA).

Futrell, Richard & Mahowald, Kyle. 2025. How Linguistics Learned to Stop Worrying and Love the Language Models. *Behavioral and Brain Sciences*. 1–98. https://doi.org/10.1017/S0140525X2510112X

Gualdo, Riccardo & Dell'Anna, Maria V. 2004. *La "faconda" Repubblica. La lingua della politica italiana (1992-2004)*. San Cesario di Lecce: Manni.

Jaeger, Florian T. 2010. Redundancy and reduction: Speakers manage syntactic information density. *Cognitive Psychology* 61(1). 23–62. https://doi.org/10.1016/j.cogpsych.2010.02.002

Jurafsky, Daniel & Martin, James H. 2026. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models*, *3rd edition draft* (versione online pubblicata il 6 gennaio 2026). https://web.stanford.edu/~jurafsky/slp3

Liesenfeld, Andreas & Dingemanse, Mark. 2024. Rethinking open source generative AI: Open-washing and the EU AI Act. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24)*, *Rio de Janeiro, 3-6 giugno 2024*, 1774–1787. New York: ACM - Association for Computing Machinery. https://doi.org/10.1145/3630106.3659005

Oh, Byung-Doh & Schuler, William. 2023. Transformer-Based Language Model Surprisal Predicts Human Reading Times Best with About Two Billion Training Tokens. In Bouamor, Houda, Pino, Juan & Bali, Kalika (a cura di), *Findings of the Association for Computational Linguistics: EMNLP 2023*, *Singapore*, *6-10 dicembre 2023*, 1915–1921. Stroudsburg: Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.findings-emnlp.128.

Orlando, Riccardo, Moroni, Luca, Huguet Cabot, Pere-Lluís, Conia, Simone, Barba, Edoardo, Orlandini, Sergio, Fiameni, Giuseppe & Navigli, Roberto. 2024. Minerva LLMs: The First Family of Large Language Models Trained from Scratch on Italian Data. In Dell'Orletta, Felice, Lenci, Alessandro, Montemagni, Simonetta & Sprugnoli, Rachele (a cura di), *Proceedings of the Tenth Italian Conference on Computational Linguistics (CLiC-it 2024)*, *Pisa, 4-6 dicembre 2024,* 707–719. Aachen: CEUR Workshop Proceedings. https://aclanthology.org/2024.clicit-1.77/

Smith, Nathaniel J. & Levy, Roger P. 2013. The effect of word predictability on reading time is logarithmic. *Cognition* 128(3). 302–319. https://doi.org/10.1016/j.cognition.2013.02.013

Wilcox, Ethan G., Pimentel, Tiago, Meister, Clara, Cotterell, Ryan & Levy, Roger P. 2023. Testing the Predictions of Surprisal Theory in 11 Languages. *Transactions of the Association for Computational Linguistics* 11. 1451–1470. https://doi.org/10.1162/tacl_a_00612

ELISABETTA BONVINO

Osservare l'apprendimento. I dati nella linguistica educativa

Nella ricerca in linguistica educativa e acquisizionale, il dato costituisce al tempo stesso il punto di partenza dell'indagine empirica e il principale terreno su cui si misurano le scelte teoriche e metodologiche. L'apprendimento linguistico è per sua natura un processo cognitivo non direttamente osservabile: il ricercatore ha accesso non all'acquisizione in sé, ma a tracce, produzioni, comportamenti, interazioni, tempi di risposta e pratiche didattiche che richiedono un'interpretazione teoricamente fondata. Il dato non coincide mai con il fenomeno che intende rappresentare, ma ne costituisce sempre una registrazione mediata da scelte concettuali e operative. Il dato, pertanto, non è mai neutro: è il risultato di un processo di selezione, costruzione e interpretazione che ne condiziona il significato e l'utilizzabilità.

Il presente contributo si propone di esplorare lo statuto, la natura e le funzioni del dato nella linguistica educativa contemporanea, interrogandosi su che cosa venga considerato "dato" nella ricerca del settore, su come i dati vengano raccolti, trattati e interpretati e, soprattutto, su quale ruolo essi assumano nella costruzione della conoscenza scientifica. A partire da una definizione operativa trasversale - che considera dato qualsiasi elemento osservabile e misurabile, dalla produzione linguistica al comportamento metacognitivo, dall'interazione in classe al questionario sociobiografico - il contributo ripercorre alcune linee di sviluppo della ricerca italiana in linguistica educativa che, negli ultimi decenni, hanno posto i dati al centro dell'analisi. Vengono presi in considerazione studi che utilizzano dati raccolti in contesti scolastici, universitari e migratori: corpora di apprendenti, registrazioni di interazioni, produzioni scritte e orali, protocolli metacognitivi, dati longitudinali, approcci quantitativi e qualitativi. La distinzione tra dati primari (evidenze grezze: registrazioni audio/video, interazioni spontanee), secondari (trascrizioni, codifiche, annotazioni) e terziari (metadati sul contesto e sulle condizioni di raccolta) evidenzia come il dato sia sempre il prodotto di operazioni interpretative stratificate, e non un'entità "data in sé".

Particolare attenzione è dedicata ai *learner corpora* e alle metodologie *corpus-based*, per il loro contributo al rafforzamento di una prospettiva empirica nella disciplina. L'analisi sistematica delle produzioni degli apprendenti ha permesso infatti di descrivere le varietà di interlingua, individuare sequenze acquisizionali, osservare fenomeni di variabilità e identificare aree di difficoltà per specifiche popolazioni. L'integrazione tra ricerca *corpus-based* e pratica didattica ha inoltre dato impulso ad approcci che coinvolgono gli apprendenti nell'esplorazione guidata di dati autentici per promuovere la scoperta induttiva di pattern lessicali e morfosintattici.

L’ambito della valutazione linguistica evidenzia come i dati (prove, punteggi, prestazioni annotate) costituiscano una componente centrale per garantire validità, affidabilità e trasparenza: la loro gestione in ambienti digitali integrati e l’eventuale supporto di sistemi di intelligenza artificiale richiedono procedure controllate, supervisione esperta e attenzione alla tracciabilità e all’equità dei processi valutativi.

Il contributo riflette anche sulla crescente eterogeneità dei dati impiegati nella ricerca contemporanea e sulla conseguente necessità di una consapevolezza metodologica esplicita nelle procedure di raccolta, annotazione e interpretazione.

LORENZA BRASILE, FABIO ARDOLINO, SILVIA CALAMAI E ROSALBA NODARI

Archivi orali di ieri, dati linguistici di oggi: il caso della Sestino (AR) degli anni ’80

Gli archivi orali sono risorse dal grande potenziale documentario (tipicamente includono, infatti, registrazioni vocali o sonore, metadati, informazioni socio-biografiche sui parlanti, appunti dal campo), ma in larga parte inutilizzate nella ricerca dialettologica: si pensi non solo a quelli frutto di ricerche di questo stesso settore, ma anche a quelli di area antropologica, sociologica o storica, i cui materiali forniscono dati linguistici altrettanto preziosi. Tali archivi possono consentire, fra le altre cose: i) studi diacronici basati su dati di parlato, e conseguenti descrizioni dialettologiche e geolinguistiche più accurate e non mediate da trascrizioni; ii) studi con metodi quantitativi, nel caso di campagne d’indagine svoltesi con molti soggetti diversi, e/o in contesti e momenti diversi, e/o su argomenti diversi; iii) descrizioni e analisi più fedeli alla realtà complessiva della/e comunità indagata/e, grazie a un campione di parlanti socialmente più vario che nelle indagini dialettologiche tradizionali (le quali, se affidate ai cosiddetti NORM, com’era frequente soprattutto nel secolo scorso, finiscono per restituire un’immagine piuttosto filtrata e parziale del dialetto locale, nonché della variazione diatopica generale; cfr. Calamai 2023; Watt, Renwick & Stanley 2023). L’ascolto degli archivi orali permette, dunque, di accedere a voci da tempi, luoghi e gruppi sociali diversi, nonché di integrare le prospettive dialettologiche e geolinguistiche tradizionali.

Il contributo proposto si concentra sul caso di Sestino, comune della provincia di Arezzo al confine con le Marche e l’Emilia-Romagna. In particolare, descrive e analizza la varietà di italiano locale che vi si parlava negli anni ’80, accessibile oggi grazie al recupero e alla digitalizzazione dell’*Archivio orale dell’Istituto Interregionale di Studi e Ricerche della Civiltà Appenninica* (cfr. Ardolino, Brasile & Calamai 2025; Dini 1990). Frutto di un’indagine etno-antropologica, l’Archivio comprende per la sola Sestino circa 300 ore di parlato semispontaneo da 158 parlanti di età, sesso e livello di istruzione differenti: un coro di voci che permette di conoscere non solo i fenomeni linguistici presenti, ma anche la loro incidenza nel parlato del singolo, la diffusione nella comunità, quindi le abitudini linguistiche effettive dei parlanti e, infine, i processi di koinizzazione in un contesto periferico rispetto ai principali (e per Sestino molteplici) centri di influenza linguistica.

Più precisamente, dalle registrazioni sestinesi emergono: i) l’uso dell’italiano (per quanto diatopicamente marcato) da parte di una comunità montana isolata, compresi gli anziani (classi 1890-1915) poco o per nulla scolarizzati; ii) la presenza, in tale varietà, di tratti linguistici da tutte e tre le aree dialettali circostanti (toscana, romagnola e perimediana), con prevalenza di quelli toscani rurali

o popolari nella morfologia e di quelli anti-toscani nella fonetica; iii) la tendenza a "modulare" l'incidenza dei tratti dialettali, piuttosto che al *code-switching*.

Dopo aver presentato brevemente l'Archivio, il suo recupero e la metodologia adottata in questo caso di riuso, la comunicazione illustrerà le caratteristiche dell'italiano parlato sestinese e lo confronterà con la letteratura dialettologica e sociolinguistica sulla località (Deli 1987; Vitali 2020: 380-386), sulla situazione toscana (ad es. Calamai 2017; Giacomelli 1975; Giannelli 1984, 2023) e sulla diffusione dell'italiano a livello orale (ad es. Bianconi 2003; Bruni 2007; Castellani 1982; De Mauro [1963] 1991, 2014; Serianni [1997] 2002). Si vedrà confermata l'importanza del riuso degli archivi orali, i cui dati sia arricchiscono in modo significativo la letteratura esistente, con la descrizione di un nuovo punto nel tempo e nello spazio, cui in futuro si potranno anche confrontare nuovi dati, sia la verificano in modo puntuale.

Riferimenti bibliografici

Ardolino, Fabio, Brasile, Lorenza & Calamai, Silvia. 2025. Uno stress test per il *Vademecum per il trattamento delle fonti orali*: questioni legali ed etiche nel recupero e nel riuso dell'Archivio orale dell'Istituto Interregionale di Studi e Ricerche della Civiltà Appenninica. *L'Analisi Linguistica e Letteraria* 33(3). 19–32.

Bianconi, Sandro. 2003. "La nostra lingua italiana comune" ovvero: la "strana questione" dell'italofonia preunitaria. In Marcato, Gianna (a cura di), *Italiano. Strana lingua? Atti del Convegno di Sappada/Plodn, Belluno, 3-7 luglio 2002*, 5–16. Padova: Unipress.

Bruni, Francesco. 2007. Per la vitalità dell'italiano preunitario fuori d'Italia. I. Notizie sull'italiano nella diplomazia internazionale. *Lingua e Stile* 2. 189–242. https://www.rivisteweb.it/doi/10.1417/25850

Calamai, Silvia. 2017. Tuscan between Standard and Vernacular: A Sociophonetic Perspective. In Cerruti, Massimo, Crocco, Claudia & Marzo, Stefania (a cura di), *Towards a New Standard: Theoretical and Empirical Studies on the Restandardization of Italian*, 213–241. Boston/Berlin: De Gruyter Mouton.

Calamai, Silvia. 2023. Sociophonetics and oral history. In Strelluf, Christopher (a cura di), *The Routledge handbook of sociophonetics*, 365–382. London: Routledge.

Castellani, Arrigo. 1982. Quanti erano gli italofoni nel 1861?. *Studi di linguistica italiana* VIII. 3–26.

De Mauro, Tullio. 1991. *Storia linguistica dell'Italia unita*. *2ª edizione*. Roma/Bari: Laterza.

De Mauro, Tullio. 2014. *Storia linguistica dell'Italia repubblicana: dal 1946 ai nostri giorni*. Roma/Bari: Laterza.

Deli, Annarosa. 1987. Tecniche e modalità d'indagine dialettologica: le aree dell'alto Marecchia e dell'alto Foglia a confronto. Come inserire l'indagine dialettale nella scuola dell'obbligo. *Formazione e società* 16. 193–198.

Dini, Vittorio. 1990. *Luoghi e voci della memoria collettiva. Per un archivio dei saperi e dei vissuti della cultura valtiberina toscana. Documentazione raccolta nei comuni di Sestino e Monterchi dall'anno 1978*. San Giovanni Valdarno: Litografia Valdarnese.

Giacomelli, Gabriella. 1975. Dialettologia toscana. *Archivio glottologico italiano* 60. 79–191.

Giannelli, Luciano. 1984. La recente evoluzione linguistica in Toscana. In *Dal dialetto alla lingua. Atti del IX Convegno per gli studi dialettali italiani, Lecce, 28 settembre-1 ottobre 1972*, 247–256. Pisa: Pacini.

Giannelli, Luciano. 2023. Volendo riscrivere la dialettologia. In Giannelli, Luciano, *Scritti vagabondi. Linguistica e dintorni*, 15–32. Arezzo: Helicon.

Serianni, Luca. 2002. Lingua e dialetti d'Italia nella percezione dei viaggiatori sette-ottocenteschi. In Serianni, Luca, *Viaggiatori, musicisti, poeti: saggi di storia della lingua italiana*, 55–88. Milano: Garzanti.

Watt, Dominic, Renwick, Margaret E. L. & Stanley, Joseph A. 2023. Sociophonetics and dialectology. In Strelluf, Christopher (a cura di), *The Routledge handbook of sociophonetics*, 263–284. London: Routledge.

Vitali, Daniele. 2020. *Dialetti emiliani e dialetti toscani. Le interazioni linguistiche fra Emilia-Romagna e Toscana e con Liguria, Lunigiana e Umbria*. Vol. 1, La Toscana e il confine con l'Umbria e la Romagna. Bologna: Pendragon.

BEATRICE COLCUC

Mappatura percettiva: terminologia, tipi di percezione, metodi di elicitazione e analisi dei dati

Nell'ambito delle indagini percettive sul linguaggio, l'espressione "dato percettivo" è usata in senso ampio per indicare materiali provenienti da metodi di elicitazione tra loro diversi, come test con stimoli, mappe mentali, questionari, interviste, commenti metalinguistici e giudizi di similarità (Canobbio & Iannàccaro 2000; Sauer & Hoffmeister 2022). Tuttavia, se non si precisano tipologie di percezione e condizioni dell'osservazione, dati di natura diversa finiscono per essere interpretati nello stesso quadro empirico (cfr. Becker, Herling & Wochele 2024). Come notavano già Krefeld e Pustka (2010), la percezione è un costrutto non sempre definito in modo univoco. In questa direzione, il modello di Iannàccaro (2015, 24) distingue tra Percezione$_1$, legata al "sentire" linguistico e alle valutazioni, e Percezione$_2$, legata alla rappresentazione sociale della comunità attraverso la lingua. Resta però aperta la questione operativa: tali distinzioni non specificano ancora quali scelte di raccolta dati rendano osservabili i tipi di percezione in modo comparabile.

Il contributo, a partire dalla pluralità di metodi e dati, amplia la riflessione alla costruzione dei compiti concreti ai quali vengono sottoposti i partecipanti alle inchieste. Anche a parità di tecnica, infatti, ciò che viene misurato dipende dall'impostazione della ricerca e dagli input concreti che incidono sulla risposta. Conta, anzitutto, la prospettiva adottata, interna alla comunità oppure esterna, secondo la distinzione emico ed etico (Pike 1967). Entro questa cornice si innestano scelte come la formulazione della consegna, la presenza o assenza di etichette, categorie chiuse o aperte, presentazione dello stimolo, cornice comunicativa e, nei compiti cartografici, caratteristiche della mappa di base. La stessa procedura può infatti produrre dati più vicini a Percezione$_1$ oppure a Percezione$_2$. Ad esempio, domande come "che caratteristiche ha la lingua X?" o "dove si parla la lingua X?" attivano spesso conoscenze disponibili e rappresentazioni sociali (Iannàccaro & Dell'Aquila 2001). Le stesse domande, se riferite a uno stimolo audio non etichettato, possono invece ancorare le risposte a proprietà linguistiche percepite nello stimolo e fornire indicazioni su ciò che i parlanti effettivamente sentono. Questo livello percettivo è rilevante sia per evitare descrizioni basate solo su categorie elaborate dal linguista, sia per comprendere meglio le strategie comunicative dei parlanti (Colcuc 2025; Colcuc & Cassisi 2024; Fiori 2023; Iannàccaro 2002). Anche quando tali scelte metodologiche sono note nella pratica, la loro mancata formalizzazione in uno schema può rendere più complessa la comparazione e l'analisi dei dati.

La mappatura viene discussa attraverso esempi che mostrano come variazioni nell'impostazione del compito spostino l'interpretazione del dato tra Percezione$_1$ e Percezione$_2$. Sulla

base di questa impostazione, il contributo propone uno schema che mette in relazione tipi di percezione, principali fenomeni studiati dalla linguistica percettiva (per esempio i confini percepiti e rappresentati, i tratti salienti, l'auto ed eteropercezione), metodi di raccolta, caratteristiche degli *input*, tipi di dati e strategie di analisi. La proposta ha un duplice scopo: da un lato, rendere più chiari i rapporti tra fenomeni, metodi, dati e terminologia all'interno degli studi percettivi; dall'altro, mostrare come, in ottica integrativa, i dati percettivi contribuiscano alla descrizione dei fenomeni linguistici.

Riferimenti bibliografici

Becker, Lidia, Herling, Sandra & Wochele, Holger (a cura di). 2024. *Manuel de linguistique populaire*. Berlin/Boston: De Gruyter.

Canobbio, Sabina & Iannàccaro, Gabriele (a cura di). 2000. *Contributo per una bibliografia sulla dialettologia percettiva*. Alessandria: Edizioni dell'Orso.

Colcuc, Beatrice. 2025. *Confini nel continuum? Percezione e rappresentazione della variazione linguistica in area dolomitica*. München/Salzburg: Ludwig-Maximilians-Universität; München/Paris: Lodron Universität Salzburg. (Tesi di dottorato.)

Colcuc, Beatrice & Cassisi, Nicola. 2024. "Cua l é sté na bela roba che é sucesso la not". Fenomeni di accomodamento linguistico in Agordino. *Rivista Italiana di Dialettologia* 47. 223–242.

Fiori, Stefano. 2023. Una prospettiva multivariata sulla produzione e percezione dei confini linguistici nelle Quattro Province. In Dal Negro, Silvia & Mereu, Daniela (a cura di), *Confini nelle lingue e tra le lingue. Atti del LV Congresso Internazionale della Società di Linguistica Italiana (SLI), Bressanone, 8-10 settembre 2022*, 31–46. Milano: Officinaventuno.

Iannàccaro, Gabriele & Dell'Aquila, Vittorio. 2001. Mapping languages from inside: notes on perceptual dialectology. *Social & Cultural Geography* 2(3). 265–280.

Iannàccaro, Gabriele. 2002. *Il dialetto percepito. Sulla reazione di parlanti di fronte al cambio linguistico*. Alessandria: Edizioni dell'Orso.

Iannàccaro, Gabriele. 2015. Tipi di percezione, linguistica della variazione e dialettologia. In Benedetto Mas, Paolo, D'Addario, Carlotta, Ghia, Alberto, Giordano, Silvia, Pons, Aline Sordella, Silvia & Trovato, Marianna (a cura di), *L'abisso saussureano e la costruzione delle varietà linguistiche*, 21–36. Alessandria: Edizioni dell'Orso.

Krefeld, Thomas & Pustka, Elissa. 2010. Per una varietistica percezionale. *Revue de Linguistique Romane* 74(295–296). 312–339.

Pike, Kenneth Lee. 1967. *Language in relation to a unified theory of the structure of human behavior*. The Hague/Paris: Mouton and Co.

ALESSANDRA FAGGIOTTO

Il dato fonologico in prospettiva multimodale: il *cued speech* come rappresentazione visiva del segmento

Sviluppato negli anni Sessanta dall'americano R. Orin Cornett nell'ambito dell'educazione di persone sorde, il *cued speech* diventa dispositivo privilegiato di accesso alla lingua orale. A partire dagli anni Settanta, si diffonde anche in Europa e viene riadattato a diverse lingue, tra cui quella francese (*Langue française Parlée Complétée*, LfPC; cfr. Leybaert 2012; Schultz 2020). Non è una lingua dei segni, né un sistema linguistico autonomo: si tratta un supporto bimodale che accompagna la produzione verbale attraverso un numero limitato di configurazioni manuali combinate con specifiche posizioni nello spazio peri-orale, con lo scopo di rendere distinguibili opposizioni fonologiche non accessibili alla sola lettura labiale.

Il principio di funzionamento del sistema si fonda su una mappatura sistematica tra gesto e unità segmentale: le configurazioni manuali distinguono classi consonantiche, mentre le posizioni articolatorie nello spazio segnalano le opposizioni vocaliche. La combinazione dei due parametri consente di rappresentare in modo discreto le sequenze consonante-vocale, rendendo esplicita la struttura segmentale del flusso parlato e disambiguando opposizioni altrimenti neutralizzate sul piano visivo (ad esempio consonanti labialmente omofone).

Il sistema del *cued speech* può allora essere considerato un caso particolarmente interessante per riflettere sulle modalità del dato fonologico. Nella pratica ordinaria, il dato fonologico è ricostruito a partire dal segnale acustico attraverso un processo di astrazione che isola unità discrete da un continuum fonetico. Nel *cued speech*, tale astrazione viene in parte esternalizzata: le opposizioni pertinenti sono rese immediatamente osservabili attraverso una codifica convenzionale ma sistematica, che seleziona solo le distinzioni funzionali del sistema linguistico. La funzione di disambiguazione, resa necessaria da condizioni di accesso non completo al segnale, porta in primo piano unità e opposizioni che nell'uso ordinario restano implicite e richiedono un'operazione di astrazione teorica.

L'organizzazione del sistema presuppone infatti una segmentazione discreta del flusso fonico e una corrispondenza stabile tra gesto e unità fonologica (cfr. Hyman 2011). Il *cued speech* e la *Langue française Parlée Complétée* non riproducono il dettaglio fonetico continuo, ma selezionano e codificano unità funzionali del sistema, rendendo esplicite le opposizioni pertinenti; possono dunque essere letta come una forma di rappresentazione visiva del segmento.

A partire da questa osservazione, l'intervento si concentra su una questione circoscritta: in che misura il dato fonologico possa essere considerato indipendente dal canale acustico che tradizionalmente ne costituisce il dominio privilegiato di osservazione. In questa prospettiva, il *cued speech* e la LfPC offrono un terreno di riflessione per interrogare il rapporto tra rappresentazione fonologica e modalità di manifestazione del dato linguistico.

Riferimenti bibliografici

Hyman, Larry M. 2011. The representation of segmental structure. In van Oostendorp, Marc, Ewen, Colin J., Hume, Elizabeth & Rice, Keren (a cura di), The Blackwell Companion to Phonology, 2–29. Malden (MA): Wiley-Blackwell.

Leybaert, Jacqueline (a cura di). 2012. *La Langue française Parlée Complétée (LPC): Fondements et perspectives*. Louvain-la-Neuve: De Boeck Supérieur.

Schultz, Jérôme (a cura di) 2020. *La LfPC regardée par les sciences.* Parigi: Association nationale pour la Langue française Parlée Complétée.

TOMMASO LAMARRA

L'astrazione lessicale attraverso metodi sperimentali: dall'elaborazione psicolinguistica di parole in isolamento al loro uso in conversazioni naturalistiche

Il presente contributo avanza una riflessione sul dato linguistico, esplorando l'astrazione lessicale come processo cognitivo attraverso due variabili semantiche e psicolinguistiche: concretezza e specificità categoriale.

La concretezza riguarda il grado di percettibilità fisica del referente denotato da un segno linguistico: le parole concrete denotano oggetti o entità esperibili attraverso i sensi, mentre le parole astratte rimandano a esperienze prevalentemente mediate dal linguaggio e dall'interazione sociale (Reilly et al. 2025). La specificità, invece, misura il livello di estensione, o inclusività categoriale, differenziando la semantica lessicale in termini di lessico sovraordinato, di base e subordinato (Bolognesi, Burges & Caselli 2020).

Tali variabili sono tradizionalmente indagate attraverso paradigmi psicolinguistici che impiegano stimoli linguistici decontestualizzati in configurazioni sperimentali controllate. Se da un lato questo approccio garantisce elevato rigore metodologico e consente di isolare gli effetti di specifiche variabili psicolinguistiche sull'uso della lingua, dall'altro tende a trascurare il ruolo cruciale dell'informazione contestuale nell'accesso ai significati lessicali e dell'interazione nella costruzione del significato (Banks et al. 2023; Vigliocco et al. 2024).

Si configura, perciò, un attrito metodologico tra il dato linguistico ricostruito e controllato in laboratorio e quello ecologicamente valido e contestualizzato, come quello ricavabile da corpora.

Questo contributo propone una riflessione teorico-metodologica su questa tensione, presentando un'indagine articolata in tre studi sperimentali che analizzano diverse tipologie di dati linguistici in tre ambienti di analisi: parole in isolamento (*Lexical e Semantic Decision Tasks* con misurazione di tempi di reazione), dialoghi simulati (valutazione cronometrica e indiretta della plausibilità percepita del dialogo) e corpora di conversazioni semi-naturalistiche (analisi quantitativa di dinamiche semantiche e pragmatiche).

Studio 1. I dati cronometrici supportano il noto *concreteness effect*, un vantaggio comportamentale nell'elaborazione delle parole concrete rispetto alle astratte, ed evidenziano un vantaggio cognitivo per le parole specifiche rispetto a quelle generali, indipendentemente dal grado di concretezza.

Studio 2. Gli stimoli sono stati inseriti in dialoghi naturalistici simulati, in grado di fornire un supporto contestuale esplicito. I partecipanti hanno valutato la plausibilità di 60 frasi italiane contenenti parole che variavano per concretezza e specificità, associate a risposte che veicolavano diverse funzioni comunicative (espressione di incertezza, certezza, richiesta di chiarimento, assertività). I risultati mostrano che le frasi astratte sono giudicate più plausibili quando accompagnate da risposte che segnalano incertezza o bisogno di chiarimento, indipendentemente dal livello di specificità. Al contrario, le frasi concrete-generali risultano più plausibili nell'elicitare richieste di chiarimento rispetto alle frasi concrete-specifiche, che si sono rivelate più pragmaticamente esaustive.

Studio 3. È stata condotta un'analisi su 6.483 frasi inglesi estrapolate dal corpus ECOLANG (Gu et al., 2025), un *dataset* di conversazioni semi-naturalistiche in cui gli interlocutori descrivono oggetti in quattro condizioni sperimentali che riflettono situazioni reali: oggetti fisicamente presenti vs. assenti; familiari vs. non familiari. I risultati mostrano un aumento della concretezza nel linguaggio impiegato per descrivere oggetti non familiari e una maggiore specificità quando gli oggetti sono al contempo non familiari e fisicamente assenti.

Nel complesso, la ricerca contribuisce al dialogo tra dati e teorie dell'elaborazione linguistica combinando evidenze sperimentali e *corpus-based* e proponendo un modello di indagine che mette in continuità paradigmi sperimentali e dati linguistici lungo una progressione di validità ecologica.

Riferimenti bibliografici

Banks, Briony, Borghi, Anna Maria, Fargier, Raphaël, Fini, Chiara, Jonauskaite, Domicele, Mazzuca, Claudia, Montalti, Martina, Villani, Caterina & Woodin, Greg. 2023. Consensus paper: Current perspectives on abstract concepts and future research directions. *Journal of Cognition* 6(1). 62. https://doi.org/10.5334/joc.238

Bolognesi, Marianna, Burgers, Christian & Caselli, Tommaso. 2020. On abstraction: decoupling conceptual concreteness and categorical specificity. *Cognitive Processing* 21(3). 365–381. https://doi.org/10.1007/s10339-020-00965-9

Gu, Yan, Donnellan, Ed, Grzyb, Beata, Brekelmans, Gwen, Murgiano, Margherita, Brieke, Ricarda, Perniss, Pamela & Vigliocco, Gabriella. 2025. The ECOLANG Multimodal Corpus of adult-child and adult-adult Language. *Scientific Data* 12(1). 89. https://doi.org/10.1038/s41597-025-04405-1

Reilly, Jaimie, Shain, Cory, Borghesani, Valentina, Kuhnke, Philipp, Vigliocco, Gabriella, Peelle, Jonathan E., Mahon, Bradford Z., Buxbaum, Laurel J., Majid, Asifa, Brysbaert, Marc, Borghi, Anna M., De Deyne, Simon, Dove, Guy, Papeo, Liuba, Pexman, Penny M., Poeppel, David, Lupyan, Gary, Boggio, Paulo, Hickok, Gregory, Gwilliams, Laura, Fernandino, Leonardo, Mirman, Daniel, Chrysikou, Evangelia G., Sandberg, Chaleece W., Crutch, Sebastian J., Pylkkänen, Liina, Yee, Eiling, Jackson, Rebecca L., Rodd, Jennifer M., Bedny, Marina, Connell, Louise, Kiefer, Markus, Kemmerer, David, de Zubicaray, Greig, Jefferies, Elizabeth, Lynott, Dermot, Siew, Cynthia S.Q., Desai, Rutvik H., McRae, Ken, Diaz, Michele T., Bolognesi, Marianna, Fedorenko, Evelina, Kiran, Swathi, Montefinese, Maria, Binder, Jeffrey R., Yap, Melvin J., Hartwigsen, Gesa, Cantlon, Jessica, Bi, Yanchao, Hoffman, Paul, Garcea, Frank E. & Vinson, David. 2025. What we mean when we say semantic: Toward a multidisciplinary semantic glossary. *Psychonomic Bulletin & Review* 32. 243–280. https://doi.org/10.3758/s13423-024-02556-7

Vigliocco, Gabriella, Convertino, Laura, De Felice, Sara, Gregorians, Lara, Kewenig, Viktor, Mueller, Marie A. E., Veselic, Sebastijan, Musolesi, Mirco, Hudson-Smith, Andrew, Tyler, Nicholas, Flouri, Eirini & Spiers, Hugo J. 2024. Ecological brain: reframing the study of human behaviour and cognition. *Royal Society open science* 11(11). 240762. https://doi.org/10.1098/rsos.240762

ALESSANDRO LENCI

Quali dati, quali teorie? Sfide e prospettive per la linguistica nell'era dei Large Language Model

I *Large Language Model* (LLM) apprendono le strutture del linguaggio a partire da dati quantitativamente e qualitativamente molto diversi da quelli a disposizione degli esseri umani. Questa differenza solleva interrogativi cruciali sul rapporto tra apprendimento statistico, esperienza linguistica e conoscenza grammaticale. In tale scenario, la teoria linguistica può e deve svolgere un ruolo centrale, andando oltre il "gigantismo" computazionale dei modelli attuali per interrogarsi sui dati rilevanti per l'acquisizione e sui meccanismi dell'apprendimento umano. La linguistica teorica è inoltre indispensabile per interpretare le generalizzazioni che i nuovi modelli neurali sembrano acquisire, distinguendo tra mere correlazioni distribuzionali e rappresentazioni linguistiche astratte. In questa prospettiva, gli LLM possono essere considerati non soltanto strumenti tecnologici, ma veri e propri paradigmi epistemologici, capaci di ridefinire il dialogo tra linguistica teorica, computazionale e cognitiva. Lungi dal segnare la fine della teoria linguistica, l'attuale sviluppo dell'intelligenza artificiale generativa pone la linguistica stessa di fronte a nuove sfide sul rapporto tra dati e teoria, tra uso e competenza, tra *performance* e spiegazione.

LUCA LÒTANO

GoPro e creazione partecipata del dato: dal punto di vista incarnato alla video-analisi condivisa

Il contributo proposto (nella forma "comunicazione orale") indaga l'uso di videocamere indossabili GoPro come dispositivo di creazione del dato in ricerche induttive (bottom-up) sulla mediazione (Piccardo 2022; Piccardo & Langé 2023) in contesti educativi plurilingui (Canagarajah 2013; Ortega 2019) e multimodali con persone migranti e non (D'Agostino 2021, 2022, 2023). Partendo dall'assunto che il dato non coincide con la mera registrazione ma emerge da un processo di scelte che ne determinano forma e statuto (Kalocsányiová & Shatnawi 2022) si indaga il completo iter di creazione del dato: dalla configurazione tecnica (videocamera indossata/handheld/posizionata) ai successivi criteri di selezione degli episodi, alle pratiche di analisi e trascrizione condivisa con i partecipanti. In questa prospettiva l'utilizzo della GoPro produce un tipo specifico di evidenza: un dato incarnato e situato, in grado di avvicinarsi al fuoco attentivo dei partecipanti e di rendere osservabili micro-momenti di negoziazione quali riparazioni, chiarimenti, riformulazioni, gesti deittici, uso di oggetti e dispositivi digitali. Kinsley, Schoonover & Spitler (2016) mostrano come il dispositivo GoPro possa sostenere una postura etnografica sensibile al *wayfinding* e al punto di vista del soggetto; van der Kleij, Adie e Cumming (2019) evidenziano l'adeguatezza della GoPro per studiare interazioni educative e processi di feedback; Porter & Wilson (2019) sottolineano il valore formativo delle tecnologie indossabili nel rendere visibile l'esperienza e attivare riflessione sull'azione: la pratica di osservazione (auto)riflessiva diventa poi tale anche per gli insegnanti/ricercatori nel riosservare se stessi "osservati" dagli apprendenti.

Il contributo propone quindi una proposta operativa innovativa in quattro passaggi: (1) progettazione delle condizioni di ripresa (introduzione dello strumento nel contesto classe; camera indossata/handheld/posizionata; tempi); (2) raccolta di video anche in assenza del docente/ricercatore, per ridurre alcune forme di "orientamento" e intercettare interazioni tra pari così da riprodurre e valorizzare pratiche fluide e plurilingue (Lüpke & Cissé 2023); (3) selezione bottom-up di episodi attraverso indizi interazionali (*breakdown* di comprensione, richieste di aiuto, riallineamenti); (4) analisi e trascrizione condivisa con i partecipanti, multimodale (parlato, alternanza di codici, prosodia, gesto, sguardo, spazio, oggetti). La camera "incarnata" e la revisione collettiva come seconda fase di produzione del dato (Vigouroux 2007) permettono così di ascoltare anche l' "inascoltato": questa proprietà del dato orienta ciò che poi diventa "evidenza". Quando i partecipanti rivedono insieme le riprese, il materiale registrato dal gruppo stesso si trasforma: emergono negoziazioni, accordi/disaccordi, spiegazioni retrospettive, attribuzioni di intenzionalità, etichette linguistiche e criteri di "comprensibilità" che non sono direttamente leggibili nel flusso dell'azione.

In altre parole, la co-visione genera un dato metapragmatico che documenta come il gruppo costruisce significato e valuta l'efficacia delle proprie strategie. Questa fase rende esplicite asimmetrie e responsabilità (chi può nominare cosa, chi autorizza un'interpretazione), ma anche produce un'ulteriore evidenza: la storia sociale del dato, cioè come viene selezionato, spiegato e reso condivisibile. Il contributo propone quindi di trattare la co-analisi non come semplice "validazione" a posteriori, ma come momento partecipato di costruzione dell'evidenza, capace di aumentare trasparenza metodologica, densità interpretativa e qualità etica del dato audiovisivo.

Dati: estratti video GoPro e videocamere (2024–2025) da contesti di insegnamento/apprendimento di italiano L2 con persone migranti, in setting formali e non formali. Il corpus include riprese in autonomia e non e sessioni di co-visione per la produzione di dato metapragmatico.

Riferimenti bibliografici

Canagarajah, Suresh. 2013. Theorizing a competence for translingual practice at the contact zone. In May, Stephen (a cura di), *The Multilingual Turn: Implications for SLA, TESOL and Bilingual Education*. New York: Routledge. https://doi.org/10.4324/9780203113493

D'Agostino, Mari. 2021. *Noi che siamo passati dalla Libia. Giovani in viaggio fra alfabeti e multilinguismo*. Il Mulino. Bologna.

D'Agostino, Mari. 2022. Giovani in movimento: multilingui, connessi, spesso analfabeti. Una nuova migrazione fra risorse e bisogni. *Italiano LinguaDue* 14(1). 5–13. https://doi.org/10.54103/2037-3597/18148

Kalocsányiová, Erika & Shatnawi, Malika. 2022. 10. Transcribing (multilingual) voices: From fieldwork to publication. In Holmes, Prue, Reynolds, Judith & Ganassin, Sara (a cura di), *The politics of researching multilingually,* 209–228. Bristol/Blue Ridge Summit: Multilingual Matters, Channel View Publications Ltd. https://doi.org/10.21832/9781800410152-014

Kinsley, Kirsten, M., Schoonover, D., & Spitler, J. 2016. *GoPro as an ethnographic tool: A wayfinding study in an academic library*. *Journal of Access Services* 13(1). 7–23. https://doi.org/10.1080/15367967.2016.1154465

Lüpke, Friederike & Cissé, Ibrahima Abdoul Hayou. 2023. *Legitimising fluid multilingual practices: A challenge for formal education worldwide.* In Reilly, Colin, Chimbutane, Feliciano, Clegg, John, Rubagumya, Casmir & Erling, Elizabeth J. (a cura di), *Multilingual Learning:*

Assessment, Ideologies and Policies in Sub-Saharan Africa, 43–64. Abingdon/New York: Routledge. https://doi.org/10.4324/9781003311553-3

Ortega, Lourdes. 2019. SLA and the Study of Equitable Multilingualism. *The Modern Language Journal* 103. 23–38. https://doi.org/10.1111/modl.12525

Piccardo Enrica. 2022. Mediation and the plurilingual /Pluricultural dimensionin language education. *ItalianoLinguaDue* 14(2). 24–45. https://doi.org/10.54103/2037-3597/19568

Piccardo, Enrica &, Langé, Gisella (a cura di). 2023. *La classe plurilingue. Insegnare con un approccio orientate all'azione*. Milano/Torino: Sanoma

Porter, Su & Wilson, Hamish. 2019. The pedagogy of technology in outdoor learning or use of the GoPro to enhance learning and teaching. In Kumer, Branko (a cura di), *The Mediation of Experiences by Technology in the Outdoors: Opening or Losing Connections with the World*. *17th EOE Conference's Selection of Papers*, *Bohinj, 19–23 settembre 2018*, 68–75. Ljubljana: CŠOD.

van der Kleij, Fabienne, Adie, Lenore & Cumming, Joy. 2019. Feasibility and Value of Using a GoPro Camera and iPad to Study Teacher-Student Assessment Feedback Interactions. In: Veldkamp, Bernard P. & Sluijter, Cor (a cura di), *Theoretical and Practical Advances in Computer-based Educational Measurement. Methodology of Educational Measurement and Assessment.* Cham: Springer. https://doi.org/10.1007/978-3-030-18480-3_18

Vigouroux, Cecile B. 2007. Trans-scription as a social activity: An ethnographic approach. *Ethnography* 8(1). 61–97. https://doi.org/10.1177/1466138107076137

KRISTEN LEONE

Elicitare il non standard: due studi di caso all'interfaccia tra semantica e sintassi

I dati elicitati costituiscono una risorsa importante per la ricerca linguistica, in particolare per lo studio di varietà non standard, poiché consentono di osservare la distribuzione di forme e costruzioni scarsamente attestate in *corpora* o atlanti linguistici. Pur presentando limitazioni metodologiche, legate alla possibile influenza del ricercatore sull'informante o alla produzione di forme metalinguisticamente controllate (Filipponio 2024; Schilling 2013), le tecniche di elicitazione si rivelano utili nell'indagare le strategie di lessicalizzazione all'interfaccia tra semantica e sintassi.

In questo contributo si presentano due studi di caso relativi al panorama italo-romanzo, nei quali i dati raccolti attraverso task sperimentali hanno consentito di osservare l'accettabilità e la produttività di alcune costruzioni in relazione a fattori semantici, sociolinguistici e tipologici.

Il primo studio riguarda le costruzioni di moto causato (Goldberg 1995) con verbi di movimento direzionale impiegati transitivamente (es. *scendere il cane*, *uscire la moto*) nell'italiano regionale della Basilicata sud-orientale. Queste costruzioni sono presenti in molte varietà dialettali e di italiano regionale dell'Italia meridionale e mostrano una distribuzione sensibile sia a fattori sociolinguistici sia alle proprietà semantiche dell'oggetto diretto (Busso & Romagno 2021; Romagno 2021). Mediante la somministrazione di un questionario volto alla raccolta di giudizi di accettabilità, è emerso un dato non evidenziato dagli studi precedenti: la produttività di queste costruzioni sembrerebbe essere influenzata anche dalla disponibilità, nel dialetto corrispondente, di verbi funzionalmente equivalenti e lessicalmente causativi che "bloccano" l'uso del corrispondente verbo intransitivo – cfr. il concetto di *statistical preemption* (Boyd & Goldberg 2011; Perek & Goldberg 2017). Questo meccanismo è osservabile nelle costruzioni contenenti *uscire*, ritenute sistematicamente meno accettabili da tutti i gruppi di parlanti. Alla richiesta di motivare il giudizio di *non accettabilità*, molti parlanti hanno infatti proposto alternative con il verbo *cacciare* 'tirare/portare fuori', impiegato anche in dialetto (1):

(1) [u ˈdd͡ziə a kkatˈt͡ʃat a ˈmoːtə ˈdu gaˈraʃə]
Lo zio ha cacciato la moto dal garage

'Lo zio ha tirato/portato fuori la moto dal garage'

Nel secondo studio si presenta invece un'indagine pilota sulla realizzazione degli stati psicologici in alcune varietà dialettali, con particolare attenzione al confronto tra la produzione di verbi sintetici e di costruzioni a verbo supporto – in cui la predicazione è "sbilanciata" sul nome, mentre il verbo fornisce i tratti di flessione grammaticale e di accordo con il soggetto (Gross 1981; Ježek 2004, 2011). Alcuni verbi possono inoltre svolgere una funzione diatetica, consentendo di esprimere diversi significati aspettuali (Pompei & Piunno 2023), come in *avere paura*, *fare paura* e *prendere paura*.

Per questo studio, gli informanti hanno tradotto nelle rispettive varietà dialettali gli stimoli proposti in un questionario contenente costruzioni [V + N] con nove nomi di stato psicologico. Di seguito, si riporta un esempio di stimolo tradotto dagli informanti dei punti di inchiesta di Montecchia di Crosara (2a) e Gravina in Puglia (2b):

(1) a. [mi ɡ ˈɔ paˈura ˈdei̯ ˈraɲi]
io ho paura dei ragni

'ho paura dei ragni'

b. [m appaˈvurəkə də lə ˈraɲɲə]

mi impaurisco dei ragni

'ho paura dei ragni'

I dati preliminari suggeriscono che la preferenza per forme sintetiche o analitiche dipenda da diversi fattori, tra cui le proprietà semantiche degli stati psicologici, i lessemi impiegati e l'area dialettale di appartenenza.

Complessivamente, i due studi di caso evidenziano come i dati elicitati possano contribuire all'analisi non solo dei singoli fenomeni, ma anche delle dinamiche di interazione tra dialetti, varietà regionali e lingua standard.

Riferimenti bibliografici

Boyd, Jeremy K. & Goldberg, Adele E. 2011. Learning What NOT to Say: The Role of Statistical Preemption and Categorization in A-Adjective Production. *Language* 87(1). 55–83. https://doi.org/10.1353/lan.2011.0012

Busso, Lucia & Romagno, Domenica. 2021. Caused Motion Constructions between standard and substandard: *entrare*, *uscire*, *scendere* and *salire* in contemporary Italian. *Italian Journal of Linguistics* 33(2). 109–146. https://doi.org/10.26346/1120-2726-177

Filipponio, Lorenzo. 2024. 'Restsprecher' and Hypercharacterizing Informants between Veglia and Capraia. In Baglioni, Daniele & Rigobianco, Luca (a cura di), *Fragments of Languages. From 'Restsprachen' to Contemporary Endangered Languages*, 213–232. Leiden/Boston: Brill.

Goldberg, Adele E. 1995. *Constructions: A Construction Grammar Approach to Argument Structure.* Chicago: The University of Chicago Press.

Gross, Maurice. 1981. Les bases empiriques de la notion de prédicat sémantique. *Langages* 63. 7–52.

Ježek, Elisabetta. 2004. Types et Degrés de Verbes Supports En Italien. *Lingvisticae Investigationes* 27(2). 185–201.

Ježek, Elisabetta. 2011. Verbi Supporto. In In Simone, Raffaele (a cura di), *Treccani - Enciclopedia dell'italiano*. Roma: Istituto della Enciclopedia Italiana Treccani. https://www.treccani.it/enciclopedia/verbi-supporto_(Enciclopedia-dell'Italiano)/

Perek, Florent & Goldberg, Adele E. 2017. Linguistic generalization on the basis of function and constraints on the basis of statistical preemption. *Cognition* 168. 276–293. https://doi.org/10.1016/j.cognition.2017.06.019

Pompei, Anna & Piunno, Valentina. 2023. Light Verb Constructions in Romance languages. An attempt to explain systematic irregularity. In Pompei, Anna, Mereu, Lunella & Piunno, Valentina (a cura di), *Light Verb Constructions as Complex Verbs,* 99–147. Berlin: De Gruyter Mouton.

Romagno, Domenica. 2021. Caused motion constructions in northern Calabria: object animacy and the role of dialect. *Lingue e linguaggio* 2(2021). 289–310. https://doi.org/10.1418/102816

Schilling, Natalie. 2013. *Sociolinguistic Fieldwork*. Cambridge: Cambridge University Press.

MARTA MAFFIA E MASSIMO PETTORINO

Quando i dati mancano: il parlato dal web nella ricerca sui *biomarker* ritmici del Parkinson

Le alterazioni del parlato associate alla Malattia di Parkinson (MP) sono ampiamente documentate a livello fonetico-fonologico, prosodico e ritmico e si inscrivono nel quadro clinico della disartria ipocinetica (Darley, Aronson & Brown 1969; Duffy 2013). Numerosi studi sperimentali, anche di natura computazionale, hanno tentato di individuare *biomarker* linguistici e acustici utili a distinguere il parlato disartrico da quello sano (tra gli altri, Harel et al. 2004; Ferrera & Gagliardi 2023).

Tuttavia, tali ricerche si fondano prevalentemente su dati raccolti in contesti clinici controllati, con compiti altamente strutturati (*sustained phonation*, compiti diadococinetici, ripetizione di sillabe), e sempre in fasi già manifeste della patologia. Meno esplorata risulta una prospettiva longitudinale orientata a intercettare segnali preclinici o prodromici e a definire *quando*, rispetto all'insorgenza dei sintomi motori, tali alterazioni linguistiche emergano. Una simile prospettiva è cruciale, al fine di individuare informazioni utili a supportare procedure di diagnosi precoce (Rusz et al. 2016).

Il presente contributo si colloca, quindi, in una riflessione più ampia sulla scarsità e accessibilità dei dati clinici longitudinali e spontanei, sostenendo che, quando i dati mancano, o sono difficilmente reperibili per vincoli etici, logistici o temporali, il parlato disponibile online costituisca una risorsa alternativa, purché sottoposta a rigorosi criteri di selezione e trattamento. Tale prospettiva è già stata usata in diversi studi sulla Malattia di Alzheimer (Petti et al. 2023; Gao et al. 2025) e anche sul parlato parkinsoniano (Favaro et al. 2024). Precedenti studi longitudinali specifici sul ritmo del parlato in soggetti con MP sono stati condotti su dati italiani (Pettorino, Maffia & Bigi 2022) e inglesi reperiti dal web (Pettorino et al. 2018, Pettorino & Maffia 2024).

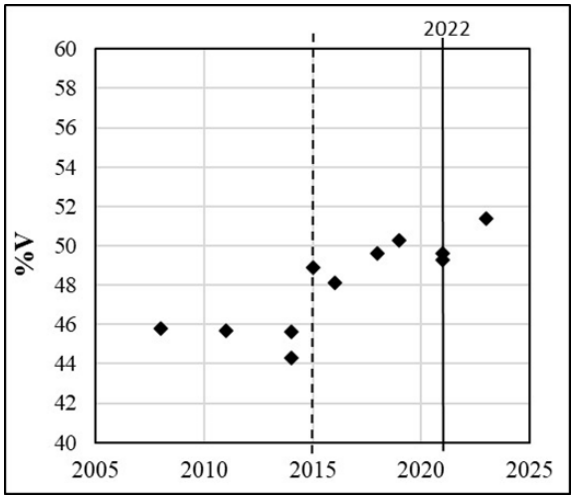
In continuità con tali ricerche, si propone in questo studio l'utilizzo di *ParkCeleb* (Favaro et al. 2024), una raccolta di campioni di parlato di personaggi noti affetti da MP, registrati in contesti pubblici negli anni precedenti e successivi alla diagnosi. Tale corpus, pur non nato con finalità cliniche, consente di ricostruire traiettorie evolutive del parlato, offrendo dati ecologici e spontanei ma anche molto eterogenei.

Dall'intero corpus sono stati selezionati, sulla base della qualità acustica e di altre peculiarità dei dati audio, 24 campioni di parlato di due soggetti (S1 e S2), entrambi anglofoni. I campioni selezionati sono stati oggetto di un'analisi spettroacustica e di una annotazione manuale in intervalli

vocalici (V) e consonantici (C), condotta con Praat (Boersma & Weenink 2023), secondo gli standard di etichettatura proposti in Maffia et al. (2021). In totale sono stati annotati circa 9200 intervalli V/C.

L'annotazione ha permesso il calcolo di alcune tra le metriche ritmiche più utilizzate in letteratura: %V, ΔC, VarcoV, VarcoC, nPVI-V, rPVI-C, VtoV. T-test per campioni appaiati sono stati applicati per valutare la significatività statistica del confronto tra i dati ritmici pre- e post-diagnosi, con il software R (vers. 4.3.1).

I risultati, in linea con quelli dei precedenti casi di studio, mostrano un'alterazione statisticamente significativa nei valori relativi alla %V, con un aumento evidente già diversi anni prima della diagnosi clinica, come riportato nelle figure 1 e 2.



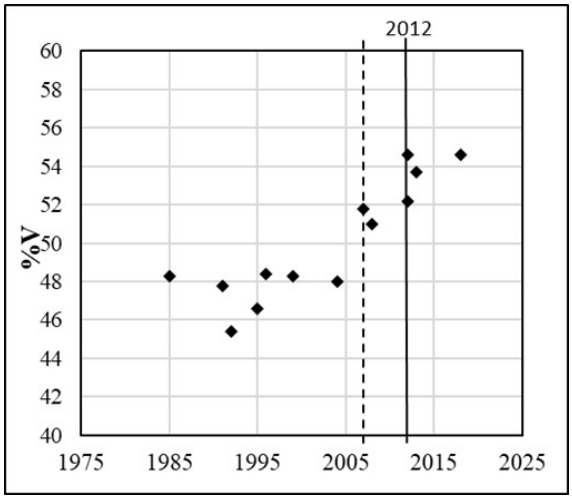


Figure 1 e 2: Andamento della %V nei dati di S1 (fig. 1) e S2 (fig. 2).
La linea continua rappresenta l'anno della diagnosi, quella tratteggiata l'anno di comparsa dell'alterazione.

I dati confermano, quindi, che tale parametro può essere considerato una spia precoce di cambiamenti neuromotori sottostanti e un possibile *biomarker* ritmico, potenzialmente utile in procedure di *screening* precoce della MP. Questo caso di studio, inoltre, conferma l'osservabilità di tali variazioni di natura ritmica anche in dati raccolti online, spontanei e non omogenei, e la possibilità che il web offre di reperire campioni linguistici antecedenti alla formalizzazione diagnostica e di condurre ricerche longitudinali nell'ambito della linguistica clinica, non basate necessariamente su enormi quantità di dati.

Riferimenti bibliografici

Boersma, Paul & Weenink, David. 2023. *Praat: Doing phonetics by computer* (version 6.3.08). www.fon.hum.uva.nl/praat/

Darley, Frederic. L., Aronson, Arnold E. & Brown, Joe R. 1969. Cluster of deviant speech dimension in the dysarthrias. *Journal of Speech and Hearing Research* 12(3). 462–496. https://doi.org/10.1044/jshr.1203.462

Duffy, Joseph R. 2013. *Motor speech disorders: Substrates, differential diagnosis, and management*. St. Louis (MO): Elsevier.

Favaro, Anna, Butala, Ankur, Thebaud, Thomas, Villalba, Jesús, Dehak, Najim & Moro-Velázquez Laureano. 2024. Unveiling early signs of Parkinson's disease via a longitudinal analysis of celebrity speech recordings. *npj Parkinson's Disease* 10. 207. https://doi.org/10.1038/s41531-024-00817-9

Ferrera, Alessandra & Gagliardi, Gloria. 2023. Acoustic biomarkers of Parkinson's Disease: a pilot study on the Italian Parkinson's Voice and Speech dataset. In Gili Fivela, Barbara, D'Apolito, Sonia, Pagliaro, Anna Chiara, Sallustio, Vincenzo, Fiorella, Maria Luisa & Siniscalchi, Marco (a cura di), *Il parlato in ambito medico: analisi linguistica, applicazioni tecnologiche e strumenti clinici [Spoken language in the medical field: Linguistic analysis, technological applications and clinical tools] (Studi AISV 11)*, 63–82. Milano: Officinaventuno. https://doi.org/10.17469/O2111AISV000005

Gao, Kunxiao, Favaro, Anna, Dehak, Najim & Moro-Velázquez Laureano. 2025. ADCeleb: A Longitudinal Speech Dataset from Public Figures for Early Detection of Alzheimer's Disease. In Klabbers, Esther, Cooke, Martin & Kinnunen, Tomi H. (a cura di), *Proceedings of INTERSPEECH 2025*, *Rotterdam*, *17-21 agosto 2025*, 5688–5692. Baixas: ISCA - International Speech Communication Association https://doi.org/10.21437/Interspeech.2025-523

Harel, Brian T., Cannizzaro, Michael S., Cohen, Henri, Reilly, Nicole & Snyder, Peter J. 2004. Acoustic characteristics of Parkinsonian speech: a potential biomarker of early disease progression and treatment. *Journal of Neurolinguistics*. 17(6). 439–453. https://doi.org/10.1016/j.jneuroling.2004.06.001

Maffia, Marta, De Micco, Rosa, Pettorino, Massimo, Siciliano, Mattia, Tessitore, Alessandro & De Meo, Anna. 2021. Speech rhythm variation in early-stage Parkinson's disease: a study on

different speaking tasks. *Frontiers in Psychology* 12. 668291. https://doi.org/10.3389/fpsyg.2021.668291

Petti, Ulla, Baker, Simon, Korhonen, Anna & Robin, Jessica. 2023. The generalizability of longitudinal changes in speech before Alzheimer's disease diagnosis. *Journal of Alzheimer's Disease* 92(2). 547–564. https://doi.org/10.3233/JAD-220847

Pettorino, Massimo & Maffia, Marta. 2024. Analyzing changes in Parkinsonian speech over time: a diachronic experimental phonetics study. *Frontiers in Aging Neuroscience* 16. 1334198. https://doi.org/10.3389/fnagi.2024.1334198

Pettorino, Massimo, Maffia, Marta & Bigi, Brigitte. 2022. A longitudinal study on Italian speech rhythm in Parkinson's disease. In Frota, Sónia, Cruz, Marisa & Vigário, Marina (a cura di), *Proceedings of the 11th International Conference on Speech Prosody*, *Lisbon, 23-26 maggio 2022*, 52–56. https://doi.org/10.21437/speechprosody.2022-11

Pettorino, Massimo, Hemmerling, Daria, Vitale, Marilisa & De Meo, Anna. 2018. Towards a speech-test for Parkinson's disease detection: a diachronic study on Michael J. Fox. In Herencsár, Norbert (a cura di), *Proceedings of the 41st International Conference on Telecommunications and Signal Processing (TSP), Athens, 4-6 luglio 2018*, 1–5. Piscataway (NJ): IEEE. https://doi.org/10.1109/TSP.2018.8441268

Rusz, Jan, Hlavnička, Jan, Tykalová, Tereza, Bušková, Jitka, Ulmanová, Olga, Růžička, Evžen & Šonka, Karel. 2016. Quantitative assessment of motor speech abnormalities in idiopathic rapid eye movement sleep behaviour disorder. *Sleep medicine*. 19. 141–147. https://doi.org/10.1016/j.sleep.2015.07.030

BORBALA SAMU ED ELISABETTA SANTORO

Il *Discourse Completion Task* come strumento di raccolta dati per la pragmatica (inter)linguistica e cross-culturale

Nel dibattito contemporaneo sui dati linguistici, la pragmatica occupa una posizione metodologicamente sensibile: l'oggetto di studio (azioni linguistiche situate, norme di appropriatezza, gestione della relazione) è intrinsecamente contestuale, mentre la ricerca empirica richiede comparabilità, replicabilità e controllo delle variabili. Questo contributo propone una riflessione critica sul "dato" in pragmatica sperimentale attraverso il caso dei *Discourse Completion Tasks* (DCT), intesi non come mero strumento di raccolta, ma come dispositivo di costruzione del dato: selezione delle situazioni, modellizzazione del contesto, elicitazione, annotazione, interpretazione.

Nella tradizione degli studi cross-culturali (Blum-Kulka, House & Kasper 1989), il DCT consente di generare produzioni di atti linguistici in condizioni sistematicamente variate e quindi comparabili tra gruppi di parlanti diversi per provenienza, ma anche per età, genere, gruppo di appartenenza ecc. Tale "vantaggio gestionale" diventa epistemico: rende osservabili regolarità di strategie sul piano pragmalinguistico e in relazione a variabili, tra cui la distanza sociale e il grado di imposizione. Tuttavia, la natura elicitata e spesso scritta del DCT produce un dato ibrido: non coincide con il parlato naturale e tende a privilegiare risposte più stereotipate, meno sensibili alla co-costruzione sequenziale, alla prosodia e alle risorse multimodali. La letteratura mostra differenze non univoche tra DCT, role play e dati naturali (Beebe & Cummings 1996; Kasper & Dahl 1991; Ogiermann 2018). Attraverso due casi di raccolta dati in pragmatica sperimentale, verranno presentate alcune problematiche concrete legate alla definizione, alla comparabilità e all'interpretazione del dato linguistico.

Un primo caso riguarda le risposte al complimento in italiano L1 e la variazione sociolinguistica. Un corpus di 1.377 risposte elicitato tramite DCT scritto, raccolto in tre aree geografiche (Piemonte, Umbria, Campania), mostra strategie ad alta frequenza comuni ma anche differenze sistematiche nella selezione di specifici pragmemi (maggiore minimizzazione in Umbria, più richieste di rassicurazione in Piemonte, maggiore accettazione ironica in Campania), ulteriormente sensibili alle variabili situazionali e al topic. La variazione emerge così come componente strutturale del dato, non come rumore.

Un secondo caso riguarda un DCT scritto sulle richieste somministrato in 14 lingue nell'ambito di una ricerca plurilingue. Qui il DCT ha consentito di controllare in modo sistematico variabili situazionali (distanza sociale e grado di imposizione) e di costruire un corpus comparabile su larga

scala. Proprio la dimensione contrastiva rende evidente il carattere costruito del dato: la standardizzazione delle situazioni garantisce comparabilità, ma impone una modellizzazione del contesto. Per questo, il DCT è stato affiancato ad altri strumenti (questionari di percezione, protocolli di riflessione, *role play*), in una logica di triangolazione metodologica (Zhang & Zhang 2020) che consente di distinguere tra repertorio di possibilità riconosciute come appropriate e realizzazioni interazionali situate.

Il contributo argomenta che i "limiti" del DCT non siano da intendere come difetti, ma come proprietà costitutive del dato elicitato. Il DCT è particolarmente informativo per ricostruire repertori e distribuzioni probabilistiche e per costruire risorse riutilizzabili in prospettiva comparativa, anche se strutturalmente meno adatto a rendere conto della sequenzialità, dell'interazione e delle risorse prosodiche e multimodali. Assumere questa differenza come dato epistemologico consente di integrare il DCT in una più ampia "scienza dei dati" pragmatica, riconoscendone il valore per una mappatura di pattern e distribuzioni, senza sovrapporre ciò che il dato elicitato mostra a ciò che richiede dati di altro tipo.

Riferimenti bibliografici

Beebe, Leslie M. & Cummings, Martha C. 1996. Natural speech data versus written questionnaire data: How data collection method affects speech act performance. In Gass, Susan M. & Neu, Joyce (a cura di), *Speech Acts Across Cultures: Challenges to Communication in a Second Language*, 65–86. Berlin/New York: Mouton de Gruyter. https://doi.org/10.1515/9783110219289.1.65

Blum-Kulka, Shoshana, House, Juliane & Kasper, Gabriele (a cura di). 1989. *Cross-cultural Pragmatics: Requests and Apologies*. Norwood: Ablex.

Kasper, Gabriele & Dahl, Merete. 1991. Research methods in interlanguage pragmatics. *Studies in Second Language Acquisition* 13(2). 215–247. https://doi.org/10.1017/S0272263100009955

Ogiermann, Eva. 2018. Discourse completion tasks. In Jucker, Andreas H., Schneider, Klaus P. & Bublitz, Wolfram (a cura di), *Methods in Pragmatics*, 229–255. Berlin/Boston: De Gruyter Mouton.

Zhang, Lawrence J. & Zhang, Donglan. 2020. Think-aloud protocols. In McKinley, Jim & Rose, Heath (a cura di), *The Routledge Handbook of Research Methods in Applied* Linguistics, 302–311. London/New York: Routledge.

SILVIA SILLERESI, ITAI BASSI, ABIGAIL ANNE BIMPEH, IMKE DRIEMEL, JOHNSON FỌLỌRUNṢỌ ILỌRI, ANASTASIA NUWORSU, GERALD OKEY NWEYA E MARIA TERESA GUASTI

Rethinking acceptability judgments in linguistic research: the case of logophoric pronouns in three West African languages

Acceptability judgments are reports of a speaker's subjective sense of well-formedness of linguistic data and have long been central to theoretical linguistics, often treated as direct evidence for grammatical competence (Chomsky 1965). Traditionally elicited by theoreticians from themselves or a small group of colleagues, such judgments have been criticized as informal and potentially biased, especially since corpus and experimental data frequently reveal distributional patterns that diverge from individual intuitions and long-standing theoretical assumptions (Myers 2017; Schütze 2005). As responses to linguistic stimuli, acceptability judgments are themselves experimental data and are sensitive to task design, contextual framing, speaker expectations, processing factors, and sociolinguistic background (Schütze 2015; Sprouse 2013). Rather than viewing such discrepancies as problematic, this paper argues that they invite a reconsideration of how acceptability data are collected, interpreted, and integrated with other types of linguistic data.

We illustrate this argument through a case study of co-reference possibilities of logophoric pronouns in three West African languages: Ewe, Igbo, and Yoruba. Logophoric pronouns (LogP) occur in the scope of an attitude predicate and unambiguously refer to the attitude holder, disallowing disjoint reference (Adésola 2005; Bimpeh et al. 2024; Clements 1975; Manfredi 1987; Pearson 2015). These languages also have ordinary pronouns (OrdP) that can appear in the same environments. A central debate concerns whether OrdP must refer to a contextually salient individual distinct from the attitude holder (Bimpeh et al. 2024; Clements 1975; Hyman and Comrie 1981; Manfredi 1987) or whether co-reference with the attitude holder is possible (Adésola 2005; Pearson 2015). An example is reported in (1) for Ewe.

(1)	*Kòfi*$_1$	*gblɔ*	*bé*	***yè***$_{1/*2}$ ***/ é***$_{\%1/2}$	*dzó.*
	Kofi	say	COMP	**LogP/OrdP**	left

'Kofi said that he left.'

Previous findings, often centered on the *de se/de re* distinction, have served as baselines for co-reference patterns but rely on very small samples - no more than five participants (Adésola 2005; Manfredi 1987; Pearson 2015) - or on data from a single language, primarily Ewe (Bimpeh 2023). As

a result, the empirical picture is inconsistent and unstable, highlighting the need for systematically collected experimental data using more robust methods and a broader speaker pool.

We report results from an acceptability-judgment task examining the referential possibilities of LogP and OrdP in two contexts: 'self' (co-reference with the attitude holder) and 'anti-self' (disjoint reference), based on data from Ewe (n = 38), Igbo (n = 38), and Yoruba (n = 33) speakers. Methodologically, we introduce several refinements over previous work: Likert scales rather than binary choices (Marty, Chemla & Sprouse 2020); stimuli embedded in naturalistic contexts—archetypal environments licensing logophoric pronouns, notably reportative contexts, such as indirect speech reports (Clements 1975); and materials presented in the original languages rather than English translations. We advocate a principled division of labor in experimental design: native linguist-speakers construct materials and identify theoretically relevant contrasts, while experimental participants are non-linguists to minimize metalinguistic bias (Schütze 2015). Cultural and sociolinguistic factors (educational level and dialectal variation) were integrated in the results' analysis.

Under these conditions, a clear pattern emerges, partly in line with previous literature based on introspective judgements or small samples. Ewe and Yoruba speakers strongly prefer LogP to indicate the attitude holder in self contexts (Adésola 2005; Bimpeh et al. 2024; Clements 1975; Manfredi 1987; Pearson 2015). For OrdP, our results align with Clements (1975) and Bimpeh et al. (2024) for Ewe and Manfredi (1987) for Yoruba: OrdP is generally dispreferred in self contexts (contra Pearson 2015 for Ewe and Adésola 2005 for Yoruba). Igbo speakers pattern differently, accepting both LogP and OrdP in self contexts - consistent with Bimpeh et al. (2024) but contrary to Hyman & Comrie (1981) and Manfredi (1987).

These findings suggest that the apparent instability of acceptability judgments reflects not methodological failure but the interaction of grammatical structure, processing factors, and socially grounded variation. A quantitatively informed, culturally sensitive, and rigorous approach indicates that introspective data may hint at the theoretical relevance of a linguistic fact, but they need to be substantiated by empirical investigation.

References

Adésola, Oluseye. 2005. *Pronouns and Null Operators – A-bar Dependencies and Relations in Yoruba*. New Brunswick: Rutgers University. (PhD thesis)

Bimpeh, Abigail A., Driemel, Imke, Bassi, Itai & Silleresi, Silvia. 2024. Obligatory *de se* logophors in Ewe, Yoruba and Igbo: Variation and competition. In Lu, Jiayi, Petersen, Erika, Zaitsu, Anissa & Harizanov, Boris (eds.), *Proceedings of the 40th West Coast Conference on Formal Linguistics, Standford, 13-15 May 2022*, 1–10. Somerville (MA): Cascadilla Proceedings Project.

Chomsky, Noam. 1965. *Aspects of the Theory of Syntax*. Cambridge, MA: MIT Press.

Clements, George N. 1975. The logophoric pronoun in Ewe: Its role in discourse. *Journal of West African Languages* 10. 141–177.

Hyman, Larry M. & Comrie, Bernard. 1981. Logophoric reference in Gokana. *Journal of African Languages and Linguistics* 3(1). 19–37. https://doi.org/10.1515/jall.1981.3.1.19

Manfredi, Victor. 1987. Antilogophoricity as domain extension in Igbo and Yoruba. *Niger-Congo Syntax and Semantics* 1. 97–117.

Marty, Paul, Chemla, Emmanuel & Sprouse, Jon. 2020. The effect of three basic task features on the sensitivity of acceptability judgment tasks. *Glossa: A Journal of General Linguistics* 5(1). 72. https://doi.org/10.5334/gjgl.980

Myers, James. 2017. Acceptability judgments. In Aronoff, Mark (ed.), *Oxford Research Encyclopedia of Linguistics*. Oxford: Oxford University Press. https://doi.org/10.1093/acrefore/9780199384655.013.333

Pearson, Hazel A. 2015. The interpretation of the logophoric pronoun in Ewe. *Natural Language Semantics* 23(2). 77–118. https://doi.org/10.1007/s11050-015-9112-1

Schütze, Carson T. 2005. Thinking about what we are asking speakers to do. In Kepser, Stephan & Reis, Marga (eds.), *Linguistic Evidence: Empirical, Theoretical, and Computational Perspectives*. 457–485. Berlin: Mouton de Gruyter. https://doi.org/10.1515/9783110197549.457

Schütze, Carson T. 2015. *The Empirical Base of Linguistics: Grammaticality Judgments and Linguistic Methodology (Classics in Linguistics 2)*. Berlin: Language Science Press. https://doi.org/10.17169/langsci.b89.101

Sprouse, Jon. 2013. Acceptability judgments. In Sprouse, Jon (ed.), *The Oxford Handbook of Experimental Syntax*, 1–25. Oxford: Oxford University Press. https://doi.org/10.1093/oxfordhb/9780198797722.001.0001

GIORGIA TROIANI, EUGENIO GORIA E BEATRICE BERNASCONI

Challenges and practices of corpus design in multilingual societies

In this talk, we argue that theory and methods of spoken corpus construction are influenced by a monolingual bias, which translates into issues in the design of multilingual corpora. After identifying such issues, we systematically compare three case studies with different multilingual settings to present possible solutions.

Saluted as "objective" representations of language in use, corpora are nonetheless the result of theoretical choices, and thus subject to the epistemological biases of their creators (Sinclair 2005). This is especially true for corpora of spoken language, which require creators to engage with the issue of representation at three levels: i) selecting who counts as a 'speaker' of a language and defining the structure of their repertoire; ii) deciding which uses of the language are eligible for recording and archiving; iii) selecting how to represent the content of a multilingual recording in transcription. Each of these decisions can result in the erasure or marginalization of various aspects of language use and lead future users of the *dataset* to reproduce the bias. Common methods for corpus design implicitly reflect monolingual assumptions on the communities involved and their linguistic practices. This is especially true in language ecologies involving cultural minorities, where the linguists' orientation may reflect ideologies and attitudes of the community itself, including purist views that ultimately subscribe to a monolingual ideology (Woolard 2016).

To illustrate these shortcomings, we first compare the methodological choices in a number of existing resources representing communities in situations of language contact, including DoReCo (Seifart, Paschen & Stave 2025), Kontatto (Dal Negro & Ciccolone 2018), Miami Corpus (Deuchar 2010), and Microcontact (D'Alessandro et al. 2024). After analysing the theoretical and methodological choices of these resources in light of the issues i-iii discussed above, we identify the implications of such choices over the representation and analysis of multilingualism.

Drawing upon notes, training materials, and recordings from corpora that we are developing, we then discuss and systematically compare how areas i-iii have been dealt with in projects with different multilingual settings. Specifically, we compare the situations of large-scale longstanding fully institutionalized multilingualism in Kazakhstan (Multimodal Corpus of Spoken Kazakh Language; Troiani & Filchenko 2026), migration-related short term contact in Italy (Stra-ParlaTO subsection of the KIParla corpus; Goria & Bernasconi, *in prep.*), and the heritage language scenario of Piedmontese in Argentina (PILAR corpus; Goria & Gasparini 2025).

Based on this comparison, we contend that no one-size-fits-all solution is available for an adequate treatment of multilingualism in corpus resources. At the same time, a greater awareness of multilingualism as a common feature in most linguistic ecologies, along with a greater consideration of ethnographic knowledge on such communities, may at least partially contribute to overcoming monolingual biases. We believe that this contribution in corpus design theory is first of all relevant to linguists and NLP practitioners, because it promotes the collection of data that closely represent language use in everyday life. Moreover, incorporating a more realistic view of multilingualism in corpora may have effects for the societies portrayed by these resources, both by providing a more accurate representation of - often marginalized - multilingual individuals and by reducing stigmatisation based on specific linguistic practices.

References

D'Alessandro, Roberta, Andriani, Luigi, Frasson, Alberto, Pinto, Manuela, Sorgini, Luana & Terenghi, Silvia. 2025. Microcontact and syntactic theory. In D'Alessandro, Roberta, Putnam, Michael T. & Terenghi, Silvia (eds), *Heritage Languages and Syntactic Theory*, 17–53. Oxford: Oxford University Press. https://doi.org/10.1093/9780191987731.003.0003

Dal Negro, Silvia & Ciccolone, Simone. 2018. Il parlato bilingue: Italiano e tedesco a contatto in un corpus sudtirolese. In Bermejo Calleja, Felisa & Katelhön, Peggy (eds.), *Lingua parlata: Un confronto fra l'italiano e alcune lingue europee,* 385–407. Berlin: Peter Lang.

Deuchar, Margaret. 2010. *BilingBank Spanish-English Bangor Miami Corpus*. Dataset. TalkBank. https://doi.org/10.21415/T5J01D

Goria, Eugenio & Bernasconi, Beatrice. *In preparation*. *The StraParlaTo corpus of spoken Italian: multilingual practices in Turin*. Turin: University of Turin.

Goria, Eugenio & Gasparini, Fabio. 2025. Towards an Oral Corpus for Heritage Piedmontese. *Oral Archives Journal* 1. 35–55. https://doi.org/10.36253/oar-3337

Seifart, Frank, Paschen, Ludger & Stave, Matthew. 2025. *Language Documentation Reference Corpus (DoReCo) 2.0* (Version 13, p. 73515 bytes) NAKALA - https://nakala.fr (Huma-Num - CNRS). https://doi.org/10.34847/NKL.7CBFQ779

Sinclair, John. 2005. Corpus and Text-Basic Principles. In Wynne, Martin (eds.), *Developing Linguistic Corpora: A Guide to Good Practice*, 17–29. Oxford: Oxbrow.

Troiani, Giorgia & Filchenko, Andrey 2026. Introducing the First Module of the Multimedia Corpus of Spoken Kazakh Language. *Journal of Open Humanities Data* 12. 67. https://10.5334/johd.529

Woolard, Kathryn. A. 2016. *Singular and Plural: Ideologies of Linguistic Authority in 21st Century Catalonia*. Oxford University Press.

LUISA TRONCONE, FLAVIO PISCIOTTA, IOLANDA ALFANO, ALFONSINA BUONICONTO, GIOVANNI DI PAOLA, GIUSEPPE PISEGNA, CARMELA SAMMARCO, ANDREA SANSÒ E MIRIAM VOGHERA

L'annotazione del dato come pratica interpretativa: il caso dei segnali discorsivi

Questo contributo propone una riflessione epistemologica sull'annotazione dei segnali discorsivi (SD), a partire da esperienze maturate nel progetto COSÌ (*Catalogo Online dei Segnali discorsivi dell'Italiano,* Voghera et al. 2025). L'annotazione del livello pragmatico rappresenta un ambito fruttuoso di riflessione sia perché è stato generalmente più trascurato (Archer, Culpeper & Davies 2008; De Felice & Strik-Lievers 2024) sia perché ivi, più che in altri livelli, giocano un ruolo importante fattori come conoscenze enciclopediche, sensibilità personale, aspettative e modelli teorici impliciti (cfr. Latour & Woolgan 1979). Un'annotazione *pragmatica*, infatti, deve avere come oggetto non solo il segno ma anche il livello della sua relazione con i parlanti in contesto (Bourgeois & Bousfield 2025; Acher, Culpeper & Davies 2008). Indagheremo pertanto come queste due componenti del dato pragmatico influenzano le pratiche di annotazione, valutando la relazione tra: a) segno e suo co(n)testo, e b) segno e soggettività dei parlanti/annotatori.

I SD rappresentano in questo senso un caso esemplare, sia per le loro difficoltà di categorizzazione (Bazzanella 2011), identificazione e classificazione, sia perché non esistono pratiche di annotazione condivise, ma solo liste (non implementate) di *desiderata* per un'annotazione ideale (cfr. Crible & Cuenca 2017: 162). Esamineremo le pratiche di annotazione di due SD che presentano criticità differenti: *allora* (*i.a.* Bazzanella et al. 2008), la cui elevata polifunzionalità ne rende complessa la distinzione delle funzioni; e *comunque* (Proietti 2000), la cui difficoltà riguarda soprattutto la disambiguazione tra la funzione di SD e quella di connettivo. Per ognuno dei due SD, sulla base di 500 occorrenze da KIParla (Mauri et al. 2019), abbiamo svolto tre tipi di annotazione, secondo lo schema elaborato nel progetto COSÌ (Voghera et al. 2025):

1) due da due linguisti esperti in solitaria,
2) una da due linguisti, in coppia e in contemporanea,
3) due da due tirocinanti, non abituati alla ricerca sui SD in generale, ma addestrati in base a linee guida specifiche per il singolo SD.

Queste sono state accompagnate dalla redazione di cinque distinti diari di bordo (Silverio et al. 2022), redatti dai quattro annotatori in solitaria e dalla coppia di annotatori, per tenere traccia di casi dubbi e riflessioni.

Il confronto tra le annotazioni, e tra annotatori in ognuna di esse, consente di esplorare l'incidenza dei fattori soggettivi. L'accordo è valutato tramite a) l'α di Krippendorff (1980) e b)

un'analisi qualitativa e quantitativa dei contesti di (dis)accordo, considerando in particolare due fattori non strettamente pragmatici: la posizione del SD nell'enunciato e la sua collocazione rispetto allo *scope* (cfr. Crible & Cuenca 2017).

I primi risultati mostrano che proprio posizione e *scope* costituiscono l'ambito privilegiato di negoziazione tra annotatori alla ricerca di criteri per una minore soggettività del dato. L'annotazione in coppia evidenzia come le decisioni vengano riformulate nella discussione: il confronto porta spesso a esplicitare criteri che nell'annotazione individuale rimangono impliciti. I diari di bordo, poi, mostrano che le categorie non sono semplicemente applicate, ma progressivamente ridefinite attraverso esempi-limite, casi prototipici e controesempi. Si osserva inoltre che i tirocinanti, meno influenzati da aspettative teoriche pregresse, tendono ad aderire maggiormente alle linee guida, e a produrre annotazioni più coerenti sul piano procedurale, ma talvolta semplificatorie rispetto ai casi problematici individuati da annotatori esperti.

L'analisi delle divergenze suggerisce che l'annotazione pragmatica non può essere intesa come mera rilevazione, ma come pratica ermeneutica (cfr. Dascal 1987; Gadamer 1960) situata nell'esperienza. Le procedure di risoluzione dei disaccordi costituiscono così degli spazi di negoziazione concettuale, dove le categorie emergono tramite approssimazioni progressive a un'annotazione ottimale dei dati.

Riferimenti bibliografici

Archer, Dawn, Culpeper, Jonathan & Davies, Matthew. 2008. Pragmatic annotation. In Lüdeling, Anke & Kytö, Merja (a cura di), *Corpus Linguistics: An International Handbook,* 613–642. Berlin/New York: De Gruyter Mouton.

Bazzanella, Carla, Bosco, Cristina, Gili Fivela, Barbara, Miecznikowski, Johanna & Tini Brunozzi, Francesca. 2008. Segnali discorsivi e tipi di interazione: il caso di *allora*. In Bosisio, Cristina, Cambiagi, Bona, Piemontese, Emanuela & Santulli, Francesca (a cura di), *Aspetti linguistici della comunicazione pubblica e istituzionale. Atti del VII Congresso AItLA, Milano 22-23 febbraio 2007*, 239–265. Perugia: Guerra.

Bazzanella, Carla. 2011. Segnali discorsivi. In Simone, Raffaele (a cura di), *Treccani - Enciclopedia dell'italiano*. Roma: Istituto della Enciclopedia Italiana Treccani. https://www.treccani.it/enciclopedia/segnali-discorsivi_(Enciclopedia-dell'Italiano)/

Bourgeois, Samuel & Bousfield, Derek. 2025. *Pragmatics in Contested Interpretation. Varied Audiences, Varied Implicatures, Varied Inferences*. Cham: Palgrave Macmillan.

Crible, Ludivine & Cuenca, Maria-Josep. 2017. Discourse Markers in Speech: Distinctive Features and Corpus Annotation. In Stent, Amanda, Taboada, Maite, Fernández, Raquel, Traum, David, Poesio, Massimo, Di Eugenio, Barbara & Stede, Manfred (a cura di), *Dialogue and Discourse Volume 8,* 149–166. Bielefeld: University of Bielefeld. https://doi.org/10.5087/dad.2017.207

Dascal, Marcelo. 1987. Defending literal meaning. *Cognitive Science* 11(3). 259–281. https://doi.org/10.1207/s15516709cog1103_1

De Felice, Irene & Strik-Lievers, Francesca. 2024. Building a Pragmatically Annotated Diachronic Corpus: The DIADIta Project. In Dell'Orletta, Felice, Lenci, Alessandro, Montemagni, Simonetta & Sprugnoli, Rachele (a cura di), *Proceedings of the Tenth Italian Conference on Computational Linguistics (CLiC-it 2024)*, *Pisa, 4-6 dicembre 2024,* 266–272. Aachen: CEUR Workshop Proceedings. https://aclanthology.org/2024.clicit-1.32/

Gadamer, Hans-Georg. 1960. *Wahrheit und Method*. Tübingen: Mohr Siebeck.

Krippendorff, Klaus. 1980. *Content Analysis: An Introduction to Its Methodology*. Thousand Oaks/London/New Delhi: SAGE Publications, Inc.

Latour, Bruno & Woolgar, Steve. 1979. *Laboratory Life. The Construction of Scientific Facts*. Princeton: Princeton University Press.

Mauri, Caterina, Ballarè, Silvia, Goria, Eugenio, Cerruti, Massimo & Suriano, Francesco. 2019. KIParla Corpus: A New Resource for Spoken Italian. In Bernardi, Raffaella, Navigli, Roberto & Semeraro, Giovanni (a cura di), *Proceedings of the Sixth Italian Conference on Computational Linguistics (CLiC-it 2019), Bari, 13-15 novembre 2019.* 243–249. Aachen: CEUR Workshop Proceedings.

Proietti, Domenico. 2000. «Comunque» dalla frase al testo. *Studi di Grammatica Italiana* XIX. 175–232.

Silverio, Sergio A., Sheen, Kayleigh S., Bramante, Alessandra, Knighting, Katherine, Koops, Thula U., Montgomery, Elsa, November, Lucy, Soulsby, Laura K., Stevenson, Jasmine H., Watkins, Megan, Easter, Abigail & Sandall, Jane. 2022. Sensitive, Challenging, and Difficult Topics: Experiences and Practical Considerations for Qualitative Researchers. *International Journal of Qualitative Methods* 21. https://doi.org/10.1177/1609406922112473

Voghera, Miriam, Sansò, Andrea, Alfano, Iolanda, Buoniconto, Alfonsina, Cataldo, Violetta, Di Paola, Giovanni, Esposito, Pasquale, Pisciotta, Flavio, Sammarco, Carmela & Troncone, Luisa. 2025. COSÌ: an online repertoire of Italian discourse markers. Contributo presentato al workshop *Talking data - Methodological and theoretical challenges raised by spoken interaction data*, *Bologna 8-10 ottobre 2025*.

PYTTI VALVERDE ROCHA DINIZ SILVA, LEONARDO WELPEL TEIXEIRA AND MARLON NUNES SILVA

School small data and synthetic data: retention, interoperability, and ethics for a situated linguistic data science

Situated within Macro-theme 3 "Is there a 'linguistic data science'?" and the subtheme addressing methodological, ethical, and legal issues in data collection, this paper investigates the epistemological, methodological, and infrastructural conditions for the constitution, documentation, retention, interoperability, and reuse of linguistic data in technology-mediated educational contexts, whose forms of production and circulation are increasingly shaped by digital transformations in knowledge practices (Simone, 2000). The study draws on a qualitative corpus consisting of oral interactions, written productions, and reflective records generated in teacher-education pathways (Barni, 2015) aimed at developing discursive multipliers for tackling bullying in school settings, as well as pedagogical materials produced with the support of large language models used in teacher training (Bender *et al.*, 2021). Methodologically, we combine (i) systematic description of production metadata (institutional context, participants' sociobiographical profiles, conditions of circulation, and training framing), (ii) situated discourse analysis, and (iii) critical problematization of the data lifecycle, from its inception to potential reuse. The first axis examines the school-based small data produced in these formative contexts. Such data are dependent on specific sociobiographical and institutional conditions and often linked to experiences of violence that affect students and teachers at relational, emotional, and institutional levels. We argue that their collection, documentation, retention, and eventual reuse, including as legacy data, cannot be guided solely by technical criteria of archiving or standardization. The scientific quality of the data requires rigorous recording of conditions of production, clear definition of analytic categories, and, simultaneously, ethical protocols that account for participants' vulnerability and the potential effects of circulation of these materials (UNESCO, 2021). The second axis problematizes synthetic data generated by language models and incorporated into training practices. These outputs, derived from probabilistic inference over large corpora, circulate as pedagogical resources, simulation instruments, and objects of analysis. This paper discusses their status within a linguistic data science: they are not natural usage data, but devices for discursive experimentation, which require transparent documentation, explicit articulation of interpretive limits, bias control, source traceability, and interoperability between technical and pedagogical infrastructures (Buckingham Shum & Ferguson, 2012). The third axis addresses retention, interoperability, sustainability, and data governance in light of FAIR principles (Wilkinson *et al.,* 2016) and Open Science practices. We contend that, in educational settings, standardization and openness must be articulated through responsible management capable of balancing accessibility,

protection of sensitive data, legal compliance, and respect for subjectivities (Van Leeuwen and Rienties, 2021). We conclude that consolidating a linguistic data science applied to education requires integrating technical governance, epistemological reflection, and research ethics. The scientific status of data derives not merely from scale or formalization, but from documentation of production conditions, the possibility of sustainable interoperability, and responsibility in retention, reuse, and circulation across different institutional and formative contexts.

References

Barni, Monica. 2015. *Insegnare e valutare l'italiano nel mondo*. Bologna: Il Mulino.

Bender, Emily M., Timnit Gebru, Angelina McMillan-Major & Margaret Mitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623. New York: Association for Computing Machinery. https://doi.org/10.1145/3442188.3445922

Buckingham Shum, Simon & Rebecca Ferguson. 2012. Social learning analytics. *Educational Technology & Society* 15(3). 3–26. https://doi.org/10.1145/2330601.2330616

Simone, Raffaele. 2000. *La terza fase. Forme di sapere che stiamo perdendo*. Roma/Bari: Laterza.

UNESCO. 2021. Ethical guidelines for artificial intelligence in education. Paris: UNESCO.

Van Leeuwen, Anouschka & A. Rienties. 2021. Ethics and transparency in learning analytics. *British Journal of Educational Technology* 52(4). 1529–1545.

Wilkinson, Mark D., Dumontier, Michel, Aalbersberg, IJsbrand Jan, Appleton, Gabrielle, Axton, Myles, Baak, Arie, Blomberg, Niklas, Boiten, Jan-Willem, Bonino da Silva Santos, Luiz, Bourne, Philip E., Bouwman, Jildau, Brookes, Anthony J., Clark, Tim, Crosas, Mercè, Dillo, Ingrid, Dumon, Olivier, Edmunds, Scott, Evelo, Chris T., Finkers, Richard, Gonzalez-Beltran, Alejandra, Gray, Alasdair J.G., Groth, Paul, Goble, Carole, Grethe, Jeffrey S., Heringa, Jaap, 't Hoen, Peter A.C., Hooft, Rob, Kuhn, Tobias, Kok, Ruben, Kok, Joost, Lusher, Scott J., Martone, Maryann E., Mons, Albert, Packer, Abel L., Persson, Bengt, Rocca-Serra, Philippe, Roos, Marco, van Schaik, Rene, Sansone, Susanna-Assunta, Schultes, Erik, Sengstag, Thierry, Slater, Ted, Strawn, George, Swertz, Morris A., Thompson, Mark, van der Lei, Johan, van Mulligen, Erik, Velterop, Jan, Waagmeester, Andra, Wittenburg, Peter, Wolstencroft, Katherine, Zhao, Jun & Mons, Barend. 2016. The FAIR guiding principles for scientific data management and stewardship. *Scientific Data* 3. 160018. https://doi.org/10.1038/sdata.2016.18

ROSSELLA VARVARA

Dati distribuzionali per la semantica derivazionale

I *word* e *token embeddings* (e già i modelli di semantica distribuzionale) sono diventati strumenti centrali nell'elaborazione automatica del linguaggio naturale e vengono sempre più adottati anche nella linguistica teorica per studi di semantica quantitativi e basati sull'uso (Boleda 2020; Lenci 2018). Sebbene gli *embeddings* vengano generalmente costruiti per forme di parola o unità sub-lessicali, studi recenti hanno mostrato che è possibile derivare rappresentazioni vettoriali anche per morfemi attraverso operazioni sistematiche sui vettori delle parole (Guzmán Naranjo & Bonami 2023; Kisselew et al. 2015; Lazaridou et al. 2013; Marelli & Baroni 2015; Padó et al. 2016; Wauquier 2022).

In questo intervento, con casi di studio sulla morfologia derivazionale dell'italiano e del francese, mostro come gli *embeddings* offrano un quadro potente e concettualmente informativo per lo studio della semantica dei processi morfologici, permettendo analisi quantitative di proprietà tradizionalmente affrontate in modo qualitativo, come la trasparenza semantica, la rivalità tra affissi, la regolarità morfologica e la polifunzionalità.

Dopo una disamina delle diverse operazioni possibili per estrarre vettori di morfemi, come ad esempio la media dei vettori differenza derivato–base (*offset vectors*) o la stima di funzioni di trasformazione lineare, discuterò come sia possibile operazionalizzare nozioni centrali della teoria morfologica mediante misure distribuzionali.

In particolare, tratterò della trasparenza semantica, definita come il grado in cui il significato di una parola complessa può essere inferito dai suoi costituenti. Diversi indicatori distribuzionali della trasparenza saranno confrontati, tra cui la similarità tra i vettori della base e del derivato, misure basate sulla correlazione e metriche di predittibilità che confrontano i vettori osservati delle parole con vettori composti a partire dalle rappresentazioni dei morfemi (Kotowski & Schäfer 2023; Lombard et al. 2022; Reddy, McCarthy & Manandhar 2011; Stupak & Baayen 2022; Varvara & Huyghe 2025).

In secondo luogo, mi occuperò della rivalità tra affissi e della regolarità semantica, due nozioni strettamente correlate che descrivono la competizione tra processi morfologici e la stabilità del loro contributo semantico. Mediante la similarità tra vettori di affissi si potrà approssimare la rivalità, mentre misure di dispersione applicate ai vettori di *offset* possono essere utilizzate per quantificare la regolarità (Bonami & Paperno 2018; Guzmán Naranjo & Bonami 2023).

Infine, affronterò la relazione tra polisemia lessicale e polifunzionalità degli affissi. Utilizzando *embeddings* contestualizzati, si esplorerà se il *clustering* distribuzionale possa distinguere parole monosemiche da quelle polisemiche e se gli affissi associati a funzioni semantiche multiple mostrino una maggiore variabilità contestuale (Salvadori & Huyghe 2023; Varvara, Salvadori & Huyghe 2024).

Le misure distribuzionali così estratte saranno confrontate con uno standard di riferimento annotato manualmente, e si discuterà il contributo che i diversi tipi di dati (distribuzionali, quantitativi annotati manualmente, introspettivi qualitativi) hanno nella ricerca teorica.

Questo intervento ha l'obiettivo di mostrare come, nell'ambito della morfologia, i modelli distribuzionali possano sia rivelare nuovi *pattern* empirici, sia affinare distinzioni teoriche. Più in generale, intende mostrare le potenzialità di questo tipo di dati per lo studio teorico della semantica derivazionale, considerando anche la possibilità di rappresentare diverse varietà linguistiche, sia diacronicamente che sincronicamente.

Riferimenti bibliografici

Boleda, Gemma. 2020. Distributional semantics and linguistic theory. *Annual Review of Linguistics* 6(1). 213–234. https://doi.org/10.1146/annurev-linguistics-011619-030303

Bonami, Olivier & Paperno, Denis. 2018. Inflection vs. derivation in a distributional vector space. *Lingue e Linguaggio* 17(2). 173–196. https://doi.org/10.1418/91864

Guzmán Naranjo, Matias & Bonami, Olivier. 2023. A distributional assessment of rivalry in word formation. *Word structure* 16(1). 87–114. https://doi.org/10.3366/word.2023.0222

Kisselew, Max, Padó, Sebastian, Palmer, Alexis & Šnajder, Jan. 2015. Obtaining a better understanding of distributional models of german derivational morphology. In Purver, Matthew, Sadrzadeh, Mehrnoosh & Stone, Matthew (a cura di), *Proceedings of the 11th International Conference on Computational Semantics,* 58–63. https://aclanthology.org/W15-0108/

Kotowski, Sven & Schäfer, Martin. 2023. Quantifying semantic relatedness across base verbs and derivatives: English out- prefixation. In Kotowski, Sven & Plag, Ingo (a cura di), *The semantics of derivational morphology. Theory, methods, evidence*, 177–218. Berlin: De Gruyter.

Lazaridou Angeliki, Marelli, Marco, Zamparelli, Roberto & Baroni, Marco. 2013. Compositional-ly Derived Representations of Morphologically Complex Words in Distributional Semantics, In Schuetze, Hinrich, Fung, Pascale & Poesio, Massimo (a cura di), *Proceedings of the 51st*

Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Sofia, 4-9 agosto 2013, 1517–1526. Stroudsburg: Association for Computational Linguistics. https://aclanthology.org/P13-1149/

Lenci, Alessandro. 2018. Distributional models of word meaning. *Annual Review of Linguistics* 4. 151–171. https://doi.org/10.1146/annurev-linguistics-030514-125254

Lombard, Alizée, Wauquier, Marine, Fabre, Cécile, Hathout, Nabil, Ho-Dac, Lydia-Mai & Huyghe, Richard. 2022. Evaluating morphosemantic demotivation through experimental and distributional methods. *Lingvisticae investigationes* 45(1). 83–115. https://doi.org/10.1075/li.00068.wau

Marelli, Marco & Baroni, Marco. 2015. Affixation in semantic space: Modeling morpheme meanings with compositional distributional semantics. *Psychological Review* 122(3). 485–515. https://doi.org/10.1037/a0039267

Padó, Sebastian, Herbelot, Aurelie, Kisselew, Max & Snajder, Jan. 2016. Predictability of distributional semantics in derivational word formation, In Matsumoto, Yuji & Prasad, Rashmi (a cura di), *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*. Osaka (Japan), 11-16 dicembre 2016, 1285–1296. The COLING 2016 Organizing Committee. https://aclanthology.org/C16-1122/

Reddy, Siva, McCarthy, Diana & Manandhar, Suresh. 2011. An empirical study on compositionality in compound nouns. In Haifeng Wang & David Yarowsky (a cura di), *Proceedings of 5th International Joint Conference on Natural Language Processing*, *Chiang Mai, 8-13 novembre 2011*, 210–218. Asian Federation of Natural Language Processing. https://aclanthology.org/I11-1024/

Salvadori, Justine & Huyghe, Richard. 2023. Affix polyfunctionality in French deverbal nominalizations. *Morphology* 33(1). 1–39. https://doi.org/10.1007/s11525-022-09401-4

Stupak, Inna & Baayen, Harald. 2022. An inquiry into the semantic transparency and productivity of german particle verbs and derivational affixation. *The Mental Lexicon* 17(3). 422–457. https://doi.org/10.1075/ml.22012.stu

Varvara, Rossella, Salvadori, Justine & Huyghe, Richard. 2024. Lexical ambiguity in contextualized word embeddings: A case study of nominalizations. *Lingue e Linguaggio* 1/2024. 141–182. https://doi.org/10.1418/112743

Varvara, Rossella & Huyghe, Richard. 2026. The linguistic factors of semantic transparency: evidence from verb-to-noun derivation in French. *Journal of Linguistics* 62(1). 147–175. https://doi.org/10.1017/S0022226725000209

Wauquier, Marine. 2022. Apports de la sémantique distributionnelle pour la morphologie dérivationnelle. *Corpus* 23. https://doi.org/10.4000/corpus.6303

ALESSANDRO VIETTI

L'incertezza strutturata: il ruolo della probabilità in linguistica

All'interno delle scienze del linguaggio, alcune discipline hanno tradizionalmente coltivato una forte vocazione empirica: dalla linguistica computazionale (Manning & Schütze 1999) alla sociolinguistica variazionista (Cedergen & Sankoff 1974), dalla psicolinguistica (Chater & Manning 2006) alla fonetica e fonologia sperimentali (Pierrehumbert, Beckman & Ladd 2000). In questi ambiti, la probabilità rappresenta da tempo uno strumento fondamentale per modellizzare i fenomeni, descrivere la variazione, definire le categorie linguistiche e, più in generale, rendere conto del funzionamento delle grammatiche (per la fonologia, per esempio, già Cherry, Halle & Jakobson (1953).

Negli ultimi decenni, la crescente disponibilità di dati e la diffusione di tecniche di analisi quantitativa sempre più sofisticate hanno esteso l'uso dei metodi probabilistici a molti altri settori della linguistica (Bod, Hay & Jannedy 2003; Kortmann 2021). Questa trasformazione metodologica invita tuttavia a una riflessione più generale: che cos'è la probabilità in linguistica? Quale ruolo svolge oggi nella descrizione, nella spiegazione e nella teoria dei fatti linguistici?

La relazione si propone di rispondere a queste domande esaminando due modi di intendere la probabilità. Nella prima prospettiva, la probabilità è innanzitutto uno strumento necessario per studiare fenomeni o sistemi in condizioni di informazione incompleta. In questo senso, il ricorso alla probabilità nasce da uno stato di parziale ignoranza del ricercatore: poiché non possiamo determinare completamente la realtà che osserviamo, il linguaggio della probabilità ci consente di quantificare la nostra incertezza epistemica, ovvero di valutare la bontà dei nostri modelli alla luce dei dati (Chater et al. 2015).

All'interno di una seconda linea di indagine, si è progressivamente affermata l'idea che la probabilità non sia soltanto uno strumento esterno di modellizzazione, ma possa rappresentare una componente costitutiva del funzionamento dei sistemi linguistici. Secondo questa ipotesi, propria di approcci orientati all'uso (Bybee 2010; Tomasello 2009), il parlante-ascoltatore non si limita ad associare regole discrete a dati variabili, ma utilizza in modo sistematico informazioni di frequenza, distribuzione e prevedibilità per inferire categorie, stabilire relazioni, anticipare forme, interpretare significati e adattare il proprio comportamento linguistico all'interazione (Voghera 2017).

La probabilità, dunque, non descrive soltanto ciò che il linguista non sa, ma anche il modo in cui il parlante conosce, apprende e usa la lingua. Da questo punto di vista, il linguaggio può essere

inteso come un sistema sociocognitivo di tipo probabilistico nel quale l'incertezza non è semplice rumore, ma una dimensione strutturata dell'organizzazione linguistica. Le unità e le categorie non scompaiono, ma vengono intese come stabilizzazioni emergenti di pratiche d'uso e dinamiche interazionali (Goldrick & Cole 2023; Pierrehumbert 2003).

L'obiettivo della relazione è di discutere questa seconda ipotesi, concentrandosi in particolare sul livello fonetico-fonologico. La domanda centrale sarà dunque se la probabilità serva soltanto a descrivere i processi cognitivi di percezione e produzione, oppure se questa giochi un ruolo anche nell'organizzazione del sistema fonologico stesso (Bybee 2001). A partire da studi variazionisti e *usage-based*, si mostrerà come la variabilità empirica delle categorie e dei processi fonologici possa essere interpretata non come deviazione accidentale da un sistema astratto e deterministico, ma come manifestazione di una grammatica probabilistica, sensibile all'esperienza linguistica.

Riferimenti bibliografici

Bod, Rens, Hay, Jennifer & Jannedy, Stefanie (a cura di). 2003. *Probabilistic linguistics*. Cambridge (MA): MIT Press.

Bybee, Joan. 2001. *Phonology and language use*. Cambridge: Cambridge University Press.

Bybee, Joan. 2010. *Language, usage and cognition*. Cambridge: Cambridge University Press.

Cedergren, Henrietta J. & Sankoff, David. 1974. Variable rules: Performance as a statistical reflection of competence. *Language* 50(2). 333–355. https://doi.org/10.2307/412441

Chater, Nick, Clark, Alexander, Goldsmith, John A. & Perfor, Amy. 2015. Towards a new empiricism for linguistics. In Chater, Nick Clark, Alexander, Goldsmith, John A. & Perfor, Amy (a cura di), *Empiricism and language learnability*, 58–105. Oxford: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780198734260.003.0003

Chater, Nick & Manning, Christopher D. 2006. Probabilistic models of language processing and acquisition. *Trends in Cognitive Sciences* 10(7). 335–344. https://doi.org/10.1016/j.tics.2006.05.006

Cherry, Colin E., Halle, Morris & Jakobson, Roman. 1953. Toward the logical description of languages in their phonemic aspect. *Language* 29(1). 36–46. https://doi.org/10.2307/410451

Goldrick, Matthew & Cole, Jennifer. 2023. Advancement of phonetics in the 21st century: Exemplar models of speech production. *Journal of Phonetics* 99. 101254. https://doi.org/10.1016/j.wocn.2023.101254

Kortmann, Bernd. 2021. Reflecting on the quantitative turn in linguistics. *Linguistics* 59(5). 1207–1226. https://doi.org/10.1515/ling-2019-0046

Manning, Christopher D. & Schütze, Hinrich. 1999. *Foundations of statistical natural language processing*. Cambridge (MA): The MIT Press.

Pierrehumbert, Janet B. 2003. Probabilistic phonology: Discrimination and robustness. In Bod, Rens, Hay, Jennifer & Jannedy, Stefanie (a cura di), *Probabilistic Linguistics*, 177–228. Cambridge (MA): The MIT Press. https://doi.org/10.7551/mitpress/5582.003.0009

Pierrehumbert, Janet B., Beckman, Mary E. & Ladd, D. Robert. 2000. Conceptual foundations of phonology as a laboratory science. In Burton-Roberts, Noel, Carr, Philip & Docherty, Gerard (a cura di), *Phonological Knowledge*, 273–304. Oxford: Oxford University Press.

Tomasello, Michael. 2009. *Constructing a language: A usage-based theory of language acquisition*. Cambridge (MA): Harvard University Press. https://doi.org/10.2307/j.ctv26070v8

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